STATISTICAL PALEOCLIMATE RECONSTRUCTIONS VIA MARKOV RANDOM FIELDS

By Dominique Guillot *, Bala Rajaratnam* and Julien Emile-Geay†

Stanford University * and University of Southern California †

Understanding centennial scale climate variability requires datasets that are accurate, long, continuous, and of broad spatial coverage. Since instrumental measurements are generally only available after 1850, temperature fields must be reconstructed using paleoclimate archives, known as proxies. Various climate field reconstructions (CFR) methods have been proposed to relate past temperature to such proxy networks. In this work, we propose a new CFR method, called GraphEM, based on Gaussian Markov random fields embedded within an EM algorithm. Gaussian Markov random fields provide a natural and flexible framework for modeling the inherent spatial heterogeneities of high-dimensional spatial fields, which would in general be more difficult with standard parametric covariance models. At the same time, they provide the parameter reduction necessary for obtaining precise and well-conditioned estimates of the covariance structure of the field, a critical ingredient in all CFRs. Well-conditioned estimates can also be obtained even when the sample size is much smaller than the number of variables, as is typically the case in paleoclimate applications. We demonstrate how the graphical structure of the field can be estimated from the data via $\ell_1$-penalization methods, and how the GraphEM algorithm can be used to reconstruct past climate variations. The performance of GraphEM is then compared to the widely used CFR method RegEM with regularization via truncated total least squares, using synthetic data. Our results show that GraphEM can yield significant improvements, with uniform gains over space, and far better risk properties. We demonstrate that the increase in performance is directly related to recovering the underlying sparsity in the covariance of the spatial field, thus affirming the usefulness of $\ell_1$ model selection in this context. In particular, we show that spatial points with fewer neighbors in the recovered graph tend to be the ones where there are higher improvements in the reconstructions. The work demonstrates how significant improvements can be made in climate reconstruction methods by leveraging advances in high-dimensional statistics.

1. Introduction and preliminaries.

1.1. Introduction. Fundamental to an informed quantification of recent climate change is an accurate depiction of past climate variability (Jansen et al., 2007). Since widespread surface temperature instrumental measurements are only available after the mid-nineteenth century, climate scientists rely on proxy data (e.g. tree rings, ice cores, sediment cores, corals) to infer past temperatures via statistical modeling (National Research Council, 2006; Jones et al., 2009) – a task known as “paleoclimate reconstruction” in the climate literature. Given an instrumental temperature dataset (see e.g. Brohan et al., 2006) and a global network of climate proxies (Mann et al., 2008, Fig 2), the temperature back in time can be estimated as a function of proxies (or vice-versa).

Keywords and phrases: Climate reconstructions, Markov random fields, Covariance matrix estimation, Sparsity, Model selection, Pseudoproxies
Various CFR methods have been proposed to infer past climate (see Tingley et al., 2012). Here we adopt an approach based on multivariate linear regression as in the regularized EM algorithm (Schneider, 2001). In that setting, the CFR problem is formalized as a missing data problem, which we now describe.

Consider a spatial grid and let \( p \) denote the number of temperature and proxy points. Let \( n = n_a + n_m \) denote the sum of the number of years of available instrumental data, \( n_a \), and missing data, \( n_m \). In practice, \( p \approx 3000, n \approx 2000 \) and \( n_a \approx 150 \) (instrumental period). We model the temperature and proxy points as a multivariate random vector \((X_1, \ldots, X_p) \sim N_p(\mu, \Sigma)\) with missing values, where \( \mu = (\mu_1, \ldots, \mu_p) \) is the mean vector and \( \Sigma = (\sigma_{ij})_{p \times p} \) is the covariance matrix of the model. We denote by \( X \) the (incomplete) \( n \times p \) data matrix where each row represents a year of observations containing \( r \) instrumental temperature observations and \( s \) proxy measurements. Hence the rows represent time order and the columns represent different spatial locations of both instrumental temperature and proxy data (see Figure 1).

Figure 2 shows that the availability of the proxy data from the network of Mann et al. (2008) (hereafter Mann08) decreases rapidly in time, and missing values constitute as much as 80% of the entries in the matrix. Reconstructing the pre-instrumental temperature field may be cast as a missing data problem, for which several strategies exist (Little and Rubin, 2002). However, the high dimensionality of the problem ("large \( p \), small \( n \)"") makes it extremely challenging to apply standard methods. For instance, it is well-known that the sample covariance matrix is a poor estimator of \( \Sigma \) in that setting (Stein, 1975, 1986; Lin and Perlman, 1985; Paul, 2007). In this paper, we explore the use of Gaussian Markov random fields (a.k.a. Gaussian graphical models) for estimating \( \Sigma \). This approach provides flexibility in terms of modeling the inherent spatial heterogeneities of the field, but at the same time reduces the number of parameters that need to be estimated, thereby leading to improved reconstructions of past temperature. We start by recounting existing reconstruction strategies, before introducing our new approach.

1.2. The EM algorithm. A popular method for the imputation of missing values is the EM algorithm (Dempster, Laird and Rubin (1977); Little and Rubin (2002). In the multivariate normal setting, given an estimate of \( \mu \) and \( \Sigma \), the EM algorithm reduces to regressing the missing values on the available ones, and thereafter updating the estimates of \( \mu \) and \( \Sigma \) by computing the sample mean and sample covariance matrix of the completed dataset. This procedure is iterated until convergence.
More precisely, let \( x \) denote the \( k \)-th row of \( X \), and let \( x_a \) and \( x_m \) denote the parts of \( x \) where data are available and missing, respectively. Let \( \mu^{(0)} \) and \( \Sigma^{(0)} \) be initial estimates of \( \mu \) and \( \Sigma \). For example, \( \mu^{(0)} \) and \( \Sigma^{(0)} \) could be the sample mean and sample covariance of the dataset completed by replacing every missing value by the mean of the available values in the corresponding columns of \( X \) (Schneider, 2001). The EM algorithm iteratively constructs a sequence \( \mu^{(l)} \) and \( \Sigma^{(l)} \) of estimates of \( \mu \) and \( \Sigma \). For every \( l \geq 0 \), the E-step consists of a linear regression

\[
(x_m - \mu_m^{(l)})^\top = B^{(l)}(x_a - \mu_a^{(l)})^\top,
\]

where

\[
B^{(l)} = \Sigma_{ma}^{-1}(\Sigma_{aa})^{-1}, \quad \Sigma^{(l)} = \begin{pmatrix} \Sigma_{aa} & \Sigma_{am} \\ \Sigma_{ma} & \Sigma_{mm} \end{pmatrix} \quad \text{and} \quad \mu^{(l)} = (\mu_a^{(l)}, \mu_m^{(l)}),
\]

are the regression coefficients and the decompositions of \( \Sigma^{(l)} \) and \( \mu^{(l)} \) associated with the decomposition of \( x \) among its available and missing parts. Denote by \( X^{(l+1)} \) the completed estimate of \( X \), obtained after the regression (1.1) has been performed in order to impute the missing values in each row of \( X \).

In the M-step of the algorithm, the estimates of \( \mu \) and \( \Sigma \) are updated by:

\[
\mu_i^{(l+1)} = \frac{1}{n} \sum_{k=1}^{n} X_{ki}^{(l+1)}, \quad \Sigma_{ij}^{(l+1)} = \frac{1}{n} \sum_{k=1}^{n} \left[ (X_{ki}^{(l+1)} - \mu_i^{(l+1)})(X_{kj}^{(l+1)} - \mu_j^{(l+1)}) \right] + C_{ij}^{(l+1)},
\]

where \( C_{ij}^{(l+1)} \) is the covariance of the residuals. Using the same block decomposition as in (1.2), we have:

\[
C^{(l+1)} = \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{mm} - \Sigma_{ma}(\Sigma_{aa})^{-1}\Sigma_{am} \end{pmatrix}
\]
The reader is referred to Little and Rubin (2002) and McLachlan and Krishnan (2008) for more details about the EM algorithm.

1.3. The regularized EM algorithm. Obtaining a precise estimate of \( \Sigma \) is a crucial step of the EM algorithm. In paleoclimate problems, the number of variables, \( p \), is generally much larger than the number of observations \( n \), and thus estimating \( \Sigma \) can be a challenging task. In particular, the sample covariance matrix is not invertible and can be a very poor estimator of \( \Sigma \). This can be a serious problem since parts of \( \Sigma \) need to be inverted to compute the regression coefficients \( B \). Different \( \ell_2 \)-type methods to regularize the problem have been proposed in the literature. Amongst them are ridge regression (a.k.a. Tikhonov regularization, Hoerl and Kennard, 1970a,b; Tikhonov and Arsenin, 1977; Hanke and Hansen, 1993; Hastie, Tibshirani and Friedman, 2008) and truncated total least squares (TTLS, Fierro et al., 1997; Van Huffel and Vandewalle, 1991; Golub and Van Loan, 1980) regression. These methods can be used to replace the regression matrix \( B^{(l)} \) in equation (1.1) by a regularized estimate, and have been implemented within the EM algorithm. The resulting algorithm is known as RegEM (Schneider, 2001), and has been widely used in paleoclimate studies (Mann et al., 2005; Rutherford et al., 2005; Mann et al., 2007a; Riedwyl et al., 2008; Mann et al., 2008, 2009; Emile-Geay et al., 2013a,b). To date, all direct regression methods have resulted in reconstructions that underestimate the amplitude of past climate variations to some extent. (e.g. von Storch et al., 2004; Smerdon et al., 2010, 2011). This “regression dilution (Frost and Thompson, 2000) is a direct consequence of modeling the temperature conditional on (noisy) proxy values (von Storch et al., 2004; Christiansen, 2010; Tingley and Li, 2012; Christiansen, 2013). Regularization may compound this problem, as with ridge regression the smoothness of the filter factors has been shown to leak energy from the leading SVD modes, resulting in overly damped estimates of past temperature (Smerdon and Kaplan, 2007). This problem may be mitigated via TTLS (Mann et al., 2007b) as it attempts to correct for regression dilution by steepening the regression slope; however, the solution is no longer guaranteed to be optimal even under broad assumptions (Carroll and Ruppert, 1996). Furthermore, a major shortcoming of TTLS as currently used in climate applications is that the truncation parameter must be specified a priori, rather than being estimated adaptively. Given the applicability of the RegEM algorithm for missing data problems in the paleoclimate context (e.g., surface temperature reconstructions for the past 2000 years), we seek to develop an imputation method that rests on a more accurate and data-adaptive estimate of \( \Sigma \) itself.

In this paper, we show how Gaussian Markov random fields (GMRFs) can be used to improve the estimate of \( \Sigma \), with important consequences for paleoclimate applications. We begin with general notions regarding GMRFs, before providing details as to how they may be used in practice.

1.4. Gaussian Markov random fields. A GMRF is a multivariate normal model which encodes conditional independence structure between variables (see Whittaker (1990); Lauritzen (1996)). More precisely, let \((X_1, \ldots, X_p)\) be a multivariate random vector with inverse covariance matrix (or precision matrix) \( \Omega = (\omega_{ij}) = \Sigma^{-1} \). The partial correlation coefficient between \( X_i \) and \( X_j \) given the rest of the variables, denoted by \( \rho_{ij|\text{rest}} \), can be obtained from the inverse covariance matrix (see Whittaker (1990) Corollary 5.8.2, Lauritzen (1996)), and is given as follows:

\[
\rho_{ij|\text{rest}} = \frac{-\omega_{ij}}{\sqrt{\omega_{ii} \omega_{jj}}}.
\]

In the case of multivariate normal data, one can show that \( \rho_{ij|\text{rest}} = 0 \) if and only if \( X_i \) is independent of \( X_j \) given the rest of the variables (Whittaker (1990) Corollary 6.3.4). The zeros in the precision matrix therefore indicate conditional independence between the corresponding variables.

The conditional independence relations in a distribution can be conveniently encoded using a graph. Recall that a graph \( G = (V, E) \) is a pair of sets \( V \) and \( E \subset V \times V \), where each element of \( V \) represents a vertex of the graph, and each point of \( E \) is a pair of elements of \( V \). We encode the conditional
independence relations by adding an edge between \( i \) and \( j \) if and only if \( X_i \) is not conditionally independent of \( X_j \) given the rest of the variables. The random vector \((X_1, \ldots, X_p)\) is then said to satisfy the pairwise Markov property with respect to the graph \( G \). We refer the reader to Whittaker (1990); Lauritzen (1996) for details on the pairwise, local and global Markov properties and other common notions in graphical models.

Once the conditional independence structure (or graphical structure) of a Gaussian random vector is known, this information can be used for estimating its covariance matrix \( \Sigma \). More specifically, given an i.i.d. sample \( x_1, \ldots, x_n \) of \((X_1, \ldots, X_p)\), and a graph \( G \), the graphical maximum likelihood estimator of \( \Sigma \) can be computed by solving

\[
\hat{\Sigma}_G = \text{argmax}_{\Sigma = \Omega^{-1} > 0} \log \det \Omega - \text{tr}(S\Omega),
\]

where \( S \) is the sample covariance matrix of \( x_1, \ldots, x_n \) given by

\[
S = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top,
\]

and \( \log \det \Omega - \text{tr}(S\Omega) \) is (up to a constant) the multivariate normal log-likelihood function. The problem (1.6) can be solved efficiently for up to a few thousand variables using, for example, regression-based algorithms (see Hastie, Tibshirani and Friedman (2008), Algorithm 17.1). The resulting matrix \( \hat{\Sigma}_G \) is generally a better estimate than the widely used sample covariance matrix, especially when the number of observations \( n \) is smaller than the number of variables \( p \).

In this paper, we combine graphical models with the EM algorithm in order to reconstruct past temperature fields. In our approach, we first model the conditional independence structure of the target field using data from the instrumental period. A sparse estimate of \( \Sigma \) is then obtained in accordance with this graphical structure at every step of the EM algorithm. This approach greatly reduces the number of parameters to estimate, leads to better conditioned and more precise estimates of \( \Sigma \), and also exploits the natural conditional independence structure of the spatial field. The regression step (1.1) can then be performed using any regularization method (or even no regularization at all). We call the resulting algorithm GraphEM. The exact algorithm is given in Section 2 together with a method to estimate the graphical structure of field. We then test the performance of GraphEM in a realistic geophysical context (Sections 3 and 4), and demonstrate that GraphEM can lead to important gains in accuracy and precision in comparison with previous methods. The characteristics of the estimated conditional independence structures are then studied in Section 5. In particular, we find a Markov random field structure in the spatial random field, thus leading to sparse graphs and a parsimonious representation of the covariance parameter. We demonstrate that the improvement in reconstruction can be attributed to this sparse estimation, and that this approach yields a more stable paleoclimate reconstruction. We conclude with a discussion section.

2. Methodology.

2.1. The GraphEM algorithm. We now describe the GraphEM algorithm. The first step of the algorithm is to obtain an estimate of the conditional independence structure of the field. Once this is done, the missing values in the field are reconstructed using the EM algorithm, where the covariance matrix is estimated by (1.6) at every iteration (see Algorithm 2 in Appendix A for a detailed description of the algorithm). The E-step and M-step of the algorithm correspond to the E and M step of the EM algorithm for a multivariate normal model with a Markov random field structure encoded by a graph \( G = (V, E) \) (see Appendix B for more details). In the case where \( G \) is a complete graph (i.e., a graph where every pair of vertices is connected), the GraphEM algorithm reduces to the classical
EM algorithm for a Gaussian model. A crucial step in GraphEM is the estimation of the graphical structure of the field, which we now describe.

2.2. Graph estimation. Different methods have been proposed in the literature to discover the conditional independence relations (or graphical structure) of a dataset, in either the Bayesian or frequentist framework (see for e.g. Banerjee, El Ghaoui and d’Aspremont (2008); Friedman, Hastie and Tibshirani (2008); Dawid and Lauritzen (1993); Letac and Massam (2007); Rajaratnam, Massam and Carvalho (2008)). In this work, we explore two different approaches: neighborhood graphs and $\ell_1$-penalized maximum likelihood (Banerjee, El Ghaoui and d’Aspremont (2008); Friedman, Hastie and Tibshirani (2008); Guillot et al. (2012); jui Hsieh et al. (2011)).

2.2.1. Neighborhood graphs. Since temperature variations at a given point are to a large extent explained by temperature of surrounding points, it is natural to use a neighborhood graph to approximate the true graphical structure of the field (i.e., a graph where two vertices are connected if and only if they are within a specified radius $R$). The choice of the radius can be either specified, or chosen from the data. As we illustrate in Figure 3, the choice of an optimal radius can be made by performing cross-validation over the instrumental period, and choosing the radius that minimizes the MSE of the reconstructed values.

Besides this natural and meaningful Markov random field structure in spatial temperature fields, a neighborhood graph approach has the distinct advantage that the underlying graph does not have to be estimated from sample-deficient high dimensional data. Dimensionality reduction is achieved with great ease and at the same time has an intuitive geophysical interpretation; sparsity is entirely governed by the neighborhood radius $R$. On the other hand, neighborhood graphs are less flexible and cannot model in an adaptive way (1) conditional independence relations resulting from anisotropic structures present in the data (such as land/ocean boundaries, mountain ranges, atmospheric flow patterns, etc.), and (2) long range dependences that arise due to teleconnections. As an illustration, Fig. 4 displays the neighborhood graphs with radius 2000 km for 6 locations on the globe.

![Fig 3: Cross-validation scores for choosing a neighborhood graph (5-fold cross-validation)](image-url)
2.2.2. $\ell_1$-penalized maximum likelihood. Another approach for obtaining a sparse estimate of the precision matrix $\Omega$ is to maximize the normal likelihood subject to an $\ell_1$ penalty on its norm. More specifically, let $S$ denote the sample covariance matrix of the data matrix $X$:

$$S = \frac{1}{n}X^\top X.$$  

The $\ell_1$-penalized maximum likelihood problem consists of solving

$$\max_{\Omega > 0} l(\Omega) - \rho \|\Omega\|_1,$$

where $\Omega = \Sigma^{-1}$ denotes the precision matrix of the data, $l(\Omega)$ is the normal log-likelihood of $\Omega$, and $\|\Omega\|_1$ is the 1-norm of $\Omega$:

$$\|\Omega\|_1 = \sum_{i=1}^{p} \sum_{j=1}^{p} |\Omega_{ij}|,$$

and $\rho > 0$ is a regularization parameter. The use of an $\ell_1$ penalty, as first introduced in the context of the LASSO regression (Tibshirani, 1996), favors the introduction of zero elements and thus leads to sparse solutions (see Hastie, Tibshirani and Friedman (2008), Section 3.4.3). At the same time, using an $\ell_1$ penalty leads to a convex problem that can be solved efficiently using modern methods of convex optimization. Once an estimate of $\Omega$ is known, the associated graph can be inferred from the pattern of zeros in $\Omega$. In this work, we employ the *graphical lasso* (glasso) algorithm of Friedman, Hastie and Tibshirani (2008) to obtain a sparse estimate of $\Omega$ by solving an $\ell_1$-penalized likelihood problem. Problem (2.2) can thus be replaced by

$$\max_{\Omega > 0} l(\Omega) - \rho_{TT}\|\Omega_{TT}\|_1 - 2\rho_{TP}\|\Omega_{TP}\|_1 - \rho_{PP}\|\Omega_{PP}\|_1,$$

where $\Omega_{TT}, \Omega_{TP}$ and $\Omega_{PP}$ are block matrices corresponding to the temperature/temperature, temperature/proxy, and proxy/proxy parts of the matrix. Since the signal contained in proxies is generally weaker than the temperature signal, it may be sensible to use different penalty parameters for different parts of the matrix when solving the $\ell_1$-penalized maximum likelihood problem. Problem (2.2) can thus be replaced by

$$\max_{\Omega > 0} l(\Omega) - \rho_{TT}\|\Omega_{TT}\|_1 - 2\rho_{TP}\|\Omega_{TP}\|_1 - \rho_{PP}\|\Omega_{PP}\|_1,$$

and $\rho > 0$ is a regularization parameter. The use of an $\ell_1$ penalty, as first introduced in the context of the LASSO regression (Tibshirani, 1996), favors the introduction of zero elements and thus leads to sparse solutions (see Hastie, Tibshirani and Friedman (2008), Section 3.4.3). At the same time, using an $\ell_1$ penalty leads to a convex problem that can be solved efficiently using modern methods of convex optimization. Once an estimate of $\Omega$ is known, the associated graph can be inferred from the pattern of zeros in $\Omega$. In this work, we employ the *graphical lasso* (glasso) algorithm of Friedman, Hastie and Tibshirani (2008) to obtain a sparse estimate of $\Omega$ by solving an $\ell_1$-penalized likelihood problem. As $\rho$ varies, the matrix $\hat{\Omega}$ displays different sparsity patterns. When $\rho = 0$ and $n \geq p$, there is no penalty and $\hat{\Omega}$ is equal to the maximum likelihood estimate $S^{-1}$ of $\Omega$. The estimate $\hat{\Omega}$ tends to a diagonal matrix as the regularization parameter $\rho$ is increased. Fig. 5 displays the neighborhood of a few locations for a graph estimated on a real temperature dataset (Brohan et al., 2006).

We note that problem 2.2 can also be easily modified to use a different penalty for different parts of the matrices. Consider, for example, the precision matrix of a temperature/proxies field. The matrix can be organized in block form

$$\Omega = \begin{pmatrix} \Omega_{TT} & \Omega_{TP} \\ \Omega_{PT} & \Omega_{PP} \end{pmatrix},$$

where $\Omega_{TT}, \Omega_{TP}$ and $\Omega_{PP}$ are block matrices corresponding to the temperature/temperature, temperature/proxy, and proxy/proxy parts of the matrix. Since the signal contained in proxies is generally weaker than the temperature signal, it may be sensible to use different penalty parameters for different parts of the matrix when solving the $\ell_1$-penalized maximum likelihood problem. Problem (2.2) can thus be replaced by

$$\max_{\Omega > 0} l(\Omega) - \rho_{TT}\|\Omega_{TT}\|_1 - 2\rho_{TP}\|\Omega_{TP}\|_1 - \rho_{PP}\|\Omega_{PP}\|_1,$$

and $\rho > 0$ is a regularization parameter. The use of an $\ell_1$ penalty, as first introduced in the context of the LASSO regression (Tibshirani, 1996), favors the introduction of zero elements and thus leads to sparse solutions (see Hastie, Tibshirani and Friedman (2008), Section 3.4.3). At the same time, using an $\ell_1$ penalty leads to a convex problem that can be solved efficiently using modern methods of convex optimization. Once an estimate of $\Omega$ is known, the associated graph can be inferred from the pattern of zeros in $\Omega$. In this work, we employ the *graphical lasso* (glasso) algorithm of Friedman, Hastie and Tibshirani (2008) to obtain a sparse estimate of $\Omega$ by solving an $\ell_1$-penalized likelihood problem. As $\rho$ varies, the matrix $\hat{\Omega}$ displays different sparsity patterns. When $\rho = 0$ and $n \geq p$, there is no penalty and $\hat{\Omega}$ is equal to the maximum likelihood estimate $S^{-1}$ of $\Omega$. The estimate $\hat{\Omega}$ tends to a diagonal matrix as the regularization parameter $\rho$ is increased. Fig. 5 displays the neighborhood of a few locations for a graph estimated on a real temperature dataset (Brohan et al., 2006).

We note that problem 2.2 can also be easily modified to use a different penalty for different parts of the matrices. Consider, for example, the precision matrix of a temperature/proxies field. The matrix can be organized in block form

$$\Omega = \begin{pmatrix} \Omega_{TT} & \Omega_{TP} \\ \Omega_{PT} & \Omega_{PP} \end{pmatrix},$$

where $\Omega_{TT}, \Omega_{TP}$ and $\Omega_{PP}$ are block matrices corresponding to the temperature/temperature, temperature/proxy, and proxy/proxy parts of the matrix. Since the signal contained in proxies is generally weaker than the temperature signal, it may be sensible to use different penalty parameters for different parts of the matrix when solving the $\ell_1$-penalized maximum likelihood problem. Problem (2.2) can thus be replaced by

$$\max_{\Omega > 0} l(\Omega) - \rho_{TT}\|\Omega_{TT}\|_1 - 2\rho_{TP}\|\Omega_{TP}\|_1 - \rho_{PP}\|\Omega_{PP}\|_1,$$
Fig 5: Estimated graphical structure of a temperature field (HadCRUT3v) for different regularization parameters.

where $\rho_{TT}, \rho_{TP}, \rho_{PP} > 0$ are regularization parameters. This problem can also be solved efficiently by using a modified graphical lasso algorithm (see Friedman, Hastie and Tibshirani (2008), equation (15)).

2.2.3. Choice of regularization parameter. We now turn to the problem of choosing a suitable penalty parameter $\rho$ in optimization problem (2.2). A high penalty forces many zero entries in the precision matrix, while a low penalty adds some edges that make little geophysical sense to the graph. An optimal choice should find a balance between those extremes.

If $\rho > \rho_{\text{max}} := \max_{i \neq j} |S_{ij}|$, it can be shown (see e.g. (Witten, Friedman and Simon, 2011, Theorem 2)) that the resulting glasso estimate of $\Sigma$ is a diagonal matrix. A relevant finite number of regularization parameter values can therefore be obtained by dividing the interval between some small value $\rho_{\text{min}}$ and the biggest relevant value $\rho_{\text{max}}$. In our numerical work, we have chosen $\rho_{\text{min}} = 0.1 \cdot \rho_{\text{max}}$, and have divided the interval $[\rho_{\text{min}}, \rho_{\text{max}}]$ into 10 values. Problem (2.2) can then be solved for each of these penalty parameters to obtain estimates $\hat{\Omega}$ of the precision matrix $\Omega$. To each estimate corresponds a graph based on the structure of zeros in $\hat{\Omega}$. When the dimension of the problem to solve is small
PALEOCLIMATE RECONSTRUCTION VIA MRF

(e.g. in regional reconstructions) or a single penalty parameter is used for the whole precision matrix (as in Equation (2.2)) an optimal parameter can be chosen using k-fold cross-validation. However, when a different penalty parameter is used for each part of the precision matrix (see Problem (2.5)), performing cross-validation for an array of regularization parameters (for e.g. a $10 \times 10 \times 10$ grid of penalty parameters) incurs a prohibitive computational cost. We propose a different method based on the sparsity (% of possible edges) of the resulting graph in that case, which we now describe.

The ultimate benefit of using a Markov random field is the important dimensionality reduction obtained by modeling the spatial structure of the field. Hence a solution consists of searching for a graph that is a) dense enough to capture the salient spatial dependences, and b) sparse enough to make the reconstruction possible and stable (by reducing the dimension of the problem to a size comparable to the sample size). This technique is implemented in our proposed version of GraphEM, and is our current method of choice for identifying the graphical structure of climate fields. The main advantage of this method is that it is much less computationally intensive than performing k-fold cross-validation to select an optimal penalty parameter. Choosing the graph based on its sparsity level makes explicit the tradeoff between working with a complex model, and a model that is sparse enough to achieve this dimension reduction.

To avoid these problems, one may use the sparsity level approach and select a triple $(\rho_{TT}, \rho_{TP}, \rho_{PP})$ of regularization parameters with a greedy search algorithm. The algorithm starts with large values of the three penalty parameters, and progressively reduces the value of each penalty parameter until a given target sparsity is obtained for each part of the precision matrix. At each step of the algorithm the value of the three penalty parameters are chosen in order to become close to the target sparsity level, i.e., the penalty parameters are reduced if the corresponding part of the graph is too sparse and increased if it is too dense. This technique requires computing the solution of Problem (2.5) at only a few points of the grid, and generally leads to graphs with good properties. In this paper, we have chosen fixed sparsity levels when performing large reconstruction ensembles, after verifying via targeted experiments that the specified sparsity levels were close to those deemed optimal by 5-fold cross-validation.

3. Validation via pseudoproxy experiments.

3.1. Background. In the climate literature, pseudoproxy experiments have become the method of choice to objectively evaluate the performance of CFR techniques against a geophysically-relevant target (see Smerdon, 2011, for a recent review). This target temperature field is often the output of coupled general circulation model (GCM) simulations for the past 1000 years or so, sampled at a fixed spatiotemporal resolution. Although GCM-simulated temperature fields do not exactly match the characteristics of observed temperature fields, they are generated in accordance with physical laws embedded in such models, and thus provide a controlled, realistic framework to test reconstruction methods.

In practice, a pseudoproxy is obtained by adding noise to a GCM-simulated temperature field at locations where proxy observations are available in the real world. Because such observations are sparse, the pseudoproxy network therefore comprises a small collection of time series. Given only knowledge of the temperature field over a 150-year calibration interval, the CFR method is then used to backcast a thousand-year long temperature field based on this relatively small sample of noisy temperature time series. Given a simulated temperature field $T(l, t)$ at location $l$ and time $t$ (standardized to have mean 0 and variance 1 over time) from a GCM model, the pseudoproxies $P(l, t)$ are constructed as follows:

$$P(l, t) = T(l, t) + \frac{1}{\text{SNR}} \cdot \xi(l, t)$$

where $\xi(l, t)$ are independent realizations of a Gaussian white noise process, and the (scalar) signal-to-noise ratio SNR controls the amount of noise in the pseudoproxy (see Christiansen (2010); Christiansen
D. GUILLOT, B. RAJARATNAM, AND J. EMILE-GEAY

Fig 6: Geographic location of the pseudoproxies in the MBH98 database

and Ljungqvist (2011) for other possible modeling approaches). Although pseudoproxies constitute an oversimplification of reality, they have been used extensively in the climate literature (Bradley, 1996; Mann and Rutherford, 2002; Christiansen, Schmith and Thejll, 2009; Tingley and Huybers, 2010a; Smerdon et al., 2011; Smerdon, 2011; Li and Smerdon, 2012; Annan and Hargreaves, 2012) to provide a realistic numerical laboratory to test the performance of CFR methods.

In our simulations, we used the NCAR CSM 1.4 model experiment (Ammann et al., 2007), which simulates the climate of the last millennium (850-1980 AD). As per previous work (Mann et al., 2007a; Li, Nychka and Ammann, 2010; Smerdon et al., 2011; Wang et al., 2014), the location of the pseudoproxies were chosen in accordance with the Mann, Bradley and Hughes (1998) database (MBH98, Fig. 6) and the value of SNR has been fixed to 0.5. Other SNR values have also been investigated but, for the sake of brevity, are not presented here. The last 150 years of data have been used as a calibration period, and the remaining 981 years of temperature data have been reconstructed using GraphEM. As a benchmark, we follow recent work (Tingley and Huybers, 2010a; Steiger et al., 2013) and use RegEM-TTLS, which was widely used in high-profile climate reconstructions (Mann et al., 2008, 2009). In RegEM-TTLS, the linear regressions in the EM algorithm are replaced by truncated total least squares (TTLS) regressions. The TTLS solution of a linear system \( Ax = b \) is obtained by expressing the total least squares solution of the linear system as a function of the SVD of the matrix \( A \), and then truncate all but a given number of eigenvalues. The number of retained eigenvalues corresponds to the truncation parameter of RegEM-TTLS (see Fierro et al. (1997) for more details).

3.2. Performance metrics. Various metrics have been used in the literature to measure the quality of CFR methods and reconstructed temperature fields (Cook et al. (1999), Bürger (2007)). Let \( T(l, t) \) denote the temperature at a location \( l \) and at time \( t \), and denote by \( \hat{T}(l, t) \), a reconstruction of \( T(l, t) \). The mean squared error (MSE) measures the mean difference between the two fields at a given location \( l \):

\[
\text{MSE}(\hat{T})(l) = \frac{1}{N} \sum_{t} (T(l, t) - \hat{T}(l, t))^2,
\]

where \( N \) is the number of time points. To measure the improvement made by our proposed graphical method, we define relative MSE difference at a location \( l \) by:

\[
\text{relative MSE difference}(l) = \frac{\text{MSE}_{\text{TTLS}}(l) - \text{MSE}_{\text{GraphEM}}(l)}{\text{MSE}_{\text{TTLS}}(l)}.
\]

Although a small MSE indicates a good reconstruction, it is not immediately clear how small the MSE has to be for the reconstruction to be considered a “good reconstruction”. A useful approach is to
compare the MSE of a given reconstruction to that of a reconstruction that is equal to a constant value over time (“constant reconstruction”). The reduction of error (RE) compares the MSE of a given reconstruction to a constant reconstruction, where the specified constant is defined to be the mean of the field over the calibration period:

\[
RE(l) = 1 - \frac{\text{MSE}(\hat{T}(l))}{\text{MSE}(\overline{T}_c(l))}
\]

where \(\overline{T}_c(l)\) is the mean value of \(T(l, t)\) over the calibration period:

\[
\overline{T}_c(l) = \frac{1}{N_c} \sum_{t' \in \text{calibration}} T(l, t'),
\]

and \(N_c\) is the number of time points in the calibration period. Other metrics have also been used in the literature. The coefficient of efficiency (CE) compares the MSE of the reconstruction to a constant reconstruction equal to the mean of the field over the verification interval:

\[
CE(l) = 1 - \frac{\text{MSE}(\hat{T}(l))}{\text{MSE}(\overline{T}_v(l))},
\]

where

\[
\overline{T}_v(l) = \frac{1}{N_v} \sum_{t' \in \text{verification}} T(l, t'),
\]

and \(N_v\) is the number of time points in the verification period. In our numerical work (Section 4), the latest 150 years of the GCM simulation (1831-1980) have been used as the calibration period, and the rest of the simulation (850-1830) as the verification period.

Finally, the bias at point \(l\) is the difference between \(\hat{T}(l, \cdot)\) and \(T(l, \cdot)\) averaged over time. A perfect reconstruction would have a MSE of 0, a CE and a RE of 1 and a bias of 0. The closer to those values, the better the reconstruction.

3.3. Design of pseudoproxy experiments. In order to test the performance and the sensitivity of GraphEM to reconstruct temperature over the whole globe, we performed 50 reconstructions, each corresponding to a different noise realization \(\xi(l, t)\). The performance of GraphEM is then compared to the performance of RegEM-TTLS. Though several other methods have been proposed within RegEM, we focus on TTLS as it is currently widely used in paleoclimate applications (see e.g Mann et al., 2008, 2009; Emile-Geay et al., 2013a,b). The truncation parameter was set to 5, but the results show little sensitivity to this choice.

4. Results. In order to study the performance of GraphEM, reconstructions were performed using both the \(\ell_1\) method and the neighborhood graph method (Section 2.2). For illustration purposes, in Sections 4.1 and 4.2, we present detailed results for the neighborhood graph method with a cutoff radius of 800 km, as suggested by cross-validation (see Figure 3). Verification statistics for other cutoff radii and for the \(\ell_1\) method are also provided in Tables 1 and 2.

4.1. Spatial reconstructions. We begin by studying the performance of GraphEM in space. Figure 7a displays the average relative MSE improvement for the 50 reconstructions, and shows that the improvement can be substantial when using GraphEM. Figure 7a also shows that this improvement is positive in almost every region, and is particularly significant over the Central and Northern Pacific ocean. Since no proxies are available in these regions, the reconstruction seems to strongly benefit from
the dimension reduction brought about by graphical modeling. Figure 7b displays the distribution of the percentage improvement in MSE. Again, we see that the improvement is positive for almost every locations with an average relative MSE improvement of about 43%, whereas improvements as large as 80% are recorded in certain regions.

Fig 7: Relative MSE improvement and relative MSE improvement distribution for SNR = 0.5.

Figures 8a and 8b display the CE statistics (averaged over the 50 noise realizations) for TTLS and GraphEM respectively. Again, in many regions, GraphEM leads to substantial improvements, particularly where the skill was very poor with TTLS. In particular, we note that the improvement in CE are most notable for land masses where there is more data. The different precision metrics averaged over space (for the unsmoothed reconstruction) are presented in Table 1 along with their standard deviation computed using the 50 reconstructions. This table confirms once more that GraphEM performs better spatially and is more stable than RegEM-TTLS.

| Method                        | MSE   | RE      | CE      | Bias   |
|-------------------------------|-------|---------|---------|--------|
| GraphEM (0.5% target sparsity) | 0.42  | 0.36    | 0.15    | 0.11   |
| GraphEM (0.7% target sparsity)| 0.41  | 0.36    | 0.16    | 0.11   |
| GraphEM (800 km radius)       | 0.39  | 0.39    | 0.19    | 0.09   |
| GraphEM (1000 km radius)      | 0.40  | 0.38    | 0.18    | 0.09   |
| GraphEM (1200 km radius)      | 0.41  | 0.36    | 0.16    | 0.09   |
| RegEM-TTLS                    | 0.84  | -0.24   | -0.61   | 0.03   |

Table 1: Mean (and standard deviation) of the precision metrics averaged over space for the global reconstructions

Finally, we note that although the results presented in Table 1 are quite similar for the different methods, the neighborhood graphs seem to perform slightly better than the $\ell_1$ method. They could therefore potentially be very useful when working with noisy data for which discovering the structure of the field from the data could be very difficult.

4.2. Global mean reconstructions. The spatial reconstructions given by TTLS and GraphEM can also be averaged over space to obtain (area-weighted) global mean reconstructions. Figure 9 displays a 95% deviation band (constructed using the 50 reconstructions) for the global mean temperature series reconstructed with RegEM-TTLS and GraphEM. A 20 year low-pass filter has been applied to the time series for illustration and interpretation purposes. The mean width of the deviation interval for GraphEM and RegEM-TTLS are 0.25 and 0.66 respectively.
4.3. Uncertainty quantification. Section 4.1 demonstrates the ability of GraphEM to reduce the uncertainties in paleoclimate reconstructions. The stability of both methods was studied via an ensemble of pseudoproxies differing only in their noise realizations. This technique provides a useful way to test the robustness of the reconstructions. However, in practice, it is necessary to obtain an estimate of the uncertainties internally (see e.g. Li, Nychka and Ammann (2010)). We therefore produce prediction intervals for both RegEM and GraphEM using a non-parametric block bootstrap method (Liu,
1988). The technique is described in Algorithm 1 below, and is illustrated for the global reconstruction of Section 4.1. Using the reconstruction \( \hat{X}_1, \ldots, \hat{X}_N \) provided by the non-parametric bootstrap, we estimate a 95% prediction interval for each reconstructed mean by computing the 2.5th and 97.5th percentiles of the empirical distribution. From the simulation, we observed that the bootstrap tends to give slightly smaller coverage than in Figure 9. Two natural ways to address this discrepancy are 1) to modify the band width to obtain the right coverage, or 2) to inflate the variance of the reconstructed values in the bootstrap (see Li, Nychka and Ammann (2010); Janson and Rajaratnam (2014) for details). Figure 10 displays the average global mean reconstruction for RegEM-TTLS and GraphEM where the band width has been adjusted by computing the mean ratio of the band width obtained in the experiment performed in Section 4.2, and the band width obtained from the bootstrap. The uncertainties have been estimated with a blocksize of 2 (other values for the blocksize were also tried, but did not produce significantly different results). Once again, the estimated uncertainties for GraphEM are significantly smaller than those of RegEM-TTLS.

5. Characteristics of paleoclimate Markov random fields. Our results demonstrate that the GraphEM approach produces substantial improvements in comparison to RegEM-TTLS almost
uniformly over space. This section takes an in-depth look at characteristics of the resulting paleo-climatic Markov random fields with the goal of understanding a) whether the GraphEM approach is indeed achieving its original aim of parameter reduction and b) whether the parameter reduction provably leads to better statistical reconstructions.

We first illustrate the achieved parameter reduction when the graph is estimated from the data. Figure 5 shows that a point is generally only connected to a few immediate neighbors, although the graph can display some far away connections (which may or may not represent geophysical relations). The main message is that the number of neighbors is relatively few compared to one that would be present if a full precision matrix were to be used. Figure 11a displays the number of temperature neighbors of each temperature point (SNR = 0.5). Since there are in total 1732 grid points, the connectivity of each point is very small compared to a full graph. Figure 11b provides the distribution of the number of temperature neighbors of each temperature point (also referred to as degree distribution). It demonstrates that the connectivity percentage (defined as # of neighbors / # of gridpoints) is on average only 3% (as compared to being fully connected to all temperature points in the case when a complete graph is used).

The same analysis can be repeated in order to understand the number of proxies each temperature point is connected to (Fig. 12). Again, it is evident that the GraphEM approach yields a parsimonious graph structure in the cross correlation matrix between temperature points and the proxies.

Figure 13 displays the distribution of the number of proxy neighbors when no noise has been added to the temperature time series when generating pseudoproxies (SNR = \( \infty \)), as compared to the
D. GUILLOT, B. RAJARATNAM, AND J. EMILE-GEAY

![Graph showing distribution of number of proxy neighbors](image)

Fig 13: Distribution of the number of proxy neighbors of each temperature point.

typical noise case that has been studied thus far (SNR = 0.5). This comparison shows an increase in neighbors/degree when noise is added (see also Table 3). Many proxy neighbors in the SNR = 0.5 case have thus been added to the graph only because of random noise, instead of the presence of a real temperature/proxy relationship. The potential for glasso to detect spurious relations in the presence of noise is to be expected (Banerjee, El Ghaoui and d’Aspremont (2008)). This problem may be mitigated by adding further constraints on the estimated graph and will be explored in future work. If locality is a prime concern, neighborhood graphs offer a natural solution.

| SNR | Mean | Median | Std |
|-----|------|--------|-----|
| ∞   | 3.10 | 3.00   | 2.13|
| 0.5 | 9.67 | 9.00   | 4.49|

Table 3

Mean, median and standard deviation of the number of proxy neighbors

On to question b): does the sparsity translate to improvements in paleoclimate reconstructions? Fig. 14 displays the improvements given by GraphEM (as compared to RegEM-TTLS) at different temperature points vs. their connectivity (number of neighbors) in the corresponding graph. The figure indicates that percentage improvement increases as the number of neighbors decreases. To formally test if this relationship is significant, as a first step, both the standard Pearson’s product moment and Spearman’s rank correlation coefficients (and their respective p-values) between % improvement and number of neighbors have been calculated (see Table 4) (quadratic and higher terms were also computed, but were not found to be significant). The results confirm that there are substantial gains in percentage improvement over RegEM-TTLS for points which have fewer neighbors, thus confirming that a sparse graphical structure in the covariance matrix leads to more accurate estimates of the field. Note that this is to be expected as fewer connections avoids over-fitting and leads to an improvement in out-of-sample MSE.

| Type  | Correlation | p-value       |
|-------|-------------|---------------|
| Spearman | -0.39       | $< 10^{-14}$  |
| Pearson | -0.36       | $< 10^{-14}$  |

Table 4

Correlation and p-values between the improvement and the sparsity of the graph

The small p-values displayed in Table 4 suggest a clear and highly significant relationship between the relative MSE improvement and the sparsity of the graph. Nevertheless, care must be taken when computing p-values due to inherent spatial correlations. In order to obtain an estimate of the distribu-
tion of the correlation coefficient, while taking into account the spatial correlation, a block bootstrap technique Efron and Tibshirani (1993); Lahiri (2003); Léger, Romano and Politis (1992); Wilks (1997) was employed. More specifically, bootstrap samples were constructed from the data displayed in Figure 14, where resampling units are made up of geographically contiguous blocks based on a regular $30^\circ \times 30^\circ$ grid.

Figure 15 displays the density given by the block bootstrap approach for the Pearson product moment correlation coefficient. A highly statistically significant relationship between the relative MSE improvement and number of neighbors is retained, even when spatial correlation is taken into account (a 95% confidence interval is given by $[-0.43, -0.28]$).

The analysis above establishes that the percentage improvement over RegEM-TTLS is greater for points which have lower connectivity (i.e., sparse in terms of degree/connectivity). A related question is to determine whether the above relationship is uniform over space, i.e., are there regions where the improvement over RegEM-TTLS is even greater than what one would expect when a linear relationship between sparsity and percentage improvement is already accounted for. An equivalent way to frame this question is to determine if percentage improvement over RegEM-TTLS is related to space, even after controlling for sparsity. In order to establish this we looked at the difference between the observed percentage improvement and the expected one when a linear relationship is fitted. Figure 16 illustrates
that in certain regions the improvements are even higher than those suggested by a simple linear relationship with sparsity - thus indicating that the main message from our investigation is not as simple. Sparsity alone does not seem to explain improvements in the reconstruction. Local spatial characteristics also seem to have an effect.

Fig 16: Difference between the observed and expected percentage relative MSE improvement, after accounting for a linear relationship between % relative MSE improvement and number of temperature neighbors.

A follow up question on the analysis above is to determine if percentage improvement over RegEM-TTLS at a spatial point is related to data availability at that particular point. A closer examination of Figure 7a provides some compelling evidence that the magnitude of the percentage improvement appears to be even greater at certain spatial points that are in the open ocean, and also at points which are farther away from proxy data. In particular, vast swaths of the entire central and northern Pacific stretching from East Asia to North and central America display significantly higher improvements in MSE. The same appears to be true for parts of the southern Atlantic. Improvements over the Indian ocean however tend to be modest. Thus there appears to be evidence that the GraphEM approach may be taking advantage of teleconnections (see e.g. Liu and Alexander (2007)), and could be potentially very useful for paleo reconstructions over the oceans, where relatively few observations are available.

6. Concluding remarks. The main objective of the paper was to explore the efficacy of recent advances in the theory of graphical models and high-dimensional inference for statistical paleoclimate reconstructions. Markov random fields provide a sparse representation of the precision matrix of spatial fields, and thus achieve the dimension reduction that is often necessary in high-dimensional settings. At the same time, they can readily represent the spatial heterogeneities of geophysical fields (e.g. land/ocean contrasts, topographical boundaries, teleconnection patterns), which would in general be difficult with parametric (e.g. Matérn family) covariance functions. This combination of flexibility and parsimony presents a major advantage over previous approaches. The approach was also subsequently tested on simulated data within a geophysically-realistic context. We also proposed a block bootstrap method to internally estimate the uncertainties in the reconstructions performed using GraphEM and RegEM-TTLS.

Our experiments show that the GraphEM approach gives consistently better reconstructions than the frequently used RegEM-TTLS (see e.g Mann et al., 2008, 2009) almost uniformly over space. We show that Gaussian Markov random fields yield demonstrably improved estimates of the underlying spatio-temporal process, which we tied to the sparsity of the estimated covariance model. In particular, the improvement brought about by GMRFs tends to be more pronounced at those grid points which are modeled in a sparse way (in terms of the connectivity/degree of graph). These spatial points also tend to be located far away from proxy sites, thus underscoring the need for, and benefit of using $\ell_1$ based sparse covariance modeling in such problems. Improvements are particularly notable in sparsely-
sampled ocean regions – a very useful feature since the high cost of data collection in such regions leaves little hope of covering these gaps in the immediate future.

Although sparsity and distance to proxy data are the source of significant improvements, they are not the whole story: the pattern of teleconnections present in the climate model clearly led to some idiosyncrasies in the pattern of MSE reduction, which we do not expect to carry over in other temperature fields. Further work using different numerical experiments will be necessary to establish to full generality of these results.

A caveat of the graphical lasso is the tendency to sometimes detect spurious edges in the graph; that is, to detect relationships that arise from the presence of noise, instead of physical links between the temperature field and the proxies (or pseudoproxies) that derive from it. This is to be expected due to the signal to noise relationship in the data, and is inherent in all statistical and signal processing recovery techniques. Several approaches are currently under way to further remedy this problem, and to ensure that the graphs selected by the graphical lasso retain a high degree of locality, even in the presence of noisy climate proxies. An alternative method consists of using a neighborhood graph, i.e., a graph where geographically close neighbors are connected. As we demonstrate in our simulations, neighborhood graphs perform well and can be used in situations where there is less hope of discovering the graphical structure of the field from the data.

Finally, and although we were primarily motivated by paleoclimate applications and the use of the EM algorithm in this context, it is worth pointing out that graphical models also applicable within Bayesian CFR methods (e.g. Tingley and Huybers, 2010b,a) and beyond the confines of climate science. GraphEM as described here provides a useful addition to the RegEM framework, one that will be applicable to any high-dimensional imputation problem, and one that can be used in tandem with other $\ell_2$ regularization approaches, especially data-adaptive ones. Future work will extend the use of Gaussian Markov random fields as process models for geophysical fields, in tandem with hierarchical models.

References.

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

Annan, J. D. and Hargreaves, J. C. (2012). Identification of climatic state with limited proxy data. Clim. Past 8 1141–1151.

Banerjee, O., El Ghaoui, L. and d’Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. J. Mach. Learn. Res. 9 485–516.

Bradley, R. S. (1996). Are there optimum sites for global paleotemperature reconstruction? NATO ASI 41 Climate variations and forcing mechanisms of the last 2000 years. 603-624. SPRINGER.

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

Bürger, G. (2007). On the verification of climate reconstructions. Climate of the Past 3 397–409.

Carroll, R. J. and Ruppert, D. (1996). The use and misuse of orthogonal regression in linear errors-in-variables models. The American Statistician 50 1-6.

Christiansen, B. (2010). Reconstructing the NH mean temperature: can underestimation of trends and variability be avoided? Journal of Climate 24 674–692.

Christiansen, B. (2013). Straight line fitting and predictions: on a marginal likelihood approach to linear regression and errors-in-variables models. Journal of Climate 27 2014–2031.

Christiansen, B. and Ljungqvist, F. C. (2011). Reconstruction of the extratropical NH mean temperature over the last millennium with a method that preserves low-frequency variability. Journal of Climate 24 6013–6034.

Christiansen, B., Schmidt, T. and Thijss, P. (2009). A surrogate ensemble study of climate reconstruction methods: stochasticity and robustness. J. Clim. 22 951-976.

Cook, E. R., Meko, D. M., Stahle, D. W. and Cleaveland, M. K. (1999). Drought reconstructions for the continental United States. J. Clim. 12 1145–1162.

National Research Council (2006). Surface temperature reconstructions for the last 2000 years. The National Academies Press, Washington, D.C.
Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Stat.***21** 1272-1317.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B***39** 1–38. With discussion.

Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Monographs on Statistics and Applied Probability **57**. Chapman and Hall, New York.

Emile-Geay, J., Cobb, K. M., Mann, M. E. and Wittenberg, A. T. (2013a). Estimating central equatorial Pacific SST variability over the past millennium. Part 1: methodology and validation. *J. Clim.***26** 2302–2328.

Emile-Geay, J., Cobb, K. M., Mann, M. E. and Wittenberg, A. T. (2013b). Estimating central equatorial Pacific SST variability over the past millennium. Part 2: reconstructions and implications. *J. Clim.***26** 2329–2352.

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

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostat***9** 432–441.

Frost, C. and Thompson, S. G. (2000). Correcting for regression dilution bias: comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society: Series A (Statistics in Society)***163** 173–189.

Golub, G. H. and Van Loan, C. F. (1980). An analysis of the total least squares problem. *SIAM J. Numer. Anal.***17** 883–893.

Guillot, D., Rajaratnam, B., Rolfs, B., Wong, I. and Maleki, A. (2012). Iterative thresholding algorithm for sparse inverse covariance estimation. In *Advances in Neural Information Processing Systems* **25** 1583–1591.

Hanke, M. and Hansen, P. C. (1993). Regularization methods for large-scale problems. *Surveys Math. Indust.* **3** 253–315.

Hastie, T., Tibshirani, R. and Friedman, J. (2008). *The elements of statistical learning: data mining, inference and prediction*, 2 ed. Springer.

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

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

Jansen, E., Overpeck, J., Brietta, K. R., Duplessy, J. C., Joos, F., Masson-Delmotte, V., Olago, D., Otto-Bliesner, B., Peltier, W. R., Rahmstorf, S., Ramesh, R., Raynaud, D., Rind, D., Solomina, O., Villalba, R. and Zhang, D. (2007). *Climate change 2007: The physical science basis. Contribution of working group I to the fourth assessment report of the Intergovernmental Panel on Climate Change* Palaeoclimate. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Janson, L. and Rajaratnam, B. (2014). A methodology for robust multiproxy paleoclimate reconstructions and modeling of temperature conditional quantiles. *Journal of the American Statistical Association***109** 63-77.

Jones, P. D., Brietta, K. R., Osborn, T. J., Lough, J. M., van Ommen, T. D., Vintner, B. M., Luterbacher, J., Wahl, E. R., Zwieter, F. W., Mann, M. E., Schmidt, G. A., Ammann, C. M., Buckley, B. M., Cobb, K. M., Esper, J., Gose, H., Graham, N., Jansen, E., Kiefer, T., Kull, C., Kuttel, M., Mosley-Thompson, E., Overpeck, J. T., Riedwyl, N., Schulz, M., Tudhope, A. W., Villalba, R., Wanner, H., Wolff, E. and Xoplaki, E. (2009). High-resolution palaeoclimatology of the last millennium: a review of current status and future prospects. *The Holocene***19** 3-49.

Jui Hsieh, C., Dhillon, I. S., Ravikumar, P. K. and Sustik, M. A. (2011). Sparse inverse covariance matrix estimation using quadratic approximation. In *Advances in Neural Information Processing Systems* **24** (J. Shave-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira and K. Q. Weinberger, eds.) 2330–2338. Curran Associates, Inc.

Lahiri, S. N. (2003). *Resampling methods for dependent data*. Springer Series in Statistics. Springer-Verlag, New York.

Lauritzen, S. L. (1996). *Graphical models*. Oxford Statistical Science Series **17**. The Clarendon Press Oxford University Press, New York. Oxford Science Publications.

Léger, C., Romano, J. P. and Politis, D. N. (1992). Bootstrap technology and applications. *Technometrics***34** 378–398.

Letac, G. and Massam, H. (2007). Wishart distributions for decomposable graphs. *Annals of Statistics***35**.

Li, B., Nychka, D. W. and Ammann, C. M. (2010). The value of multi-proxy reconstruction of past climate. *J. Amer. Statist. Assoc.***105** 883–911.

Li, B. and Smerdon, J. E. (2012). Defining spatial comparison metrics for evaluation of palaeoclimatic field reconstructions of the Common Era. *Environmetrics***23** 394–406.

Lin, S. and Perlman, M. (1985). A Monte-Carlo comparison of four estimators of a covariance matrix. *Multivariate Analysis* **VI** 411–429.

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

Liu, R. Y. (1988). Bootstrap procedures under some non-I.I.D. models. *Ann. Statist.***16** 1345–1741.

Liu, Z. and Alexander, M. A. (2007). Atmospheric bridge, oceanic tunnel, and global climatic teleconnections. *Rev.
MANN, M. E., BRADLEY, R. S. and HUGHES, M. K. (1998). Global-scale temperature patterns and climate forcing over the past six centuries. Nature 392 779-787.

MANN, M. E. and RUTHERFORD, S. (2002). Climate reconstruction using ‘pseudoproxies’. Geophys. Res. Lett. 29 139-1.

MANN, M. E., RUTHERFORD, S., WAHL, E. and AMMANN, C. (2005). Testing the fidelity of methods used in proxy-based reconstructions of past climate. Journal of Climate 18 4097-4107.

MANN, M. E., RUTHERFORD, S., WAHL, E. and AMMANN, C. (2007a). Robustness of proxy-based climate field reconstruction methods. Journal of Geophysical Research (Atmospheres) 112 D12109+.

MANN, M. E., RUTHERFORD, S., WAHL, E. and AMMANN, C. (2007b). Reply to comments on “Testing the fidelity of methods used in proxy-based reconstructions of past climate” by Smerdon and Kaplan. J. Clim. 20 5671-5674.

MANN, M. E., ZHANG, Z., HUGHES, M. K., BRADLEY, R. S., MILLER, S. K., RUTHERFORD, S. and NI, F. (2008). Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia. Proceedings of the National Academy of Sciences 105 13252-13257.

MANN, M. E., ZHANG, S., RUTHERFORD, S., BRADLEY, R. S., HUGHES, M. K., SHINDELL, D., AMMANN, C., FALUVEGI, G. and NI, F. (2009). Global signatures and dynamical origins of the little ice age and medieval climate anomaly. Science 326 1256-1260.

McLACHLAN, G. J. and KRISHNAN, T. (2008). The EM algorithm and extensions, second ed. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.

PAUL, D. (2007). Asymptotics of the leading sample eigenvalues for a spiked covariance model. Stat. Sinica. 17 1617-1642.

RAJARATNAM, B., MASSAM, H. and CARVALHO, C. M. (2008). Flexible covariance estimation in graphical Gaussian models. Ann. Statist. 36 2818-2849.

RIEDWYL, N., KÜTTEL, M., LUTERBACHER, J. and WANNER, H. (2008). Comparison of climate field reconstruction techniques: application to Europe. Clim. Dyn.

RUTHERFORD, S., MANN, M. E., OSBORN, T. J., BRADLEY, R. S., BRIFFA, K. R., HUGHES, M. K. and JONES, P. D. (2005). Proxy-based northern hemisphere surface temperature reconstructions: sensitivity to method, predictor network, target season, and target domain. J. Clim. 18 2308-2329.

SCHNEIDER, T. (2001). Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. J. Clim. 14 853-871.

SMERDON, J. E. (2011). Climate models as a test bed for climate reconstruction methods: pseudoproxy experiments. WIREs Clim Change.

SMERDON, J. E. and KAPLAN, A. (2007). Comments on ”Testing the fidelity of methods used in proxy-based reconstructions of past climate”: the role of the standardization interval. J. Clim. 5666-5670.

SMERDON, J. E., KAPLAN, A., CHANG, D. and EVANS, M. N. (2010). A pseudoproxy evaluation of the CCA and RegEM methods for reconstructing climate fields of the last millennium. Journal of Climate 23 4856-4880.

SMERDON, J. E., KAPLAN, A., ZORITA, E., GONZÁLEZ-ROUCO, J. F. and EVANS, M. N. (2011). Spatial performance of four climate field reconstruction methods targeting the Common Era. Geophys. Res. Lett. 38.

STEIGER, N. J., HAKIM, G. J., STEIG, E. J., BATTisti, D. S. and ROE, G. H. (2013). Assimilation of time-averaged pseudoproxies for climate reconstruction. Journal of Climate 27 426-441.

STEIN, C. (1975). Estimation of a covariance matrix. In Rietz Lecture. 39th Annual Meeting IMS, Atlanta, GA.

STEIN, C. (1986). Lectures on the theory of estimation of many parameters. Journal of Mathematical Sciences 34 1373-1403.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58 267-288.

TIKHONOV, A. N. and ARSENNY, V. Y. (1977). Solution of ill-posed problems. Scripta Series in Mathematics. V. H. Winston and Sons, Washington.

TINGLEY, M. P. and HUYBERS, P. (2010a). A Bayesian algorithm for reconstructing climate anomalies in space and time. Part 2: comparison with the regularized expectation-maximization algorithm. J. Clim. 23 2782-2800.

TINGLEY, M. P. and HUYBERS, P. (2010b). A Bayesian algorithm for reconstructing climate anomalies in space and time. Part 1: development and applications to paleoclimatic reconstruction problems. J. Clim. 23 2759-2781.

TINGLEY, M. P. and LI, B. (2012). Comments on “Reconstructing the NH mean temperature: can underestimation of trends and variability be avoided?”. Journal of Climate 25 3441-3446.

TINGLEY, M. P., CRAIGMILE, P. F., HARAN, M., LI, B., MANNSHARDT, E. and RAJARATNAM, B. (2012). Piecing together the past: statistical insights into paleoclimatic reconstructions. Quaternary Science Reviews 35 1 - 22.

VAN HUFFEL, S. and VANDERWAAL., J. (1991). The Total Least Squares Problem: Computational Aspects and Analysis. Frontiers in Applied Mathematics 9. SIAM, Philadelphia, PA.

VON STORCH, H., ZORITA, E., JONES, M. N., DIMITRIEV, Y., GONZÁLEZ-ROUCO, F. and TETT, S. F. B. (2004). Reconstructing past climate from noisy data. Science 306 679-682.

WANG, J., EMILE-GEAY, J., GUILLOT, D., SMERDON, J. E. and RAJARATNAM, B. (2014). Evaluating climate field reconstruction techniques using improved emulations of real-world conditions. Climate of the Past 10 1–19.
APPENDIX A: DESCRIPTION OF THE GRAPHEM ALGORITHM

Algorithm 2 The graphical EM algorithm (GraphEM)

**Input**: Incomplete $n \times p$ matrix $X$, graph $G$

1: Initialize $X^{(0)}$ by replacing the missing values in $X$ by the sample mean of each variable over the instrumental period;

2: Compute initial estimates $\mu^{(0)}$ and $\Sigma^{(0)}$ of $\mu$ and $\Sigma$ by computing the sample mean and sample covariance of $X^{(0)}$;

3: Initialize $i \leftarrow 0$;

4: Initialize $\Sigma^{G(0)} = \Sigma^{(0)}$;

5: repeat

6: Compute $X^{(i+1)}$ by performing a linear regression of the missing values on the available ones for each row of $X$, using the current estimate $\mu^{(i)}$ of $\mu$ and the current graphical estimate $\Sigma^{G(i)}$ of $\Sigma$ (see 1.1);

7: Compute $\mu^{(i+1)}$ by computing the sample mean of $X^{(i+1)}$;

8: Compute $\Sigma^{(i+1)}$ as in (1.3);

9: Compute the new graphical estimate $\Sigma^{G(i+1)}$ by solving (1.6) with $S = \Sigma^{(i+1)}$, i.e.,

$$(A.1) \quad \Sigma^{G(i+1)} = \arg\max_{\Sigma = \Omega^{-1} > 0} \log \det \Omega - \text{tr}(\Sigma^{(i+1)} \Omega);$$

10: $i \leftarrow i + 1$;

11: until convergence

**Output**: Completed matrix $\hat{X}$, estimate $\hat{\mu}$ of $\mu$, estimate $\hat{\Sigma}$ of $\Sigma$

APPENDIX B: DERIVATION OF THE GRAPH-EM ALGORITHM

We follow the notation in Little and Rubin Little and Rubin (2002). The complete data belongs to a regular exponential family given by a Gaussian Markov random field with graph $G = (V, E)$ (as compared to a complete model in the classical EM algorithm). The sufficient statistics are given by:

$$S = \left( \sum_{i=1}^{n} y_{ij}, j = 1, \ldots, k; \sum_{i=1}^{n} y_{ij}y_{ik} \quad \text{with} \quad (j,k) \in E \right).$$

Let $\theta^{(t)} = (\mu^{(t)}, \Sigma^{(t)})$ denote the current estimate of the parameters. The E-step is given as follows:

$$(B.2) \quad E \left[ y_{ij} \mid Y_{\text{obs}}, \theta^{(t)} \right] = \sum_{i=1}^{n} y_{ij}^{(t)} \quad j = 1, \ldots, k$$

and

$$(B.3) \quad E \left[ y_{ij}y_{ik} \mid Y_{\text{obs}}, \theta^{(t)} \right] = \sum_{i=1}^{n} (y_{ij}^{(t)}y_{ik}^{(t)} + c_{jki}^{(t)}) \quad (j,k) \in E$$

with

$$(B.4) \quad y_{ij}^{(t)} = \begin{cases} y_{ij} & \text{when } y_{ij} \text{ is observed} \\ E \left[ y_{ij} \mid y_{\text{obs},i}, \theta^{(t)} \right] & \text{when } y_{ij} \text{ is missing} \end{cases}$$
and

\begin{equation}
\begin{cases}
0 & \text{if at least one of the } y_{ij} \text{ or } y_{ik} \text{ is observed} \\
0 & \text{if } j \perp_G k|\text{obs},i \\
\text{Cov}\left[y_{ij}, y_{ik}\mid y_{\text{obs},i}, \theta^{(t)}\right] & \text{if both } y_{ij} \text{ and } y_{ik} \text{ are missing and } j \not\perp_G k|\text{obs},i,
\end{cases}
\end{equation}

where \( j \perp_G k \) means that \( j \) and \( k \) are separated in the graph \( G \) (see e.g. (Lauritzen, 1996, Example 3.2)). At a first glance, it would appear as if there is little difference between the treatment in the graphical vs. the complete case. A closer look reveals that there are some notable differences, the first being in the calculation of the sufficient statistics. Second, note that the definition of \( y_{ij}^{(t)} \) and \( c_{jki}^{(t)} \) below are different: \( y_{ij}^{(t)} \) when \( y_{ij} \) is missing is given as follows:

\begin{equation}
E\left[y_{ij}\mid y_{\text{obs},i}, \theta^{(t)}\right] = \mu_{ij}^{(t)} + \left(\Sigma_{\text{obs},i}^{G}\right)^{(t)} \left(\Sigma_{\text{obs},\text{obs}}^{G}\right)^{(t)} \left(y_{\text{obs},i} - \mu_{\text{obs}}^{(t)}\right),
\end{equation}

where \( \left(\Sigma^{G}\right)^{(t)} \) corresponds to a graphical covariance matrix \( \Sigma \). When both \( y_{ij} \) and \( y_{ik} \) are missing and \( (j,k) \in E \),

\begin{equation}
\text{Cov}\left[y_{ij}y_{ik}\mid y_{\text{obs},i}, \theta^{(t)}\right] = \left(\Sigma_{jk}^{G}\right)^{(t)} - \left(\Sigma_{\{j,k\},\text{obs}}^{G}\right)^{(t)} \left(\Sigma_{\text{obs},\text{obs}}^{G}\right)^{(t)} \left(\Sigma_{\text{obs},\{j,k\}}^{G}\right)^{(t)}.
\end{equation}

Note however that \( \text{Cov}\left[y_{ij}y_{ik}\mid y_{\text{obs},i}, \theta^{(t)}\right] = \Sigma_{jk|\text{obs},i}^{(t)} \). Thus,

\begin{equation}
c_{jki}^{(t)} = \Sigma_{jk|\text{obs},i}^{(t)} = 0 \quad \text{if } j \perp_G k|\text{obs},i.
\end{equation}

The M-step in the GraphEM algorithm therefore consists of using the sufficient statistics for the complete data derived in (B.2) and (B.3) to determine the graphical mle. In particular, the estimate of the mean parameter is given by the sample mean and estimate of the graphical covariance is given in equation (A.1).