Quaternary Science Reviews 35 (2012) 1–22

Contents lists available at SciVerse ScienceDirect

Quaternary Science Reviews

journal homepage: www.elsevier.com/locate/quascirev

Invited Review

Piecing together the past: statistical insights into paleoclimatic reconstructions

Martin P. Tingley [a,b,c,*], Peter F. Craigmile [d], Murali Haran [b,e], Bo Li [f], Elizabeth Mannshardt [b,d,g], Bala Rajaratnam [b,h,i]

[a] National Center for Atmospheric Research, USA
[b] Statistical and Applied Mathematical Sciences Institute, USA
[c] Department of Earth and Planetary Sciences, Harvard University, USA
[d] Department of Statistics, The Ohio State University, USA
[e] Department of Statistics, Pennsylvania State University, USA
[f] Department of Statistics, Purdue University, USA
[g] Department of Statistical Sciences, Duke University, USA
[h] Department of Statistics, Stanford University, USA
[i] Department of Environmental Earth System Science & The Woods Institute for the Environment, Stanford University, USA

ARTICLE INFO

Article history:
Received 15 June 2011
Received in revised form
21 December 2011
Accepted 11 January 2012
Available online 10 February 2012

Keywords:
Paleoclimate
Bayesian methods
Hierarchical modeling
Spatial modeling
Space–time modeling

ABSTRACT

Reconstructing a climate process in both space and time from incomplete instrumental and climate proxy time series is a problem with clear societal relevance that poses both scientific and statistical challenges. These challenges, along with the interdisciplinary nature of the reconstruction problem, point to the need for greater cooperation between the earth science and statistics communities – a sentiment echoed in recent parliamentary reports.

As a step in this direction, it is prudent to formalize what is meant by the paleoclimate reconstruction problem using the language and tools of modern statistics. This article considers the challenge of inferring, with uncertainties, a climate process through space and time from overlapping instrumental and climate sensitive proxy time series that are assumed to be well dated – an assumption that is likely only reasonable for certain proxies over at most the last few millennia. Within a unifying, hierarchical space–time modeling framework for this problem, the modeling assumptions made by a number of published methods can be understood as special cases, and the distinction between *modeling assumptions* and *analysis* or *inference* choices becomes more transparent.

The key aims of this article are to 1) establish a unifying modeling and notational framework for the paleoclimate reconstruction problem that is transparent to both the climate science and statistics communities; 2) describe how currently favored methods fit within this framework; 3) outline and distinguish between *scientific* and *statistical* challenges; 4) indicate how recent advances in the statistical modeling of large space–time data sets, as well as advances in statistical computation, can be brought to bear upon the problem; 5) offer, in broad strokes, some suggestions for model construction and how to perform the required statistical inference; and 6) identify issues that are important to both the climate science and applied statistics communities, and encourage greater collaboration between the two.

1. Introduction

This paper is a product of our participation in the 2009–2010 program on "Space–Time Analysis for Environmental Mapping, Epidemiology and Climate Change",[1] organized by the Statistical and Applied Mathematical Sciences Institute (SAMSI), an NSF sponsored research center in North Carolina. Our focus at SAMSI was on the statistical challenges surrounding the reconstruction of past climate from incomplete instrumental and proxy data sets, and part of the motivation for writing this piece stems from the various controversies surrounding the interpretation and assimilation of instrumental and proxy-temperature time series. Much of the controversy points to the potential benefits of greater collaboration between statisticians and paleoclimatologists in the analysis and interpretation of climate data, a sentiment that is echoed in the recent United Kingdom parliamentary report on the University of East Anglia's Climate Research Unit (CRU):

> "*We cannot help remarking that it is very surprising that research in an area that depends so heavily on statistical methods has not*

* Corresponding author. National Center for Atmospheric Research, 1850 Table Mesa Drive, Boulder, 80305, USA. Tel.: +1 6173299384.
E-mail address: tingley@fas.harvard.edu (M.P. Tingley).

[1] For more information, see www.samsi.info/programs/research-programs/past.

0277-3791/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.quascirev.2012.01.012

been carried out in close collaboration with professional statisticians. Indeed there would be mutual benefit if there were closer collaboration and interaction between CRU and a much wider scientific group outside the relatively small international circle of temperature specialists. "[2]

It seems pertinent that a group of statisticians interested in the climate reconstruction problem, in collaboration with a climate scientist, present both a formal description of the reconstruction problem and offer suggestions for how this field can be advanced via a reasoned use of modern statistics. We will not present a new reconstruction, or propose, test, or apply a specific analysis model. Instead, we provide a detailed presentation of hierarchical statistical models and describe how the different levels should be specified in the context of paleoclimatic reconstructions. More general reviews of climate reconstructions of the last few millennia can be found, for example, in the 2006 National Research Council report on the subject (NRC, 2006), or the Jones et al. (2009) review in *The Holocene*. Hughes and Ammann (2009) provide a broad overview of the state of paleoclimate reconstruction methods, and, as we do, offer suggestions on how to move forward. This article builds upon and provides the necessary background to implement the hierarchical models mentioned in Hughes and Ammann (2009).

Inferring past climate from raw observations of the natural world is a grand challenge. We focus on one particular aspect of the problem: given climate sensitive proxy time series that are assumed to be well dated, how should they be combined along with the instrumental record to arrive at estimates, with uncertainties, of a climate process through space and time? We consider the challenges involved in modeling a space–time process such as annual mean surface temperature anomalies, as well as the difficulties involved in inferring such a process from a number of different data sources, all of which are noisy and incomplete. It is our aim to clearly define the scope of the problem and the nature of the challenges, identify and describe the relevant statistical tools and techniques, and indicate how they can be used in particular applications. In addition, we describe how numerous published methods fit within the proposed hierarchical framework. Posing the paleoclimatic reconstruction problem in the language of modern statistics will help elucidate those areas in which statisticians have expertise that can be brought to bear upon this problem, and will encourage greater collaboration between the climate science and statistics communities.

The assumption that the proxy series are well dated is likely only reasonable for certain types of proxy over at most the last few millennia. The treatment of time-uncertain proxy time series is an active field of research (Haslett et al., 2006a; Auestad et al., 2008; Haam and Huybers, 2010), and becomes particularly important when considering proxy archives such as pollen and sea floor sediment cores that, in contrast with tree rings and ice cores, do not form laminations with a known frequency. Likewise, raw observations of proxy archives frequently undergo considerable processing before being put forth as a climate sensitive time series. For example, raw pollen counts or percentages are transformed via comparison with modern analogues (e.g., Haslett et al., 2006b), and some estimate of the biological growth effect must be removed from individual tree ring series before they are combined into a climate sensitive site chronology (e.g., Briffa et al., 1992; Melvin and Briffa, 2008; Schofield, in preparation). Recent work (e.g., Haslett et al., 2006b), has focused on forward-model based approaches to processing raw observations into climate sensitive series. This article will not focus on either time uncertainty or this processing of raw proxy observations into climate sensitive series, but we will provide brief comments on how progress on those problems can be incorporated into the framework outlined below.

It is important to recognize that we are not the first group of statisticians to become interested in this problem, and hopefully we are not the last. There have been numerous time series analyses of paleorecords in the statistics literature, such as Visser and Molenaar (1988); West (1997); Harvill and Ray (2006); and Haslett et al. (2006b). More recently, Li et al. (2010) present a hierarchical model and apply it to pseudoproxies derived from climate models, while Brynjarsdóttir and Berliner (2011) reconstruct surface temperatures using borehole temperature profiles. Likewise, several recent papers from the climate literature have proposed hierarchical models in the context of reconstructing past climate. Lee et al. (2008) propose a state-space or Kalman filter model for inferring large-scale spatial average temperatures, which we interpret as a hierarchical model (see Section 8.3). Lee et al. (2008) include estimates of climate forcing series in the inference model, and the specification of separate models for the target process and the data. In contrast, Tingley and Huybers (2010a,b) propose a simple hierarchical statistical model without forcings to infer a climate field in both space and time. While there are examples in the published literature of hierarchical models and Bayesian analysis applied to paleoclimate data (e.g., Haslett et al., 2006b; Li et al., 2010; Tingley and Huybers, 2010a), what has been lacking, until know, is a more general argument for and exposition of Bayesian hierarchical modeling for inferring past climate.

In Section 2, we introduce a representative subset of the data from Mann et al. (2008a) in order to illustrate the challenges posed by paleoclimatic and instrumental data, and to motivate the modeling approach we favor. We then present a general, hierarchical statistical space–time modeling framework appropriate for the reconstruction problem in Section 3. The key specifications of this class of models are the space–time structure of the target climate process, which we discuss in Section 4, and the relationships between the statistical processes characterizing the data sources and the target process, which we describe in Section 5. We then discuss issues regarding the observations in Section 6, including the influence of observational errors and the treatment of missing data. Performing inference on this class of space–time models is non-trivial and can be computationally intensive, and we provide suggestions on how to overcome these difficulties in Section 7. Within the hierarchical modeling framework, a number of published reconstructions methods can be interpreted as special cases, and thus our approach yields a unifying framework for paleoclimatic reconstructions. We discuss several commonly used methods in Section 8, and then close with some general remarks and discussion in Section 9.

2. A motivating data set

Mann et al. (2008a) present a reconstruction of hemispheric and global surface temperatures over the last two millenia using 1209 proxy time series (described in the supplement, Mann et al., 2008b) and the $5^\circ \times 5^\circ$ gridded surface temperature data product from the University of East Anglia's Climatic Research Unit (Brohan et al., 2006).[3] We illustrate a number of challenges posed by paleoclimate data by considering a subset of this data, geographically restricted to Northern North America and Greenland, and consisting of only the instrumental, tree ring density, tree ring width, ice core $\delta^{18}O$, and annual lake sediment varve thickness time series (Figs. 1 and 2).

There are a number of climate quantities that a researcher might wish to reconstruct from a data set of this sort, including time series of

[2] Taken from www.uea.ac.uk/mac/comm/media/press/CRUstatements/SAP.

[3] The proxy data is available at www.meteo.psu.edu/mann/supplements/MultiproxyMeans07/, and the instrumental data set at www.cru.uea.ac.uk/cru/data/temperature/.

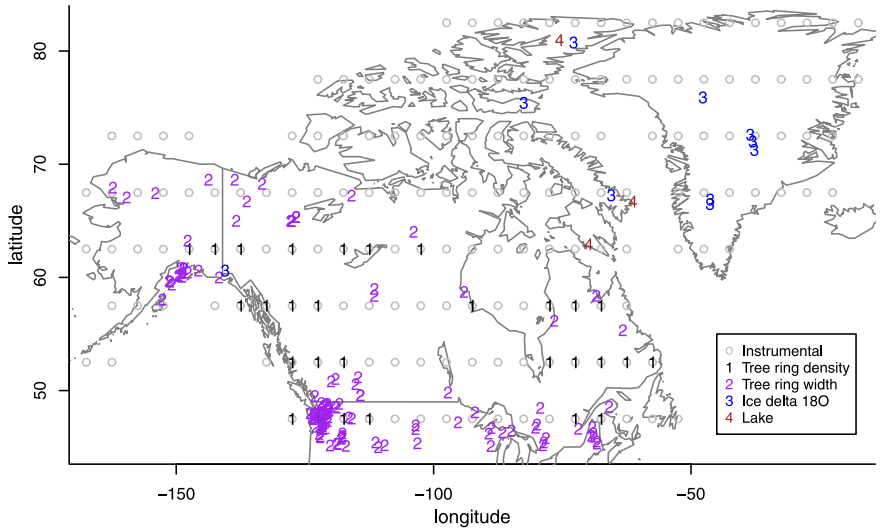

Fig. 1. A subset of the data used in Mann et al. (2008a). Circles indicate the centroids of the grid used to produce the annual, spatially-averaged, temperature anomaly time series from the instrumental record (Brohan et al., 2006). Numbers indicate the locations of, or the centroids of the regions represented by, the various proxy time series.

large-scale spatial averages of the climate field (e.g., Moberg et al., 2005; Lee et al., 2008; Mann et al., 2008a; Kaufman et al., 2009), or the spatial pattern of a climate variable as a function of time (e.g., Mann et al., 1998; Cook et al., 1999; Luterbacher et al., 2004). Alternatively, the goal may be to infer an index, such as El Niño, that reflects broad aspects of climate (e.g., Emile-Geay et al., submitted for publication-a, submitted for publication-b). We focus here and below on the reconstruction of climate processes that can be modeled as continuous in space and discrete in time – such as annual mean surface temperatures.

Broadly speaking, the challenges associated with reconstructing a space–time climate process fall into two categories. First, the climate system varies on the full spectrum of temporal and spatial scales, and displays complex spatial, temporal, and spatiotemporal covariance structures. For example, El Niño variability has a rich spatiotemporal signature in the surface temperature anomaly process, and reconstruction techniques should somehow account for such phenomena. Second, paleoclimatic reconstructions tend to combine a variety of different sources of data, each with particular characteristics. With respect to a representative subset of the Mann et al. (2008a) data set (see Figs. 1 and 2), we note the following:

- The different data sets likely reflect the target process averaged over different temporal scales, and likely have different functional relationships with the target process. First order autoregressive [AR(1)] fits to each data series reveal that the distribution of AR(1) coefficients for the instrumental and tree ring density series are similar, and generally smaller than those for the tree ring width series (Fig. 3). While the instrumental observations represent annual averages, the stronger temporal dependencies of the tree ring width series (Fig. 3) suggest a longer temporal averaging of the underlying climate process. Indeed, it is unclear if an AR(1) model is even appropriate for the tree ring width series. The distribution of the optimal order of AR(p) fits for the data series, according to the Bayesian Information Criterion (using the ARfit Matlab package of Neumaier and Schneider, 2001; Schneider and Neumaier, 2001) is on average about one for the instrumental and tree ring density series, but about two for the tree ring width series (Fig. 3). These results indicate that the time series dependencies are not the same for all data sources. In addition, the proxies may preferentially reflect the climate during a subset of the year, such as the growing season (tree-based proxies), or the season with the most precipitation (ice cores).
- While the instrumental and the tree ring density data sets represent averages over grid boxes, the $\delta^{18}O$, tree ring width, and varved lake sediment data sets correspond to observations at specific spatial locations. Different data types thus represent the target process on different spatial scales, and the locations of the proxy time series do not generally correspond to the centroids of the instrumental grid.
- The locations of the proxy data series, particularly the tree ring width records and the ice core records, are clustered in space.
- The number of observations available for each year decreases rapidly moving back in time (Fig. 2), from the data-dense instrumental period to less than 20 observations in 1400.

The modeling framework outlined in the following sections takes into account and reflects these features of the target climate process, and the instrumental and proxy data sources.

3. Hierarchical statistical models

The paleoclimate reconstruction problem involves inferring a target, latent[4] space–time climate process, such as surface temperature anomalies, conditional on the observed instrumental and proxy time series and other available covariates. The different data sources may have different uncertainties and different relationships with the target climate process, while each data source, as well as the target process, typically displays spatial and temporal dependencies. These properties of the data and climate system motivate a *hierarchical statistical approach* for the paleoclimate reconstruction problem.

We construct the probability model for the observations as a product of conditional distributions, each of which depends on the different instrumental and proxy observations, and various unobserved parameters. The unobserved parameters include scalars such as variance parameters, as well as a number of latent statistical processes – such as the target space–time climate process and measurement error-free data processes. Once the data are observed,

[4] The term *latent* is commonly used in statistical models to indicate a quantity that is unobserved.

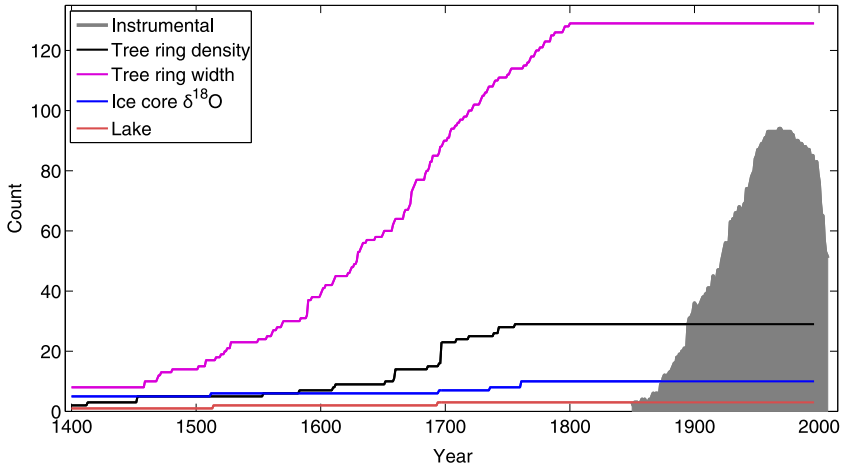

Fig. 2. The number of each type of observation as a function of time, for the subset of the Mann et al. (2008) data set depicted in Fig. 1.

this probability model, $f(\text{data}|\text{parameters})$, translates to a "likelihood" function that contains the information in the data about the unobserved parameters (including the statistical processes). Specification of the statistical model in this way allows for the construction of scientifically-driven space–time relationships between different components of the model (referred to as *modeling* choices), in isolation from the *analysis* choices – the particular tools and techniques used to infer the unknown statistical parameters in the model. As an example, assuming a linear relationship between two variables is a *modeling* assumption, while the decision to use least squares, method of moments, Bayesian inference, or some other tool to perform the statistical inference is an *analysis* choice.

This article will focus on Bayesian inference as the analysis choice. Given specifications of the prior distribution, $\pi(\text{parameters})$, for all unknowns parameters and the likelihood, $f(\text{data}|\text{parameters})$, the posterior distribution of the unknown parameters given the data is,

$$\pi(\text{parameters}|\text{data}) \propto f(\text{data}|\text{parameters})\pi(\text{parameters}). \quad (1)$$

Since the advent of Markov chain Monte Carlo (MCMC) methods in the 1980s, Bayesian data analysis methods have grown in popularity, mostly due to their ability to easily propagate measures of uncertainty in complicated scientific problems. There are now many examples in the earth and climate sciences; see, e.g., Berliner et al. (2000b); Wikle and Anderson (2003); Tebaldi et al. (2005); Song et al. (2007); Kopp et al. (2009); Li et al. (2010); Tingley and Huybers (2010a).

3.1. A general framework for the paleoclimate reconstruction problem

Let

$$\mathbf{Y} = \{Y(\mathbf{s},t) : \mathbf{s} \in \mathcal{D}, \quad t \in \mathcal{T}\} \quad (2)$$

denote the latent space–time climate process, where $\mathcal{D}$ designates the spatial and $\mathcal{T}$ the temporal domain of interest. Depending on the spatial coverage of the reconstruction, the spatial domain $\mathcal{D}$ can be discrete or continuous, usually in two or three dimensions, while the temporal domain $\mathcal{T}$ is usually a subset of the integers (for example, annual averages). For time-uncertain problems, we may instead consider $\mathcal{T}$ as a subset of the reals.

In terms of the data sources, we specify a separate model for each type or class of observation, such as ground-based thermometers, satellite observations, tree ring density observations, ice core δ^{18}O observations, and so forth. Let

$$\mathbf{Z}_{I,j} = \{Z_{I,j}(\mathbf{s},t) : \mathbf{s} \in \mathcal{D}_{I,j}, t \in \mathcal{T}_{I,j}\} \quad j = 1,\ldots,N_I, \quad (3)$$

denote the N_I different types of instrumental observation, where $\mathcal{D}_{I,j}$ and $\mathcal{T}_{I,j}$ denote the spatial and temporal domains, respectively, for the jth instrumental data type. These records need not be observed at the same set of spatial or temporal scales as the latent space–time climate process, and indeed each data source may be observed on a different scale. For example one instrumental record may be observed annually, whereas another may be recorded daily;

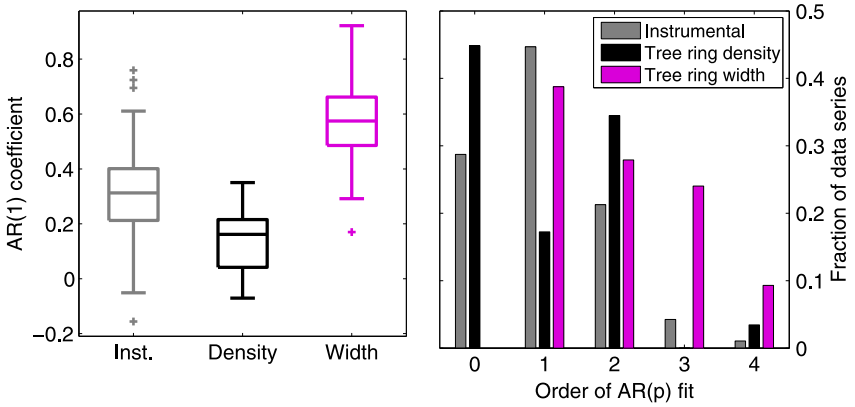

Fig. 3. Left panel: box plots of AR(1) coefficients inferred from the instrumental, tree ring density, and tree ring width time series depicted in Figs. 1 and 2. Results are not shown for the ice core and lake data sets due to the small number of series. Right panel: optimal order of AR(p) fit, according to the Bayesian Information Criterion (analysis performed with the ARFit Matlab package described in Neumaier and Schneider (2001) and Schneider and Neumaier (2001)).

another data product may be gridded, compared to one that refers to exact spatial points. Similarly, let

$$\mathbf{Z}_{P,k} = \{Z_{P,k}(\mathbf{s},t) : \mathbf{s}\in\mathcal{D}_{P,k}, t\in\mathcal{T}_{P,k}\}, k = 1,\ldots,N_P \quad (4)$$

denote the N_P different types of proxy records (for example, the spatially located tree ring density series, tree ring width series, and ice core series in Fig. 1), with $\mathcal{D}_{P,k}$ and $\mathcal{T}_{P,k}$ denoting the spatial and temporal domains, respectively, for the kth type of proxy observation.

One way to build a statistical model for paleo-reconstruction would be to write down the distribution of each instrumental, $\mathbf{Z}_{I,j}$, and proxy, $\mathbf{Z}_{P,k}$, data source directly, conditioning on the latent climate process $\mathbf{Y}$. However, we suggest specifying the relationships between the various types of observation and the latent climate field using a two-stage model. To motivate this two-stage modeling approach, consider a spectrum of pollen counts extracted from a sample of a lake sediment core. A simple model for the pollen–climate relationship may state that larger proportions of pollen from a particular, indicator taxon correspond to warmer temperatures. In practice, a researcher often extracts a fixed number of grains, which are then sorted into taxa. Conditional on the overall count and the latent, true probability of a given grain belonging to the indicator taxon, the observed count of the indicator taxon follows a binomial distribution (e.g., Ohlwein and Wahl, 2012). The observed counts are thus used to estimate the parameter of a binomial distribution, and uncertainty is introduced by the limited sample size and effects such as the preferential degradation of certain pollen species. In addition, the model relating the binomial parameter to the climate is likely an imperfect representation of the factors that affect the pollen spectra, in the sense that, given the actual (as opposed to estimated) parameters of the model, there remains uncertainty about the state of the climate system. The same holds true for other proxy types: the model which relates a standardized site chronology of tree ring widths to climate contains uncertainty, while measurements errors, the changing number of trees as a function of time, and uncertainty in the standardization algorithm all introduce additional uncertainty.

In each case, uncertainty arises from two distinct sources: the limitations of the model relating the proxy to the climate, and limitations of the observations, including measurement errors and finite sample size. While a two-stage model for the data adds complexity to the modeling framework, it also allows for these two different sources of uncertainty to be modeled separately from one another. Models that relate the climate to various data sources are discussed in Section 5, while issues concerning the observations are discussed in Section 6. In addition, the two-stage model provides flexibility in modeling the missing data mechanism (Section 6.3), and accounting for the fact that inference may be required at one spatial and temporal scale (for example, annual means of grid box averages), while observations are on different scales (for example, seasonal averages at specific locations; see Sections 4.2 and 5.2).

To account for these two different sources of uncertainty, we introduce and condition upon an intermediate set of space–time processes. Let

$$\mathbf{W}_{P,k} = \{W_{P,k}(\mathbf{s},t) : \mathbf{s}\in\mathcal{D}_{P,k}, t\in\mathcal{T}_{P,k}\}, k = 1,\ldots,N_P, \quad (5)$$

denote the N_P latent, error-free, proxy processes, each associated with the corresponding proxy data type, $\mathbf{Z}_{P,k}$. In the context of the pollen example, $\mathbf{W}_{P,1}$ would correspond to the true proportions of various pollen species (as a function of time and space), while $\mathbf{Z}_{P,1}$ would correspond to the measured spectra. Note that the spatial and temporal domains of the $\mathbf{W}_{P,k}$ are somewhat arbitrary, and it may be useful to specify the spatial domain of the $\mathbf{W}_{P,k}$ as larger than those for the corresponding $\mathbf{Z}_{P,k}$. For example, setting the spatial domain of $\mathbf{W}_{P,1}$ to be all locations where the first proxy type could potentially be measured may be of use in selecting future sampling sites.

Similarly, let

$$\mathbf{W}_{I,j} = \{W_{I,j}(\mathbf{s},t) : \mathbf{s}\in\mathcal{D}_{I,j}, t\in\mathcal{T}_{I,j}\}, \quad j = 1,\ldots,N_I \quad (6)$$

denote the N_I latent, error-free, instrumental processes, each associated with the corresponding type of instrumental observation, $\mathbf{Z}_{I,j}$. In the case of the CRU gridded temperature anomaly product (Fig. 1; Brohan et al., 2006), the two-stage model provides flexibility in modeling the key features of the data, including the spatial averaging of the underlying temperature field, the spatially and temporally varying availability of station observations within the grid boxes, and uncertainties associated with the raw station data (see Section 5).

The likelihood is then defined in terms of a product of the following:

1. The joint distribution of the latent space–time climate process $\mathbf{Y}$;
2. The joint distribution of the error-free instrumental and proxy processes, $\{\mathbf{W}_{I,j} : j = 1\ldots N_I\}$ and $\{\mathbf{W}_{P,k} : k = 1\ldots N_P\}$, conditional on $\mathbf{Y}$;
3. The joint distribution of the instrumental and proxy data, $\{\mathbf{Z}_{I,j}\}$ and $\{\mathbf{Z}_{P,k}\}$, conditional on the error-free processes $\{\mathbf{W}_{I,j}\}$ and $\{\mathbf{W}_{P,k}\}$ and the climate process $\mathbf{Y}$.

These distributions will also depend on a number of unknown statistical parameters (such as autoregressive coefficients, spatial ranges, and measurement error variances) and may also depend on covariates (such as latitude, longitude, proximity to a coastline, or spatial maps indicating where trees grow over the globe). To allow for Bayesian inference, it is necessary to specify a prior distribution for the unknown statistical parameters, which we label $\boldsymbol{\theta}$. We make the simplifying assumption that the measurement error mechanisms are conditionally independent across data sources, and do not depend on the climate process $\mathbf{Y}$. The posterior distribution then follows from Eq. (1):

$$\pi(\mathbf{Y},\{\mathbf{W}_{I,j}\},\{\mathbf{W}_{P,j}\},\boldsymbol{\theta}|\{\mathbf{Z}_{I,j}\},\{\mathbf{Z}_{P,k}\}) \propto f(\mathbf{Y}|\boldsymbol{\theta})g(\{\mathbf{W}_{I,j}\},\{\mathbf{W}_{P,k}\}|\mathbf{Y},\boldsymbol{\theta})\left[\prod_{j=1}^{N_I} h_{I,j}(\mathbf{Z}_{I,j}|\mathbf{W}_{I,j},\boldsymbol{\theta})\right]\left[\prod_{k=1}^{N_P} h_{P,k}(\mathbf{Z}_{P,k}|\mathbf{W}_{P,k},\boldsymbol{\theta})\right]\pi(\boldsymbol{\theta}). \quad (7)$$

Techniques such as MCMC sampling can then be used to draw from the posterior distribution of the latent processes – including the latent climate process $\mathbf{Y}$, which is the main object of interest – and unknown statistical parameters, conditional on the data.

The specification of each component of the model will be discussed in subsequent sections.

4. Modeling the latent space–time climate process

Defining the probability distribution of the space–time climate process $\mathbf{Y} = \{Y(\mathbf{s},t) : t\in\mathcal{T}, \mathbf{s}\in\mathcal{D}\}$ is an important step in the construction of the statistical model, especially with regard to providing estimates of parameter uncertainty. Usually the chosen distribution is continuous, but there are some situations where the

distribution should be discrete (for example, modeling the presence or absence of sea ice). In the continuous case, it is common to assume that the process is Gaussian, so that the joint distribution at any combination of spatiotemporal locations is multivariate normal. For a climate variable such as annual mean surface temperature anomalies, the Gaussian process assumption is likely reasonable because of averaging. For other quantities, however, there is strong evidence against this assumption. For example, distributions of precipitation are right-skewed, often with elevated levels of zero precipitation (e.g., Bellone et al., 2000; Berrocal et al., 2008) – in this case, a cubic root transform may be approximately Gaussian (e.g., Stidd, 1953; Li et al., 2008). In the modeling of extremes (for example, maximum temperatures), marginally we expect observations to be well modeled by the family of generalized extreme value distributions (e.g., Coles, 2001). Non-Gaussian models necessarily involve more complicated analysis schemes. While the space–time process may be non-Gaussian, it is common practice to model the statistical parameters characterizing the distribution using Gaussian processes (e.g., Leith and Chandler, 2010) or transformations thereof; counter examples that use non-Gaussian distributions include Roe and Baker (2007) and Frame et al. (2005).

For large enough spatial regions and long enough temporal ranges, it is acknowledged that most space–time climate processes, including temperature anomalies, are *non-stationary* in both space and time – the joint distribution (or more weakly the mean and covariance) at locations and times $(\mathbf{s}, t)$ and $(\mathbf{s}', t')$ cannot be expressed in terms of the offset between the locations and time points, $(\mathbf{s} - \mathbf{s}', t - t')$. There are two common ways that non-stationarities are modeled in practice (while we demonstrate with Gaussian processes, the same ideas extend naturally to non-Gaussian processes):

1. The traditional space–time statistical modeling approach (e.g., Sahu and Mardia, 2005; Le and Zidek, 2006; Gneiting et al., 2007) expresses $\mathbf{Y}$ as

$$Y(\mathbf{s}, t) = \mu(\mathbf{s}, t) + \varepsilon(\mathbf{s}, t), \quad s \in \mathcal{D}, t \in \mathcal{T}. \tag{8}$$

   The first term, $\boldsymbol{\mu} = \{\mu(\mathbf{s}, t) : \mathbf{s} \in \mathcal{D}, \quad t \in \mathcal{T}\}$, captures the spatially- and temporally-varying mean effects – such as trends in space and/or time, and dependencies on fully observed covariates. The mean-zero stochastic process $\boldsymbol{\varepsilon} = \{\varepsilon(\mathbf{s}, t) : \mathbf{s} \in \mathcal{D}, t \in \mathcal{T}\}$ then captures the spatially- and temporally-varying covariance effects of the residuals, usually based on the simple idea that the covariance between two spatiotemporal locations decreases as a function of separation in both space and time. In the simplest models, $\boldsymbol{\mu}$ may capture all the non-stationarities in the climate process, so that ε may be specified as a stationary process in space and time.

2. In climate science, it is common to express the multivariate Gaussian distribution for $\mathbf{Y}$ in terms of a reduced number of spatial basis functions, such as the leading empirical orthogonal functions (EOFs), canonical correlation patterns, or large-scale teleconnection signals (e.g., the El Niño-Southern Oscillation (ENSO) pattern of variability, or the North Atlantic Oscillation pattern). Such models generally assume that the climate process can be well approximated as a time-varying linear combination of the leading basis functions, while assuming that the basis functions themselves are stationary-in-time.

A useful strategy in practice is to combine both methods, by including covariates, basis functions, and a space–time covariance function. The mean in the decomposition defined by Eq. (8) is first specified with the aim of capturing any observed non-stationarities of the climate process. As a constant or simple trend model in space or time is unlikely to be a good fit, the mean $\boldsymbol{\mu}$ is often better modeled as a linear combination of meaningful covariates, which may include functions of time and space, such as year, month, latitude, and longitude; more involved functions such as indicators of land, sea, or coastline; solar, volcanic, and green house gas forcing time series (e.g., Li et al., 2010); indexes such as ENSO; or basis functions such as EOFs (e.g., Berliner et al., 2000a). Models for the mean can also include simple parameterizations of the physics of the space–time process (e.g., Wikle et al., 2001).

As a simple example suppose that we perform some form of principal components analysis (using time as a replicate) upon an empirical estimate of the spatial covariance matrix $\hat{\Sigma}$, estimated from some instrumental data or the output from a general circulation model collected at m spatial locations $\mathbf{s}_1, \ldots, \mathbf{s}_m$. Let $\{\mathbf{V}_l : l = 1, \ldots, L\}$ denote the L leading eigenvectors. In the absence of other effects (which can be added later), one model for the mean as a function of space at time $\boldsymbol{\mu}_t = (\boldsymbol{\mu}(\mathbf{s}_1, t), \ldots, \boldsymbol{\mu}(\mathbf{s}_m, t))^T$ is,

$$\boldsymbol{\mu}_t = \sum_{l=1}^{L} \beta_l \mathbf{V}_l, \quad t \in \mathcal{T}. \tag{9}$$

This model captures the L spatial effects that account for most of the variance in $\mathbf{Y}$ over space, but does not account for any temporal or spatiotemporal patterns that may be prominent in the latent space–time climate process $\mathbf{Y}$. A natural extension would incorporate time-varying dependencies, by letting the regression coefficients β_l depend on time. In either case, the remaining space–time dependencies can then be accounted for by using a space–time covariance function for ε that captures the residual dynamics and interactions observed in space and time.

There is a tradeoff between modeling the mean and modeling the covariance, and in practical applications it is often not clear what should be modeled in the mean and what should be modeled in the covariance. As pointed out in Cressie (1993) "What is one person's (spatial) covariance structure may be another person's mean structure" (p. 25). An elaborate model for the mean structure may simplify the covariance specification. In contrast, a rich covariance model can compensate for a misspecified mean structure, and this is an important reason for modeling covariances in the first place. For example, a different but natural way to use the leading EOFs from the above example would be to set $\boldsymbol{\mu}$ equal to zero (or some function of other covariates) and suppose that $\varepsilon_t = (\varepsilon(\mathbf{s}_1, t), \ldots, \varepsilon(\mathbf{s}_m, t))^T$ satisfies,

$$\varepsilon_t \sim \mathcal{N}\left(0, \sum_{l=1}^{L} \gamma_l \mathbf{V}_l \mathbf{V}_l^T\right), \quad t \in \mathcal{T}, \tag{10}$$

where $\mathcal{N}(\cdot, \cdot)$ denotes the multivariate normal distribution. Regardless of which model is chosen, the selection of the number of EOFs, L, introduces uncertainty in the specification of the model. Assuming a model given by Eq. (10) does not preclude the addition of an extra space–time covariance matrix to model residual space–time dependencies not captured by the EOFs. Starting simply, in the absence of other external information, it makes sense to initially assume that the residual covariance structure is stationary in space and time. Diagnostic analyses can then indicate whether or not the analysis warrants the use of a richer class of covariance structures.

4.1. Assuming separability

In specifying space-time covariances it is important to consider the interplay between space and time. A space–time process is *separable* if the space–time covariance function can be factored into the product of a purely spatial covariance function and a purely

temporal covariance function. The issue of space–time separability is not new to the climate sciences. Hasselmann (1993) and Stouffer et al. (2000) provide motivation and illustrations of non-separability in space–time climate processes, while Li et al. (2009) investigates the space–time covariance structure of the precipitation field.

Assuming a separable covariance structure simplifies calculations substantially, while at a potential cost of not encapsulating significant space–time interactions. In some situations, space–time predictive performance has been shown to be relatively unaffected by assuming a separable covariance structure (e.g., Genton, 2007), but to our knowledge the issue of separability has not been investigated in the context of paleoclimatic reconstructions. Tingley and Huybers (2010a,b), in the first application of a space–time process-level model to the temperature reconstruction problem, specifies a space-time separable model for convenience but does not test if the data support this assumption.

There are a number of statistical tests for separability and other properties of space-time covariance structures (e.g., Mitchell et al., 2005; Fuentes, 2006; Li et al., 2007a). In the paleoclimate reconstruction problem, the space-time sparsity of the data is an obstacle to testing for separability. Indeed, it could be argued that such data-sparsity precludes us from assuming a non-separable covariance model that fits well to the data. One useful exercise would be to investigate the space-time covariance structure of the climate process using a climate model as a testbed. Tests for separability and stationarity could be readily applied to the spatially and temporally complete model output, and the results would provide a useful guide for the specification of covariance forms in the context of the paleoclimate reconstruction problem.

To demonstrate for our application that annual mean surface temperatures are likely non-separable, we consider the CRU gridded instrumental data set (Brohan et al., 2006), confined to the region shown in Fig. 1. The covariance form assumed in Tingley and Huybers (2010a,b), is separable, and corresponds to an exponential covariance function in space and an AR(1) process in time:

$$\mathrm{Cov}(Y(\mathbf{s},t),Y(\mathbf{s}',t')) = \frac{\tau^2\phi^{|t-t'|}}{1-\phi^2}\exp(-\rho\|\mathbf{s}-\mathbf{s}'\|). \quad (11)$$

Now suppose that the temporal autocorrelation parameter ϕ varies over space. Letting $\phi(\cdot)$ denote this spatially varying parameter, the space-time covariance becomes

$$\mathrm{Cov}(Y(\mathbf{s},t),Y(\mathbf{s}',t')) = \frac{\tau^2\phi(\mathbf{s}')^{|t-t'|}}{1-\phi(\mathbf{s})\phi(\mathbf{s}')}\exp(-\rho\|\mathbf{s}-\mathbf{s}'\|), \quad t-t'\geq 0, \quad (12)$$

and the same form but with $\mathbf{s}$ and $\mathbf{s}'$ interchanged for $t-t'<0$. As long as $\phi(\cdot)$ is not constant in space, the covariance function is no longer separable because it introduces an interaction between space and time. The left panel of Fig. 4 displays the estimated lag-one autocorrelation at each grid location – evidence of non-separability will follow if this autocorrelation varies spatially. An estimate of the standard error derived using a block bootstrap (cf. Lahiri, 2003) resampling is shown in the right panel of Fig. 4.[5] Taken together, the two panels of Fig. 4 indicate that the autocorrelation varies over space, which implies that the space-time process is not separable. This result should be considered in any process model for annual mean surface temperature anomalies, with the caveats that the proxy data is much sparser than the instrumental data, and that (as stated above) predictive performance is not always affected by incorrectly assuming a separable covariance form (e.g., Genton, 2007).

[5] To calculate the block bootstrap estimate of the standard error, we split the time series into blocks of four and resample these blocks to obtain a time series the same length as the original series. We estimate the lag-one autocorrelation for this resampled series. Repeating, we calculate the standard deviation of the estimated autocorrelations over the many resampled series.

4.2. Issues of spatial and temporal support

Care should be taken when specifying the spatial and temporal support of the latent space-time climate process. For example, the spatial domain $\mathcal{D}$ need not be a simple subset of two- or three-dimensional space. Statisticians have been active in developing statistical models for data observed on the globe (see, e.g., Jones, 1963; Das, 2000; Huang et al., 2002; Stein, 2005), and the measure of distance used in the specification of a spatial covariance function on a sphere must be chosen carefully. Banerjee (2005) illustrates that great circle distances can result in invalid (singular) covariance matrices, and instead suggests the use of chordal distance, the distance between two locations in $\mathbb{R}^3$.

In some settings, it may be of interest to infer the latent climate process at one level of spatial or temporal averaging (e.g., annual means averaged over regular spatial grid boxes), despite each data source having different spatial and temporal supports – see Fig. 3 for evidence that different data sources reflect the climate averaged over different temporal supports. This so-called "change-of-support" problem will be touched upon below (Section 5.2); see Gelfand et al. (2001) for a more general discussion.

5. Forward models for climate proxies

We propose linking the instrumental and proxy observations to the latent space-time climate process via a two level statistical model. This two level approach dissociates the treatment of issues related to the observations, such as measurement error and missing data, from modeling the scientific understanding of how the error-free instrumental and proxy processes, $\{\mathbf{W}_{I,j}\}$ and $\{\mathbf{W}_{P,k}\}$, are causally affected by the underlying climate process, $\mathbf{Y}$. In this section we discuss models for the latter, namely defining $g(\{\mathbf{W}_{I,j}\},\{\mathbf{W}_{P,k}\}|\mathbf{Y},\boldsymbol{\theta})$ in Eq. (7). The form of this model is usually called the *forward model* for the data processes (e.g., Hughes and Ammann, 2009).

5.1. Forward modeling

Specification of forward models requires an accurate understanding of how each data type records information about the latent climate process. In what follows we make the common and simplifying assumption that the latent instrumental and proxy processes are conditionally independent, given the latent climate process and unknown parameters:

$$g(\{\mathbf{W}_{I,j}\},\{\mathbf{W}_{P,k}\}|\mathbf{Y},\boldsymbol{\theta}) = \left[\prod_{j=1}^{N_I} g_{I,j}(\mathbf{W}_{I,j}|\mathbf{Y},\boldsymbol{\theta})\right] \times \left[\prod_{k=1}^{N_P} g_{P,k}(\mathbf{W}_{P,k}|\mathbf{Y},\boldsymbol{\theta})\right]. \quad (13)$$

This conditional independence assumption implies that any correlation between two proxy or instrumental processes, such as pollen spectra and the CRU temperature anomalies (Brohan et al., 2006), is the result of a common dependence on the underlying climate, $\mathbf{Y}$.

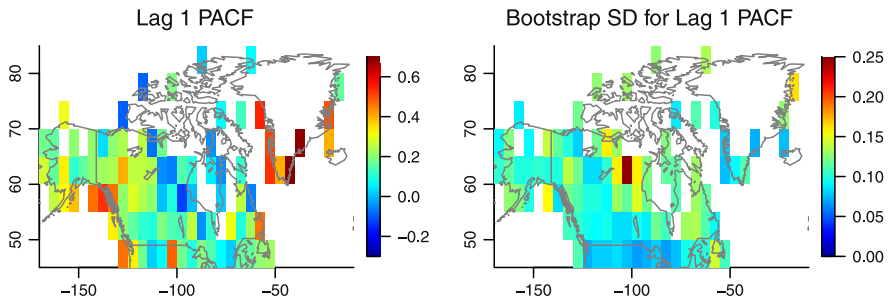

Fig. 4. Left panel: estimates of the first order partial autocorrelation function for the instrumental time series from Fig. 1. Right panel: bootstrap estimates of the standard deviation of the estimated first order partial autocorrelation. Note that this *temporal* property varies *spatially*, indicating that the process is not separable in space and time.

Taken in turn, each latent process, $\mathbf{W}_{I,j}$ or $\mathbf{W}_{P,k}$, is the dependent variable and $\mathbf{Y}$ is treated as the independent variable. In the special case of a linear model, this formulation is sometimes called "indirect regression" in the climate literature (e.g., Christiansen et al., 2009). There is little consistency in terminology, however, as ter Braak (1995) refers to methods that specify the conditional distribution of the latent climate process given the observations as "inverse approaches". As it is the climate that drives the development of the proxies (and not vice versa), treating the climate as the independent variable is the natural modeling choice (Christiansen, 2011a; Tingley and Li, 2011). Bayesian inference then provides a formalism for inverting the forward model to provide inference on the target climate process, given the latent data processes.

In what follows we focus on forward models for the latent proxy processes, but note that a similar approach can be taken for the latent instrumental processes. In the context of instrumental temperatures, for example, both GISSTEMP (NASA GISS, Hansen et al., 2010) and the Climate Research Unit's products (Brohan et al., 2006) are constructed by heavily processing raw station data into monthly grid box averages. These algorithms can be interpreted in terms of a forward model that describes the gridded values as a function of the latent (and spatially continuous) space-time temperature process.

No forward model can perfectly describe the development of any real proxy as a function of climate, so each is at best an approximation to the true relationship. As the adequacy of the forward model(s) directly affects the reconstruction results, it is important to carefully consider the tradeoff between complexity and feasibility when choosing a model. For example, Li et al. (2010) specify forward models that reasonably reflect the temporal averaging associated with each proxy type, but do not include a spatial component at the process level. In contrast, Tingley and Huybers (2010a) specify much simpler forward models, but include a spatial component at the process level. All currently favored reconstruction methods make use of linear relationships (see Section 8), and moving away from such assumptions is a key area in which reconstruction methodologies may be improved. The so-called divergence issue (e.g., D'Arrigo et al., 2008), wherein certain trees at high northern latitudes have demonstrated a reduced correlation with local temperatures over the past several decades, is indicative of the complex relationship between climate and proxy observations, which may be better captured via the development of realistic forward models.

In general, a forward model for the kth proxy type is a function of the latent climate space-time process, $\mathbf{Y}$, but must also express the uncertainty in the relationship between $\mathbf{Y}$ and the latent (and measurement error-free) proxy process. In the simplest case, the forward model may be a function of the climate process, with additive errors that are independent of the climate process. The functional form of the forward model captures the biological or physical processes through which the formation of the proxy archive (for example, the wood of a tree or the calcium carbonate of a coral) is modulated by changes in the climate system, and it is expected to capture the main features of the response of the proxy with respect to climate. The stochastic component of the forward model captures variability in the proxy that is not captured by the functional dependence on the climate. Common models for the uncertainty include an additive white noise process or additive time series processes (e.g., Li et al., 2010 use an AR(2) process). Forward models that include uncertainty in the timing (or locations) of the observations are also possible.

While there are many types of proxies (tree-rings variables, pollen assemblages, borehole temperature profiles, chemical compositions of ice cores, corals and speleothems, and so forth), each with different characteristics that must be reflected in distinct forward models, in the interest of space we discuss three commonly used classes of proxy.

5.1.1. *Forward models for tree rings*

Tree growth is a complex biological process, depending non-linearly on many climate variables, including temperature, amount of sunlight, and soil moisture parameters; see, for example the Vagonov-Shasking model of Evans et al. (2006). The feasibility of incorporating realistic non-linear forward models directly into a hierarchical climate reconstruction scheme is uncertain, due to the large number of climate variables these models require as inputs, and the computational challenges posed by the non-linearities. Recently, Tolwinski-Ward et al. (2011) proposed a reduced form of the Vaganov-Shashkin model which may be more appropriate for use within hierarchical climate reconstructions.

Techniques such as the regional curve standardization (e.g., Briffa et al., 1992) are generally applied to tree ring time series to remove non-climatic growth effects. Such pre-processing steps can be thought of as models which describe climate as a function of the observations: a number of mathematical operations are applied to the observations, and the results are interpreted as indicating climate. A more logically sound approach would be to incorporate these processing steps into the forward model (Schofield, in preparation).

5.1.2. *Forward models for pollen*

Forward models for pollen that may be of use in the context of paleoclimatic reconstructions are likewise in various states of development. Adam and West (1983) proposed an exponential regression model to estimate the relationship between pollen assemblage ratios and temperatures. A statistically advanced analysis of pollen data is presented in Haslett et al. (2006b), which uses a Bayesian analysis to invert a forward model for pollen data in order to reconstruct pre-historic Irish climate. Guiot et al. (2009) present an analysis of pollen data dating back to the last glacial maximum, and use a Bayesian inference technique to invert a vegetation forward model. More recently, Wahl et al. (2010) have made progress on a binomial logistic regression model for pollen ratio data, designed to reflect the "S" shape and [0,1] range of pollen ratios.

5.1.3. Forward models for borehole temperature profiles

Unlike tree rings and pollen, borehole temperature profiles measure the results of a purely physical process: heat diffusing down through an ice sheet or bed rock. Borehole temperature profiles are measured in the same units as the climate process under study, and have been used extensively to infer past climate borehole temperature (see NRC, 2006; Jones et al., 2009, and references therein). However, the measurement is of a vertical profile of temperature with respect to depth, while the required quantity is the time-history of surface temperature. The heat equation (Carslaw and Jaeger, 1959), subject to some initial and boundary conditions, describes how surface temperatures diffuse down through the surface to from a vertical profile, so is thus a natural forward model for relating the measured quantity to the target climate process.

The pre-observation mean-surface air temperature model (POM-SAT), discussed in Harris and Chapman (2001) and Harris (2007) is perhaps the most developed forward model for borehole temperature, and accounts for the geothermal heating from the Earth's core. Li et al. (2010) and Brynjarsdóttir and Berliner (2011) construct statistical models using POM-SAT; Li et al. (2010) emphasizes that the process must mimic the "smooth" appearance of actual borehole profiles, whereas Brynjarsdóttir and Berliner (2011) considers multiple boreholes and accounts for possible spatial dependencies.

5.2. Spatial and temporal misalignment and change-of-support

To set up the reconstruction problem in a logically consistent way, it is necessary to differentiate between the continuous, unobserved climate process, and the spatial scales of both the data and the inference. As an example, the goal might be to infer past temperatures averaged over $5^\circ \times 5^\circ$ grid boxes – the same level of spatial smoothing as the CRU temperature compilation – using the data depicted in Fig. 1. Spatially, the CRU data can be thought of as annual averages over regular grid cells, while the tree ring width time series have not been aggregated: both the locations and the extent of spatial averaging varies between the data types (Fig. 1). This is an example of a spatial *change-of-support* problem (Banerjee et al., 2004). The same issue arises with respect to temporal change-of-support, as some data sources may reflect the climate averaged over only part of a year (a tree or coral's growing season) or over several years (pollen counts from lake sediment samples), while inference may be required at the annual timescale. Indeed, two measurements derived from the same proxy archive may have substantially different time series properties, perhaps indicative of differing degrees of temporal averaging (Fig. 3). These aspects of the proxy–climate relationship can be encoded by specifying that each $\mathbf{W}_{I,j}$ or $\mathbf{W}_{P,k}$ process at a particular location $\mathbf{s}$ and time t is a known function of the process $\mathbf{Y}$ over a region of space and/or an interval of time.

Temporal change-of-support has frequently been discussed in the climate literature, though often not within the context of forward models. In certain reconstructions (e.g., Moberg et al., 2005) the proxies are divided into two classes, which are treated as reflecting high and low frequency climate variability, respectively. A similar approach is taken in Rutherford et al. (2005), where the observations are filtered into high and low frequency components, each of which is analyzed separately. In the forward-modeling approach of Li et al. (2010), different proxies are explicitly modeled as reflecting the target process averaged over different time scales (see Section 8.2).

We stress that the model for the latent instrumental and proxy processes, conditional on the climate process, must reflect the scientific understanding of the spatial and temporal domain of the climate process that causally affects the development of a proxy, while $f(\mathbf{Y})$ describes the distribution of the space-time climate process. Even though a coral or tree ring chronology may be highly correlated with a climate variable at some distant location or averaged over a continent or hemisphere, the growth of the organism is causally influenced by strictly *local* climate. In such cases, the latent space-time climate process $\mathbf{Y}$ displays long-range dependencies ("teleconnections") or is highly correlated in space, so that the information given by the site chronology at one location informs $\mathbf{Y}$ over a larger region. These are features of the process level, and should therefore be accounted for in the specification of $f(\mathbf{Y})$; see Cressie and Tingley (2010) for a discussion of this issue with respect to borehole temperature profiles.

6. Modeling the observations and other data-level issues

The hierarchical model outlined in Section 3.1 includes two levels for modeling the data. Models for the latent (and measurement error-free) instrumental and proxy processes, conditional on the space-time climate process, account for the fact that proxies such as borehole temperature profiles or pollen ratios are imperfect records of the climate (Section 5). This section focuses on models for the instrumental and proxy observations, conditional on the error-free instrumental and proxy processes – these "observation models" need to account for data-level effects such as "measurement error", screening or selection effects, and missing data mechanisms.

6.1. Regression dilution and models for measurement error

All observational data sources can be thought of as incorporating measurement error. For the kth type of proxy data (or similarly for instrumental data), the simplest measurement error model is of the form,

$$Z_{P,k}(\mathbf{s},t) = W_{P,k}(\mathbf{s},t) + \varepsilon_{P,k}(\mathbf{s},t), \tag{14}$$

where $\{\varepsilon_{P,k}(\mathbf{s},t)\}$ is a white noise error process, and the relation applies at those locations $\mathbf{s}$ and times t where the kth proxy data type is observed. Eq. (14) assumes that the measurement error process is additive white noise that is independent of the corresponding data process, and such assumptions may not be valid in all cases. Possible identifiability issues with respect to the parameters of the $\{\varepsilon_{P,k}(\mathbf{s},t)\}$ process above and parameters of the corresponding forward model, $g_{P,k}(\mathbf{W}_{P,k}|\mathbf{Y},\boldsymbol{\theta})$, can be resolved by placing an informative prior on the variance of $\{\varepsilon_{P,k}(\mathbf{s},t)\}$, derived from knowledge of the actual measurement process (see Section 7.2 below and, for example, Section 2.2.4 of Santner et al., 2008).

A contentious issue in paleoclimate reconstructions concerns the general underestimation of the temporal variance of the reconstructed time series (e.g., Christiansen et al., 2009, 2010; Rutherford et al., 2010). We note that any methodology that involves predicting instrumental observations from proxy observations (a regression, as opposed to a linear forward model) using ordinary least squares will produce this effect: the variance of the predictions is strictly smaller than the variance of the actual response, even if the model is correct (see, for example, Fig. 9-1 of NRC, 2006). This effect is accentuated if the predictor variables (the proxies) suffer from extensive measurement error, as estimates of the regression coefficients will then be biased (perhaps severely so) towards zero – an effect known as regression dilution (e.g., Frost and Thompson, 2000; Tingley and Huybers, 2010b). The issue of measurement error is not new to paleoclimatologists, and the widely-used RegEM algorithm (e.g., Schneider, 2001; Mann et al., 2007; Christiansen et al., 2009, 2010; Rutherford et al., 2010) involves regularization via either ridge regression or truncated total least squares, both of which implicitly account for errors in both the predictor and response variables (see Section 8).

In the context of climate reconstruction, the calibration interval (the last 150 years) is generally warmer than times in the past. Regression dilution thus results in reconstructions of past temperatures that are biased towards warm values. Ammann et al. (2010) illustrate how measurement errors in the proxies produce biases in both the variance and the mean value of a reconstruction. Indeed, Ammann et al. (2010) identify the problem of regression dilution, and propose a correction, based on the results in Fuller (1987) and Carroll et al. (2006), that involves minimizing out-of-sample prediction bias over subsections of the calibration interval. These solutions are more stable than the total least squares solution, as the latter can potentially over-correct the bias if there is no accurate information on the error variances for both the predictor and response variables (Carroll and Ruppert, 1996), which is often the case in paleoclimate reconstructions. Many issues with respect to regression dilution in the presence of correlated errors in both predictor and response variables remain unresolved in Ammann et al. (2010).

An alternative strategy for modeling measurement errors is afforded by the hierarchical approach proposed above, which can explicitly model the errors in each type of data and can thus avoid the attenuation effects of model misspecification. Tingley and Huybers (2010b) shows that, if the process and data-level models are correct, individual draws from the posterior (see Section 7) have, on average, the correct variability, while the variance of the posterior mean remains attenuated.

6.2. Data quality and screening for correlation over the instrumental period

The various proxies used to infer past climate are collected from different sources and combined in different ways – a tree ring chronology may be based on cores from upwards of 20 trees, whereas a coral chronology does not usually feature such replication. The different data sources thus inform the target climate process with different levels of uncertainty; see NRC (2006) and Jones et al. (2009) for further discussion of related data issues.

A procedure often referred to as screening or pre-processing (Osborn and Briffa, 2006; Mann et al., 2008a) is frequently used to select a subset of proxy time series to be used in a reconstruction from a potentially much larger candidate pool. As examples, the supplement to Mann et al. (2008) describes how only those candidate proxy time series that exhibit a significant correlation ($p < .10$) with one of the two closest instrumental time series were used in the ensuing reconstruction, while McShane and Wyner (2011) makes use of the Lasso to select a subset of the candidate proxy series. There are both advantages and disadvantages to such screening procedures. On the one hand, excluding candidate proxy series that contain little or no information about the target process will improve the reconstruction, and this is the main justification for screening candidate proxies (e.g., Mann et al., 2008a). On the other hand, such procedures can result in reconstructions with artificially high levels of skill over the interval used for screening, and no skill whatsoever elsewhere. This is even more likely in the context of autocorrelated time series, where it is well known that the empirical correlation between two independent, autocorrelated time series is highly variable and can possibly appear large (Yule, 1926; McShane and Wyner, 2011). If the candidate pool of proxies is a sufficiently large number of independent AR(1) series, then at least some will meet an arbitrarily strict correlation requirement like that used in Mann et al. (2008a). Of course, a reconstruction based on such AR(1) series will have no skill outside of the screening interval. Christiansen and Ljungqvist (2011) apply a Monte Carlo screening procedure, based on generating surrogate series with the same autocorrelation properties as the test proxy series, that reduces the likelihood of including proxies with little or no predictive power outside of the screening interval. An alternative approach is to withhold a portion of the instrumental data which is then used to test or validate the reconstruction (e.g., NRC, 2006). Without careful testing of significance, screening candidate proxies according to correlation with the instrumental records can affect the statistical results of a paleoclimatic analysis. For example, Bürger (2007) discuss how the significance levels presented in Osborn and Briffa (2006) should be modified to take into account the effects of the screening procedure used in the study.

A further issue with respect to screening proxies based on correlation with the instrumental record concerns the assumption that the relationship between the two is always linear. Within the forward-modeling framework (Section 5), each data type should be modeled separately, and appropriate models may be non-linear (e.g., Tolwinski-Ward et al., 2011).

6.3. The role of missing data

Answers to the scientific questions that paleoclimatic reconstructions seek to answer rely on the availability of data sets that are accurate, long, continuous, and of broad spatial coverage (NRC, 2006). In practice, such data sets are rarely available, and the influences that the pattern of missing data can have on a reconstruction is an issue which has begun to receive considerable attention in the climate literature (e.g., Smerdon et al., 2008; Ammann et al., 2010; Tingley and Huybers, 2010b). The problem of missing values in climate related studies is not unique to paleoclimate reconstructions. Indeed, the missing data problem arises almost immediately in standard (as opposed to paleo) climate process constructions. For instance, the post-1850 instrumental temperature record is afflicted by missing data issues in both space and time, as large swaths of the globe's surface (the South Pacific, for example) are under-sampled, particularly in the early part of the record. In general, direct physical measurements of climate processes are limited temporally and spatially, with measurement quality and availability generally decreasing back in time (NRC, 2006; Emile-Geay et al., submitted for publication-a, submitted for publication-b).

The concept of missing data has different meaning or interpretations to different end users, and often tends to be context specific. Missing values in a time series of instrumental records at a given location could mean lack of data at some point in the time series, perhaps due to an instrument malfunctioning. On the other hand, it could also mean that the date from which records began is more recent as compared to another location. While the missingness in the latter case is more structural, in the sense that there cannot be instrumental observations at a particular location before an instrument is placed there, these two types of missingness can be treated in a similar manner.

Paleoclimate proxy records likewise suffer from missing data problems, or simply a lack of availability of long proxy time series. Some of this problem is structural in the sense that the very nature of the proxies tends to affect how far they go back. For example, the length of an ice core record is limited by the physical properties of the ice sheet or glacier from which it is drilled. It is also well known that the spatiotemporal pattern of available proxy data is severely skewed toward land areas, and data availability decreases very rapidly in time (see Fig. 2 and Mann et al., 2008a). The availability of proxy data mainly over land masses limits our ability to reliably estimate past climate over the oceans (e.g., Hartmann, 1994), while the decreasing number of proxy observations generally induces greater uncertainty for the earlier intervals of paleoclimatic reconstructions (e.g., Mann et al., 1998). Similarly, available proxy data sets do not generally feature broad spatial coverage, which hinders reconstructions at the regional level, as smaller sample sizes can lead to statistical estimates

which are strongly dependent on a particular type of proxy (Emile-Geay et al., submitted for publication-a, submitted for publication-b).

Within the context of paleoclimate reconstructions, missing data is thus an unavoidable reality. Several commonly used methods, however, require that the data be "block missing," with all instrumental time series covering the same interval, and likewise for the proxies. As discussed in Section 8, such methods can involve a pre-processing step to impute missing values in the data prior to performing the reconstruction. This data infilling can have unintended consequences on the analysis, as in general, the imputed values exhibit reduced variability relative to the true, missing values (Dempster et al., 1977; Gelman et al., 2003).

We discuss two statistical approaches to the treatment of missing data that have been proposed or used in the context of paleoclimate reconstructions.

6.3.1. Imputation based on linear regression

One common interpretation of the paleoclimate reconstruction problem is to regard instrumental observations of any climate variables before 1850 (and some after this point) as "missing." In this view point, popularized by the RegEM algorithm (Schneider, 2001), the paleoclimate reconstruction endeavor reduces to an imputation or missing value problem: the goal is to infer the "missing" values in available instrumental time series prior to 1850. Labeling the unknown instrumental variables prior to 1850 as "missing values" is a nomenclature used in the literature (e.g., Schneider, 2001); in practice, the reconstruction proceeds by calibrating the proxy data against available instrumental data (post 1850) and then using that calibration to predict instrumental records prior to 1850.

If there are no missing values in the joint proxy-instrumental data set, then both the mean of each time series and the covariance matrix between them can be readily estimated. Likewise, given the mean and covariance, and the assumption that the data at each time point is multivariate normal, the missing observations for each year can be imputed using standard regression expressions. In the presence of missing values, the simultaneous estimation of the mean and covariance and the imputations of the missing values is a non-linear problem with no general closed-form solution. The Expectation–Maximization (EM) algorithm of Dempster et al. (1977) is an iterative procedure which overcomes the missing data problem by first starting with an initial estimate of the mean and covariance of the incomplete data, and then estimating regression coefficients that are used to infill the missing values. Once a complete data set is available, updated estimates of the mean and covariance are calculated from the complete data. The new estimates are in turn used to obtain new imputed values and the process is repeated until convergence. Schneider (2001) proposes several schemes for regularizing the covariance estimation in the Expectation step if the number of available observations is small; see Section 8 for further details.

The EM-based approach, which jointly models the proxy and instrumental time series and treats the instrumental records in the past as a large block of missing data, has several strengths. It is a general and transparent framework for reconstructing paleoclimate processes and the computational costs can be substantially lower than those required for Bayesian inference with a hierarchical model. There are, however, technical limitations to regularized variants of the EM algorithm. Providing uncertainty estimates for the imputations can be a challenge (see, e.g., Schneider, 2001; Tingley and Huybers, 2010b). In addition, the EM algorithm is a local search algorithm so is not guaranteed to find the global maximum of the likelihood. Some of these limitations of the can be circumvented by exploiting the fact that the bulk of the missing data in paleoclimatic reconstruction problems generally have a staircase pattern (Rajaratnam, 2010). In a perfect staircase pattern, each time series ends at the same time point (say, this year), the time series extend back to different points in the past, and each time series is complete (no missing observations between the first and last). Rajaratnam (2010) demonstrates that if the missing values follows this pattern, then closed-form expressions for the mean and covariance function can be calculated analytically. This closed-form approach is much faster than the EM-approach, ensures that the solution is the global maximum likelihood estimator, and facilitates the calculation of the variance of the resulting imputations. Even when the monotone incomplete pattern assumption is violated, the closed-form approach can be used to produce an extremely good starting value for the EM algorithm.

6.3.2. Imputation via space-time modeling

The hierarchical modeling approach developed above provides a natural framework for imputing missing values. Indeed, under this formulation, the entire latent space-time climate process, $\mathbf{Y}$, is unobserved. The forward model and the observation model describe how each observation relates to $\mathbf{Y}$, while the model for $\mathbf{Y}$ indicates how information about the climate process is shared across space and time. Depending on the particulars of the analysis, the missing data process can be modeled in the specification of the forward model, the observation model, or both. Whereas (Reg)EM imputes missing values in incomplete instrumental time series, the specification of a space-time covariance form allows for predictions of the process at *any* location – even those where there are no observations at any time. A reasonable space-time statistical model allows for appropriate uncertainty quantification in the imputation of the missing values, and imputations at locations or times many decorrelation length scales from the nearest observations will naturally be highly uncertain. In addition, the space-time covariance model can be used to estimate the reduction in uncertainty, as a function of space and time, given an additional observation with a known uncertainty. For a detailed description of the differences between RegEM and a space-time modeling approach, see Tingley and Huybers (2010b).

6.3.3. Inference in the presence of missing data

While each of these approaches provides a concrete set of tools for handling missing values, a word of caution is in order. These methods implicitly assume that the missing data mechanism itself has not distorted the properties of the observed data. In other words, there is an implicit assumption that the missing data mechanism (or more precisely the distribution of the missing data mechanism) does not depend on the values of the missing observations. Understanding the mechanisms that lead to missing data is critical in assessing their influence on statistical inference and the type of biases that missing data can introduce. The notion "missing at random" (MAR; Little and Rubin, 2002), sometimes referred to as "ignorability," characterizes this problem, and assuming that observations are MAR allows one to safely ignore the missing data mechanism. Indeed, in the paleoclimate context the methods proposed in Schneider (2001), Rajaratnam (2010) and Tingley and Huybers (2010a) all make the MAR assumption, so will retain desirable properties of standard statistical estimators only if this assumption is correct. Violations of the MAR assumption can, in each case, lead to estimates of the climate process that are biased or otherwise sub-optimal. Rubin (1976) and Little and Rubin (2002) provide comprehensive treatments of statistical inference in the presence of missing values.

Recent work with RegEM by Smerdon et al. (2008) indicates that the MAR assumption is likely incorrect in the paleoclimate context. Observations are predominantly missing in the pre-instrumental period, when temperatures are generally colder than in the data-dense instrumental period: there is a dependency between the

probability of data being missing and the value of those missing observations. A possible solution is to include CO_2 concentration (which is also correlated with temperatures) as a covariate in the process level, as is done in Lee et al. (2008) and Li et al. (2010).

7. Inference and computation

Models such as those described in Sections 3–6 are complicated, multi-level, and often incorporate non-trivial space-time dependencies. Inference for such models is generally performed via a Bayesian approach (cf. Gelman et al., 2003; Carlin and Louis, 2009). Markov chain Monte Carlo (MCMC) is a convenient general-purpose algorithm for carrying out Bayesian inference and is therefore central to fitting the kind of hierarchical models we describe in this paper. In this section we briefly describe MCMC algorithms and some issues related to prior specification for hierarchical models.

7.1. Markov chain Monte Carlo

Much of what we have described above involves building hierarchical Bayesian models where inference is based on the posterior distributions of the parameters. This posterior distribution, as shown in Eq. (1), is rarely available in closed form. Furthermore, posterior distributions for hierarchical models are typically multi-dimensional and complicated. Hence, learning about the properties of the distributions, for instance the means and variances or a tail probability, is non-trivial. This is because computing an expectation requires high-dimensional integration involving the posterior distribution.

Fortunately, MCMC methods provide a very general recipe for drawing samples according to a given distribution, and then using those samples to estimate properties of the distribution. Thus scientists can routinely obtain estimates of the entire posterior distribution of interest, including all marginal and joint distributions, allowing a diverse set of scientific questions to be answered. As a simple example, if MCMC samples are available from the posterior distribution of a parameter, say $\boldsymbol{\theta}$, the average of these samples converges to the true mean of the posterior distribution of $\boldsymbol{\theta}$ as the MCMC sample size increases. MCMC-based inference has therefore allowed statisticians and modelers in numerous fields to fit increasingly sophisticated models.

The use of MCMC has become common in many disciplines (see, for instance, the influential Gilks et al. (1996) or Brooks et al. (2011)), and there are now a number of statistical software packages that construct MCMC algorithms when presented with a fully specified model or a function proportional to the posterior distribution. The software package WinBUGS (Lunn et al., 2000), for example, provides a framework for fitting a very wide range of Bayesian models. The algorithms require minimum user intervention in principle, which makes WinBUGS very useful for non-expert users of MCMC, and its GeoBugs module allows for certain classes of spatial models to be fit. For some of the more involved space-time models discussed in this paper, however, WinBUGS may be slow and sometimes impractical. More recently a number of more specialized R (Ihaka and Gentleman, 1996) packages have been developed, for example geoR (Ribeiro and Diggle, 2001) for maximum likelihood-based inference for spatial data, spBayes (Finley et al., 2007) for Bayesian inference for an array of Gaussian process models, and ramps (Smith et al., 2008) for joint linear models for point-level and aggregated spatial data.

While these software packages are useful for certain models, it is often the case that MCMC algorithms have to be constructed and implemented on a case-by-case basis for complicated hierarchical models. Issues involved in the design of an MCMC algorithm include finding ways to make the algorithms efficient, determining the accuracy of estimates based on these algorithms to ensure that the scientific conclusions are valid, and determining an appropriate length (sample size) for the MCMC algorithm. The literature in this area is vast; we point interested reader to Robert and Casella (2004); Brooks et al. (2011), and references therein. Flegal et al. (2008), for instance, provides a simple and theoretically-justified approach for assessing the accuracy of MCMC-based estimates and for using these estimates to determine MCMC sample size. The level of expertise required in constructing MCMC algorithms for sophisticated models is perhaps another reminder that analyses of the kind discussed in this paper may be done most expediently via long term collaborations between climate scientists and statisticians.

7.2. Choice of prior distributions for Bayesian models

Careful prior specification is an integral part of a Bayesian analysis. The ideal situation is one where scientific expertise or past information provides a reasonable judgment on the distribution of possible values of a parameter in the model. In our notation, this prior distribution is denoted by $\pi(\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the vector of parameters. We note that for simplicity the prior distribution is typically specified independently for each parameter (each component of $\boldsymbol{\theta}$), though this is by no means strictly necessary. Translating expertise into a probability distribution may still be a challenge, although recent research has led to many successful approaches for formal "prior elicitation" (see O'Hagan et al., 2006). When formal prior elicitation proves to be a challenge, more informal approaches for characterizing expert judgment may be used, and simple parametric forms are chosen based on domain expertise. In situations where scientific expertise may be equivocal, the Bayesian approach allows for multiple analyses based on different priors; the agreement or differences in results based on these different priors may be of scientific interest in their own right. In many cases, however, the parameters may be hard to interpret or there may be a paucity of reasonable scientific knowledge that can inform prior selection for them.

When scientific expertise is unavailable or hard to formalize, so-called "objective" or "reference" priors (see, for instance, Kass and Wasserman, 1996; Berger, 2006) may be useful. Such priors purport to be objective in terms of the information they provide about the parameters. While there are success stories, deriving these priors can pose serious mathematical challenges in general, and doing so is often infeasible for complicated multi-stage hierarchical models, such as those presented in this paper. In practice, therefore, it is common to rely on so-called "vague" or "uninformative" prior distributions that have high variance and are therefore, at least in a simplified view, largely uninformative about a parameter. These vague priors, however, can actually be much more influential than desirable, particularly in the case of variance parameters (see, for instance Gelman, 2006, and references). Even when scientific expertise is being used, erroneous or overly confident expert priors (for example, priors that might be implicitly using some of the same information or data being analyzed), may have undue influence on the results of an analysis (cf. Frame et al., 2005). In fact, in virtually all the cases discussed above, a thorough Bayesian analysis requires careful study of the effects of various prior specifications on the resulting analysis. "Robust Bayesian inference" involves studying the sensitivity of results to uncertain inputs, such as prior specification, model specification, and utility functions (see, e.g., Box, 1980; Berger, 1994). Prior sensitivity analyses can often be carried out fairly efficiently via sampling–importance–resampling (cf. Gelman et al., 2003, p. 450), where the samples produced from a posterior distribution under one prior specification are simply resampled (with a weighting scheme) to produce the posterior under a different prior specification. This avoids having to

construct new MCMC algorithms for each new prior distribution. Other approximation approaches are also available (cf. Kass et al., 1989). We note that while prior specification may seem like a daunting task, there is value in the formalism that comes from having to explicitly specify prior judgments on parameters and seeing, based on the posterior distribution, how the data have informed our understanding of these parameters.

The number of unknown quantities in a hierarchical model for inferring past climate will almost certainly exceed the number of observations available to fit the model – recall that the entire target field is treated as latent, while data availability can be sparse in both space and time. Modeling the spatial and temporal covariance of the target field at the process-level ameliorates this issue by reducing the number of *independent* unknown parameters. More generally, provided that the priors are proper (that is, each integrates to one), inference remains well defined even if the number of independent parameters to be estimated exceeds the number of independent observations. In this case, the influence of the priors becomes more apparent, the posterior distributions will display little learning relative to the priors, and diffuse priors will lead to diffuse posterior distributions.

7.3. Computational issues posed by large space-time data sets

The dimension of the covariance matrix of a process-based space-time model for $\mathbf{Y}$ is given by the total number of spatiotemporal locations where observations exist or inference is required. In the case of a 1000 year paleoclimate reconstruction on a global 5° by 5° grid, this dimension will be well over one million. Both maximum likelihood and Bayesian approaches require repeated evaluations of the likelihood, and thus involve operations such as the repeated inversion or Choleski factorization of the covariance matrix – a calculation which scales as n^3, where n is the size of the process vector $\mathbf{Y}$. Computation for such large space-time data sets can quickly become prohibitively expensive and slow.

Several modeling and computational approaches have been proposed to allow for fast inference when spatial or spatiotemporal data sets are large. We briefly list a few of these approaches and provide references for the interested reader. The methods may be roughly classified into: (i) approaches that work with the covariance directly by either exploiting a sparse covariance matrix structure (see, for instance, the covariance tapering approach in Furrer et al., 2006; Kaufman et al., 2008) or by using a "reduced-rank" or process convolutions approach to constructing a covariance (cf. Higdon, 1998; Cressie and Johannesson, 2008), and (ii) inference based on approximations to the likelihood, including spectral domain methods. While approaches that fall into category (i) easily allow for both maximum likelihood and Bayesian approaches, methods that belong to category (ii) have most often been used in the maximum likelihood framework (cf. Whittle, 1953; Vecchia, 1988; Caragea, 2003; Stein et al., 2004; Fuentes, 2007). In principle, they may also be used for Bayesian approaches, but rigorous inference may pose some challenges since it may be hard to justify studying posterior distributions based on approximations to likelihoods (it may be the case that the approximation used does not correspond to a valid probability model).

7.4. Inference and uncertainty quantification

Inferring a point estimate of the latent climate process $\mathbf{Y}$, along with an estimate of the associated uncertainty, is a non-trivial task. That said, many interesting aspects of the space-time process cannot easily be deduced from a single, point estimate of the process and an associated uncertainty. There is often interest in determining the likelihood that particular years or decades were the warmest (or coldest) in the interval covered by the reconstruction, and a general interest in understanding the extent to which recent decades are "extreme" (e.g., Jansen et al., 2007). In addition, there may be interest in investigating the spatial variance as a function of time (e.g., Osborn and Briffa, 2006), temporal or spatial trends in the process, or the likelihood that a particular interval was warmer or cooler than some measure of the baseline. There is also considerable interest in understanding the temporal evolution of the process at different temporal scales, and results are often presented after smoothing the inferred process through time (e.g., Jansen et al., 2007). While a point estimate of a quantity like the temporally smoothed time series of spatial averages can be derived from the point estimate of the process, estimating the uncertainty in such a derived quantity can be non-trivial. Finally, any inference based on conditional expectations or ordinary linear regression will result in estimates of the process, or the time series of spatial means, that has a lower temporal variance than the actual climate process (see Section 6.1).

The breadth of questions that a researcher might want a reconstruction to answer points to the need for inference on the full statistical distribution of the process. A straightforward solution is to perform the analysis in a way that produces draws or samples of the space-time process. The two standard methodologies for producing such samples are bootstrapping and posterior predictive sampling. Li et al. (2007b) present an analysis of the 14 proxy series used in Mann et al. (1999) that extend back at least 1000 years, and use a parametric bootstrap method to produce ensembles of reconstructions. These ensembles are then used to assess the extent to which recent temperatures are anomalous in the context of the previous 1000 years. Alternatively, Bayesian methods like those presented above, in Tingley and Huybers (2010a) and in McShane and Wyner (2011), can be used to produce posterior draws of the space-time climate process (or, in the case of McShane and Wyner (2011), a time series) conditional on the data and modeling assumptions. Regardless of the method used to produce them, ensembles of reconstructions are rich end products that can be used to answer a vast array of scientific questions. Indeed, ensembles can be used to produce both a point estimate and an uncertainty for any function of the target process, and thus can provide insight into a diverse array of questions, such as the extent to which certain intervals were extreme, or the evolution of the process on an array of temporal scales. Both point-wise uncertainty intervals for each element of a reconstructed time series, and path-wise uncertainty intervals for entire time series, can likewise be derived from the ensembles (e.g., McShane and Wyner, 2011).

8. Special cases from the literature

Expressing a number of published methods within the hierarchical modeling framework developed above reveals that the *modeling assumptions* are often quite similar, while the *inference techniques* tend to differ. While it is somewhat difficult to frame certain of these techniques as hierarchical models, as not all consider a process level distinct from a data level, we do so to the extent possible in order to illustrate the modeling assumptions made by each, to differentiate these assumptions from the tools used to conduct the inference, and to explore logical shortcomings which are avoided by hierarchical modeling.

The terms *composite plus scale* (CPS) and *climate field reconstruction* (CFR) are often used to differentiate between reconstructions of a large-scale spatial mean, and reconstructions of a space-time process (e.g., Jones et al., 2009). Note that inference on the spatial mean time series is a natural by-product of successfully inferring the full space-time process. Several studies (Cressie and Tingley, 2010; Tingley and Huybers, 2010a,b) have called into

question the appropriateness of inferring the time series of spatial averages without considering the spatial covariance of the underlying process, as the resulting confidence intervals for the spatial mean time series can be severely biased. Including a spatial or space-time covariance function at the process level provides for internally consistent uncertainty estimation in the presence of clustered proxy data, by naturally down-weighting the information from nearby proxies and accounting for the uncertainty introduced by areas which are not sampled.

CPS can be thought of as a special case of CFR, where the spatial domain of the target process is a single point, and there is a single instrumental and a single proxy time series. Note, however, that the methods used to construct the proxy and instrumental composite time series may have implicit or explicit spatial elements. For example, the proxy composite can be formed by weighting the proxy time series by estimates of the spatial extent represented by each series (e.g., Esper et al., 2002; Mann and Jones, 2003).

In what follows, we largely consider climate field reconstruction methods, but include three methods – those of Lee et al. (2008), Li et al. (2010) and McShane and Wyner (2011) – which, while lacking a spatial component, have features similar to the hierarchical modeling framework described in Section 3. The methods are discussed in order (by our judgement) from most to least hierarchical and Bayesian, as certain of the less hierarchical methods can then be presented as special cases. We end the section with a more general discussion of the weaknesses common to regression-based eld reconstruction methods, and the advantages of hierarchical modeling combined with Bayesian inference.

8.1. BARCAST (Tingley and Huybers, 2010a)

BARCAST, described in Tingley and Huybers (2010a), is a hierarchical model that infers a spatially and temporally complete climate process from incomplete proxy and instrumental time series. At the process level, $\mathbf{Y}$ is modeled as a Gaussian process with constant mean and a space-time covariance structure that is separable, AR(1) in time, and isotropic, stationary and exponential in space. The covariance between two spatiotemporal points is thus given by Eq. (11).

At the data level, there is no distinction between measurement error and the stochastic relationship between $\mathbf{Y}$ and the latent data processes. Following the notation developed in Sections 5 & 6, the one-stage data-level model involves the specification of $h_{I,1}(\mathbf{Z}_{I,1}|\mathbf{Y},\boldsymbol{\theta})$ for the single instrumental data set, and $h_{P,k}(\mathbf{Z}_{P,k}|\mathbf{Y},\boldsymbol{\theta})$, $k = 1,\ldots,N_p$, for each proxy data type. The instrumental observations are modeled as reflecting the $\mathbf{Y}$ process with additive white noise,

$$\mathbf{Z}_{I,1}(t)|\mathbf{Y}(t),\sigma^2_{I,1} \sim \mathcal{N}\left(\mathrm{H}_{I,1}(t)\cdot\mathbf{Y}(t),\sigma^2_{I,1}\mathrm{I}\right), \tag{15}$$

where $\mathrm{H}_{I,1}(t)$ is a selection matrix the picks out the locations for which there are instrumental observations at year t, I is the identity matrix, and $\sigma^2_{I,1}$ is the instrumental measurement error variance. There is an implicit assumption that the instrumental data process is free of systemic errors, which could be modeled by including a specification for both $g_{I,1}(\mathbf{W}_{I,1}|\mathbf{Y},\boldsymbol{\theta})$ and $h_{I,1}(\mathbf{Z}_{I,1}|\mathbf{W}_{I,1},\boldsymbol{\theta})$.

BARCAST is designed to assimilate an arbitrarily large number of types of proxy data, but specifies an equivalent one-stage data-level model for each type:

$$\mathbf{Z}_{P,k}(t)|\mathbf{Y}(t),\beta_{1,k},\beta_{0,k},\sigma^2_{P,k} \sim \mathcal{N}\left(\beta_{0,k}+\beta_{1,k}\mathrm{H}_{P,k}(t)\cdot\mathbf{Y}(t),\sigma^2_{P,k}\mathrm{I}\right), \tag{16}$$

where the notation follows Eq. (15). As there is no distinct model for measurement errors, there is no way to disentangle uncertainty in the assumed linear relationship between the proxy observations and climate process from the uncertainty inherent in measuring the proxies.

BARCAST assumes that all measurements are on the same spatial (and temporal) scale, and that each type of observation reflects the underlying $\mathbf{Y}$ process locally in both space and time – in other words, there is no spatial or temporal averaging. This approach presents logical challenges, particularly when the instrumental data set is the CRU $5^\circ \times 5^\circ$ gridded product (Brohan et al., 2006) – which is best interpreted as representing temperatures averaged over grid cells. As Eq. (15) does not involve spatial averaging of $\mathbf{Y}$, the specification for the instrumental observations implies that the latent process inferred in the analysis is actually the spatially averaged climate process, where the degree of spatial averaging changes as a function of latitude. As discussed in Section 2, some types of proxy data are best understood as reflecting strictly local information about the underlying, unsmoothed $\mathbf{Y}$ process; in such a scenario, the analysis scheme fails to account for the differing spatial supports of the data sources. Further shortcomings of the underlying model, including the assumption of independent and identically distributed (iid) errors, are discussed in Section 4 of Tingley and Huybers (2010a).

The inference used for BARCAST is Bayesian with proper but weakly informative priors placed on all unknown parameters. The end result of the analysis is a set of draws from the joint posterior of the climate process and parameters, conditional on the data and model assumptions.

8.2. Method of Li et al. (2010)

Li et al. (2010) describe a hierarchical model for inferring the Northern Hemisphere annual mean temperature time series, and do not include a spatial component; the process $\mathbf{Y}$ thus reduces to a time series of spatial averages. Li et al. (2010) include as covariates in the process level estimates of three climate forcing time series: green house gas concentration $\mathbf{G}$, solar irradiance $\mathbf{S}$, and volcanic forcing $\mathbf{V}$. Several variants of the model are discussed; we consider the most general.

The underlying spatial mean time series $\mathbf{Y}$ is modeled as an AR(2) process with the mean term a linear function of the climate forcings,

$$\mathbf{Y}|\boldsymbol{\theta} \sim \mathcal{N}(\beta_0+\beta_1\mathbf{S}+\beta_2\mathbf{G}+\beta_2\mathbf{V}_0,\Sigma_\mathbf{Y}), \tag{17}$$

where the covariance matrix $\Sigma_\mathbf{Y}$ corresponds to that of an AR(2) time series (see, e.g., Brockwell and Davis, 2002, p. 91), and the vector $\boldsymbol{\theta}$ consists of the process- and data-level parameters. While the time series $\mathbf{G}$ and $\mathbf{S}$ are assumed to be free of error, Li et al. (2010) include a measurement error process for the observed volcanic forcing $\mathbf{V}$ in terms of the actual series $\mathbf{V}_0$, which we do not discuss here.

In terms of the data level, Li et al. (2010) assume that there is no uncertainty in the instrumental mean time series, and thus there is no forward model or observation model for the instrumental data. Each of the three different proxy types used by Li et al. (2010) – tree rings, pollen counts, and boreholes – are assumed to represent the spatial mean time series averaged over different, known, temporal scales. As with BARCAST, there is no measurement error model, so this source of uncertainty is not disentangled from the uncertainty in the proxy–temperature relationship. Let $\mathbf{Z}_{P,k}(i)$ be the ith time series of the kth proxy type ($k = \{1,2,3\}$). The data level assumptions for the tree ring and pollen series are,

$$\mathbf{Z}_{P,k}(i)|\mathbf{Y},\boldsymbol{\theta} \sim \mathcal{N}(\mu_{k,i}+\beta_{k,i}\mathrm{M}_k\cdot\mathbf{Y},\Sigma_k)$$

where Σ_k is the covariance matrix corresponding to an AR(2) process, and the averaging matrix M_k is assumed to be known for

each proxy type. Whereas BARCAST assumes that the data-level regression parameters are common across proxy time series of a given type, Li et al. (2010) specifies different parameters for all proxy series. Note, however, that the parameters of each Σ_k are common for each proxy type, but distinct across proxy types and from those characterizing $\Sigma_{\mathbf{Y}}$. As the time series used in Li et al. (2010) are all of the same length, no selection matrices (the H in Section 8.1 above) are required. In the case of the borehole series, the covariance matrix corresponds to Gaussian white noise, and the averaging matrix is applied to the covariance matrix as well. The data-level specification for the borehole series is thus,

$$\mathbf{Z}_{P,3}(i)\big|\mathbf{Y},\boldsymbol{\theta}\sim \mathrm{M}_3\cdot\mathcal{N}\left(\mu_{3,i}+\beta_{3,i}\mathbf{Y},\Sigma_3\right). \tag{18}$$

The inference used in Li et al. (2010) is Bayesian, and the priors for all unknowns are proper but weakly informative. In particular, the priors for the AR(2) coefficients are set to ensure that the various noise sequences are causal (Brockwell and Davis, 2002, p. 85). Similar to BARCAST, the end product is an ensemble of draws from the posterior distribution of the spatial mean time series.

8.3. Method of Lee et al. (2008)

The state-space model of Lee et al. (2008) is similar to that of Li et al. (2010), as both include climate forcings in the process level. In contrast to Lee et al. (2008), Li et al. (2010) assume AR(1) errors in the process, a temporally local relationship between a single proxy composite time series and the spatial mean time series (i.e. the matrices M in the preceding subsection reduce to identity matrices), and iid errors in the data level. Lee et al. (2008) perform the inference using a Kalman filter, with relevant parameters estimated in a separate step via an optimization algorithm.

8.4. LOC (Christiansen, 2011a)

The LOC method of Christiansen (2011a) can be interpreted as a site-by-site application of a special case of the method described in Li et al. (2010), or as a special case of BARCAST (Tingley and Huybers, 2010a). To recover LOC, a number of the parameters in either Li et al. (2010) or Tingley and Huybers (2010a) are set to zero and parameters are inferred via maximum likelihood estimation. In particular, LOC, unlike BARCAST, does not include an explicit model for the spatial covariance. For further discussion of the links between LOC and hierarchical modeling, and the advantages of Bayesian inference in this context, see Christiansen (2011b) and Tingley and Li (2011).

8.5. Point-to-point regression (Cook et al., 1999)

The point-to-point regression (PPR) methodology was proposed and used in Cook et al. (1999) to reconstruct gridded Palmer Drought Severity Index (PDSI) data from tree ring measurements. While it is difficult to fit into the hierarchical framework developed in Section 3, aspects of PPR naturally point towards the benefits of including a process-level model with a parametric spatial covariance form. We provide a cursory description of the methodology that focuses on the links to the hierarchical modeling framework developed above, and refer the interested reader to Cook et al. (1999).

PPR is based on the intuitive notion that inference on the PDSI time series at a given location should be dominated by proxy time series that are, in some sense, "close" to that location. The basic method is to fit a separate regression model to predict each instrumental PDSI time series from nearby proxy time series:

$$\mathbf{Z}_{I,i} = Z_{P,i}\boldsymbol{\beta}_i + \varepsilon_i. \tag{19}$$

Conforming (to the extent possible) with earlier notation, $\mathbf{Z}_{I,i}$ denotes the instrumental PDSI time series at the ith location, $Z_{P,i}$ the matrix composed of the tree ring time series (the sole type of proxy data) that are used to infer the PDSI at the ith location, $\boldsymbol{\beta}_i$ is a vector of regression coefficients, and ε_i a vector of error terms. Inference is then via ordinary least squares regression.

Our interest is in the choice of the (screened) proxy series to consider as predictors for the PDSI at the ith location. Cook et al. (1999) consider all proxy time series within a distance $r_i = \max\{r, r_{i,5}\}$ of the ith grid point, where r is some fixed value and $r_{i,5}$ is the minimum distance that encloses at least five proxy time series. We note that r is set to reflect the regional nature of drought phenomenon and is larger than the grid spacing of the instrumental PDSI grid (Cook et al., 1999). As a result, a given proxy time series is used as a predictor for PDSI at multiple spatial locations: information is shared across space. Within a forward-modeling framework, PPR thus assumes that the ith proxy time series reflects information about PDSI on a spatial scale given by r_i. Trees, however, are influenced strictly by local climate, and (in our interpretation) the regional nature of the regressions in Cook et al. (1999) represent a confounding of process- and data-level models within a methodological framework that only considers the latter. Note also that if the target PDSI process displays spatial correlation, then information can be shared spatially to improve estimates of PDSI during the instrumental period as well.

It is possible to construct a hierarchical model which encodes the same basic assumptions as PPR, does not confound process and data levels, and expresses the proxy–PDSI relationship in terms of a linear forward model. Indeed, such a model can be constructed by making a number of key alterations to BARCAST. At the process level, the AR(1) coefficient should be set to zero (PPR does not consider temporal autocorrelation) and the exponential spatial covariance replaced with a form like the spherical covariance (e.g., Table 2.1 of Banerjee et al., 2004), which decays to zero covariance at finite separation (PPR assumes information is shared only regionally in space). At the data level, the instrumental error variance should be set to zero (PPR does not consider errors in the instrumental PDSI observations), and a separate data-level of the form in Eq. (16) should be specified for each proxy time series (PDSI performs a separate regression for each PDSI time series). This hierarchical formulation of PPR clarifies that the sharing of information through space is a consequence of the spatial structure of the process, not the characteristics of the data. In addition, the multi-colinearity problem that Cook et al. (1999) solve by replacing a matrix of proxy time series with a smaller matrix of principal components is avoided, as the proxy–PDSI relationships are specified locally.

8.6. Methods based on multivariate linear regression

A number of commonly used paleoclimatic field reconstruction methods (see, for example for example Christiansen et al., 2010, Table 1) are based on very similar multivariate linear regression models, but make use of different analysis or inference techniques. Given an (incomplete) matrix of instrumental observations, where each column corresponds to one time point and each row to one location, and an (incomplete) matrix of proxy observations, the key assumption behind this family of methods is that there is a linear relationship between corresponding columns of the instrumental and proxy matrices, with additive white noise.

These methods generally do not include a model for the temporal autocorrelation of the target field, so that the columns of the instrumental data matrix are assumed to be independent of one another, and likewise for the proxy data matrix. In addition, spatial

dependencies are not explicitly modeled via the specification of a spatial covariance function, but are accounted for in the estimate of (or approximation to) the data covariance matrices. As no spatial model is included, this family of methods seeks inference on a climate process at only those spatial locations where there are some minimum number of instrumental observations.

The stated advantage of the group of methods discussed in this section is that they consider all linear relationships between the observation time series, and can thereby exploit correlations between distantly separated proxy and instrumental time series (e.g., Jones et al., 2009). As discussed with respect to PPR, it is our view that doing so confounds process- and data-level models, and thus fails to reflect well-established scientific knowledge about either. If the process displays long-range dependencies ("teleconnections"), then this information should be included in the model for $\mathbf{Y}$ (see Section 4), while the data level should specify scientifically plausible relationships between the observations and the target process.

The basic assumption underlying each of these methods is that, for each year, the column vector of observations $\mathbf{Z}(t) \equiv (\mathbf{Z}_I(t)^T, \mathbf{Z}_P(t)^T)^T$, where the subscripts I and P indicate the instrumental and proxy observations, respectively, is multivariate normal:

$$\mathbf{Z}(t)|\boldsymbol{\mu}, \Sigma \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_I \\ \boldsymbol{\mu}_P \end{pmatrix}, \begin{pmatrix} \Sigma_{II} & \Sigma_{IP} \\ \Sigma_{PI} & \Sigma_{PP} \end{pmatrix}\right). \tag{20}$$

No initial structure is assumed for the mean vector $\boldsymbol{\mu}$ or covariance matrix Σ, each of which has been partitioned into instrumental and proxy components. If the instrumental observations are assumed to simply represent the local values of the underlying climate process (with or without measurement error), then the process-level assumption, common to all methods discussed in this section, is that the climate field at each year is an iid realization of a Gaussian process.

The data level for this class of methods specifies a relationship between the proxy and instrumental observations (c.f. Sections 5 and 6 which link the observations to the target process), and takes one of two forms, differentiated above as *linear forward modeling* and *regression*. Within the linear forward-modeling framework (Section 5), the proxy observations at each year t are expressed as a linear function of the instrumental observations,

$$\mathbf{Z}_P(t) = \boldsymbol{\beta}_0 + \beta_1 \mathbf{Z}_I(t) + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}\left(0, \Sigma_{P|I}\right), \tag{21}$$

where $\boldsymbol{\beta}_0$ is a vector of intercept terms, β_1 a matrix of regression coefficients, and $\Sigma_{P|I} = \Sigma_{PP} - \Sigma_{PI}\Sigma_{II}^{-1}\Sigma_{IP}$. The conditional distribution of $\mathbf{Z}_P(t)$ given $\mathbf{Z}_I(t)$, $\boldsymbol{\mu}$, and Σ from Eq. (20) is,

$$\mathbf{Z}_P(t)|\boldsymbol{\mu}, \Sigma, \mathbf{Z}_I(t) \sim \mathcal{N}\left(\boldsymbol{\mu}_P + \Sigma_{PI}\Sigma_{II}^{-1}(\mathbf{Z}_I(t) - \boldsymbol{\mu}_I), \Sigma_{P|I}\right). \tag{22}$$

Linking Eqs. (21) and (22), $\boldsymbol{\beta}_0 = \boldsymbol{\mu}_P - \Sigma_{PI}\Sigma_{II}^{-1}\boldsymbol{\mu}_I$, $\beta_1 = \Sigma_{PI}\Sigma_{II}^{-1}$, and the estimation of the regression coefficients thus requires the inversion of a partition of the sample covariance matrix. This linear forward-modeling framework formulation, with the proxy observations on the left hand side of the equation, has also been termed "indirect regression" (e.g., Christiansen et al., 2010).

The more widely-used treatment in the paleoclimate literature is a standard regression formulation that predicts the instrumental observations from the proxy observations,

$$\mathbf{Z}_I(t) = \boldsymbol{\beta}_0 + \beta_1 \mathbf{Z}_P(t) + \varepsilon_t \quad \text{where} \quad \varepsilon_t \sim \mathcal{N}\left(0, \Sigma_{I|P}\right). \tag{23}$$

As before, the conditional distribution of $\mathbf{Z}_I(t)$ given $\mathbf{Z}_P(t)$ is multivariate normal, and the estimation of the regression coefficients involves inverting a partition of the sample covariance matrix.

The basic issue with these two models [Eqs. (21) and (23)] is that in practice the problem is often severely underdetermined – the number of years for which both instrumental and proxy observations are available is short relative to the number of time series (this is known as the $p >> n$ problem in statistics, where p is the number of variables and n the number of observations). As a result, standard estimates of the covariance matrix Σ in Eq. (20) are singular and the inversion required to estimate the regression coefficients in either Eq. (21) or Eq. (23) does not exist. A related problem can arise if the number of years of overlap is not much larger than the number of time series, or if the predictor time series are highly correlated with one another. In either of these cases, the estimated covariance matrix is nearly singular, and its inversion unstable to small perturbations of the observed values. Any inference technique therefore requires some form of data reduction or regularization in the estimates of the regression coefficients. In addition, the pattern of missing data poses challenges as the data time series are generally of different lengths.

A number of techniques have been used or proposed to solve these technical difficulties in the context of reconstructing past climate. The list presented here is not exhaustive, and we do not discuss all possible variants and combinations of these ideas.

8.6.1. Principal component regression

Either the proxy matrix, the instrumental matrix, or both, is replaced by the corresponding leading principal components. With a smaller number of time series, the problem becomes over determined and standard ordinary least squares regression is applicable. As an example, the method of Mann et al. (1998) combines a principal component decomposition of the instrumental record with linear forward modeling [Eq. (21)] to relate each proxy time series to the retained principal components of the instrumental data (see, e.g., Lee et al., 2008, for a more detailed presentation). Luterbacher et al. (2004) considers the leading principal components of both the instrumental and proxy data sets, and uses the set of retained proxy principal components as predictors in regression models of the form Eq. (23) to separately predict each instrumental principal component. Inference in each case is via ordinary least squares.

The calculation of principal components for a data set (be it instrumental or proxy) generally requires that all time series cover the same set of years – that is, all records must be of the same length with no missing values. This is rarely the case in the context of paleoclimate reconstructions, and work-around solutions to this issue include first imputing missing instrumental values over the last 150 years and then calculating principal components of the instrumental data set (e.g., Rutherford et al., 2005; Mann et al., 2007), or the step-wise treatment of the proxy data in Rutherford et al. (2005).

8.6.2. Principal component regression with Bayesian inference

McShane and Wyner (2011) use the leading ten principal components of the proxy data matrix to infer the northern hemisphere annual mean temperature time series, which is modeled as an AR(2) process, and use Bayesian inference to fit the model. Posterior draws are then used to estimate both point-wise and path-wise credible intervals for the northern hemisphere mean temperature time series.

8.6.3. Canonical correlation analysis

This approach to climate reconstruction is discussed in Christiansen et al. (2009) and Smerdon et al. (2010), and is based on ideas presented in Barnett and Preisendorfer (1987). Consider an n

by r matrix of instrumental time series, Z_I, and an n by p matrix of proxy time series, Z_P. In the paleoclimate reconstruction context, this corresponds to r instrumental time series, p proxy time series, and an n year calibration interval. Canonical correlation analysis (CCA) proceeds by finding the length r vector $\mathbf{C}_{I,1}$ and length p vector $\mathbf{C}_{P,1}$ that maximize the correlation between $Z_I \cdot \mathbf{C}_{I,1}$ and $Z_P \cdot \mathbf{C}_{P,1}$. The vectors $\mathbf{C}_{I,1}$ and $\mathbf{C}_{P,1}$ are the first pair of canonical spatial patterns, while the inner products of each with the corresponding data matrix are the first pair of canonical time series. The second pair of canonical spatial patterns and time series are calculated in the same manner as the first, under the additional constraint that the second pair of canonical time series be orthogonal to the first pair. Following Barnett and Preisendorfer (1987), Smerdon et al. (2010) first project the matrices of proxy and instrumental time series onto the space spanned by the leading principal components, and then fit a regression model of the form Eq (23) via ordinary least squares to predict the leading canonical instrumental time series from the leading canonical proxy time series. Objective criteria are proposed for calculating the number of proxy and instrumental principal components to retain in the first step of the analysis, and the number of canonical time series to retain in the second.

8.6.4. RegEM (Schneider, 2001)

Discussed in Section 6.3 as a variant of the Expectation-Maximization algorithm of Dempster et al. (1977), RegEM has been widely applied to paleoclimate reconstruction problems (e.g., Rutherford et al., 2003, 2005; Zhang et al., 2004; Mann et al., 2005; Steig et al., 2009). RegEM seeks to impute missing instrumental observations by specifying a regression model of the from Eq. (23) for each year that features incomplete instrumental data. In order to estimate the regression coefficients, some form of regularization is necessary to ensure the existence or stability of the requisite matrix inverse. Both ridge regression (Hoerl and Kennard, 1970), also known as Tihkonov regression (Tikhonov and Arsenin, 1977) and truncated total least squares regression (van Huffel and Vandewalle, 1991; Fierro et al., 1997) have been used in this context. These two regularized regression techniques are both robust[6] to the presence of measurement errors in the predictor variables (Schneider, 2001), so in that sense RegEM is based on a more general model than the basic multivariate regression models of Eq. (21) or (23). In practice, however, the measurement error model is generally not explicitly specified (contrast with the explicit measurement error in model Eq. (15)), so we consider these tools as robust inference techniques rather than methods based on distinct modeling assumptions. As the two common regularized regressions used in RegEM are robust to the presence of measurement errors in the response variables, results should likewise be robust to the regression dilution effect (see Section 6.1).

Ridge regression involves inflating the diagonal of the correlation matrix of the predictor variables, and thus ensures the existence and stability of the required matrix inverse (for a Bayesian interpretation see, for example, Tingley and Huybers, 2010b, p. 2788). Ridge regression arises as a regularization technique when the predictor and response variables are subject to homogeneous errors (Golub et al., 1999; Schneider, 2001).

Truncated total least squares regression proceeds by first replacing the joint proxy-instrumental covariance matrix with a reduced-rank representation using only the eigenvectors corresponding to the k largest eigenvalues, where k is less than the number of predictors, and then using a pseudo-inverse in the estimation of the regression coefficients. If the number of retained eigenvectors is equal to the number of predictor variables, then the result is the standard (non-truncated) total least squares solution, which minimizes the mean-squared orthogonal distance from the data point to the line of best fit (Golub and van Loan, 1980; van Huffel and Vandewalle, 1991). Total least squares provides unbiased estimates of the regression coefficients if the predictor and response vectors each contain iid errors with equal variance (Golub and van Loan, 1980; Fierro et al., 1997), and a simple modification can result in unbiased estimates if the ratio of the variances is known.

Some applications of RegEM have involved first low band-pass filtering all data time series, applying a RegEM variant separately to the low and high frequency components, and then splicing together the results (e.g., Mann et al., 2005; Rutherford et al., 2005). While this hybrid RegEM technique has been shown to out perform non-hybrid implementations of RegEM in practice (see, for example Schmidt et al., 2011, including the supplementary material), we note that each of the two band-pass filtered data sets violate, by construction, the RegEM assumption that observations from subsequent years are independent of one another. Further research is required to understand why hybrid RegEM results in skillful reconstructions, despite the fact that it does not have a sound theoretical justification.

There continues to be vigorous debate in the literature concerning the best strategy for implementing and regularizing the EM algorithm in the context of paleoclimate reconstructions (e.g., Schneider, 2001; Mann et al., 2005, 2007; Rutherford et al., 2005, 2010; Christiansen et al., 2009, 2010). We stress here that this debate largely concerns inference tools: the fundamental statistical model in each case is Eq. (23), with the caveat that ridge regression and truncated total least square are robust to measurement errors in the predictor variables.

8.7. Multivariate regression methods versus hierarchical modeling

The regression methods for climate field reconstruction discussed in Section 8.6, and the related PPR (Section 8.5), certainly have strengths. For example, they are generally easier to implement and much faster, computationally, than fitting a hierarchical Bayesian model. Each of the methods provides a disciplined framework for fitting regressions in a high-dimensional setting when the problem at hand is often highly ill-posed. Indeed, it is worthwhile to consider the interconnections between the multivariate regression methods.

Regression using the leading principal components of the predictor variables is similar to ridge regression, in the sense that both methods differentially weight the principal components of the matrix of predictor variables. Ridge regression reduces the magnitudes of the regression coefficients corresponding to the principal components, with relatively higher shrinkage for those principal components corresponding to small eigenvalues. In contrast, principal component regression truncates the coefficients of the principal components with smaller variances to zero. Principal component regression, but not ridge regression, is thus a "sparsity inducing type" regularization as it truncates many coefficients to zero (see, e.g., Hastie et al., 2009, for more details). Although principal component regression sets coefficients to zero, it does not in general introduce structural zeros or sparsity after translating back to the original coordinates (unless the original predictors are orthogonal). It should also be noted that principal components are not robust to the presence of outliers in the data; indeed, principal components can be used to screen for such outliers (see, e.g., Chapter 10 of Jolliffe, 2002). Both ridge regression and principal component regression tend to give similar results and principal component regression inherits some of the attractive features of

[6] A *robust* estimator or inferential procedure is not unduly sensitive to violations of modeling assumptions.

ridge regression – including a bias-variance tradeoff to reduce mean-squared prediction error.

Regression based on CCA shares many of the properties of principal component regression, but uses a different algorithm for data reduction. The key difference is that principal component regression treats the predictor and response variables separately, whereas CCA is based on the cross-correlation between the response and predictor variables. Thus CCA is more applicable in scenarios which require data reduction on both the response and predictor variables. In particular, by performing data reduction based on the relation between the response and predictor data sets, CCA is a natural choice in a regression context (Mardia et al., 1979).

There are a number of weaknesses shared by these commonly used regression-based field reconstruction methods, many of which are avoided by the hierarchical models proposed in Section 3 combined with Bayesian inference.

8.7.1. Biases

Each of these methods produces biased estimates of the regression coefficients, even if the underlying model is correct. This cannot be viewed as a major shortcoming, however, given that allowing a small amount of bias can result is substantial reductions in the mean-squared error (e.g., Hoerl and Kennard, 1970). In addition, similar biases can result from using Bayesian inference, where prior distributions provide regularization. The advantage of Bayesian inference is the transparency which results from considering prior specification as part of model construction: the extent of regularization is clearly stated, while conjugate priors can often be interpreted in terms of an equivalent number of additional observations with certain properties (see, for example, Gelman et al., 2003, p. 51).

8.7.2. Treatment of missing data

With the exception of the RegEM variants, the regression-based methods are only readily applicable to data sets that are block missing – that is, all proxy time series are of the same length, with no missing values, and likewise for the instrumental data set. Various work-around solutions to this problem, such as an initial application of RegEM to impute the missing values in the instrumental data, severely complicate the propagation of uncertainties. The hierarchical modeling framework, in contrast, is amenable to changing patterns of missing data (see Section 8.1 and Tingley and Huybers, 2010a).

8.7.3. Conflation of data and process

By considering all linear relationships between proxy and instrumental time series, regardless of geographical location, the methods discussed in Section 8.6 confound the process-level model with the data-level model. We stress that doing so deeply violates scientific understanding of the data and climate system. A tree may be predictive of climate at a distant location, but a tree is causally impacted by strictly local climate. It is the climate system itself that displays long-range dependencies. PPR (Section 8.5 Cook et al., 1999) accounts for the regional nature of drought variability by allowing tree ring chronologies to act as predictors of PDSI over a some user defined distance, and likewise confounds the data model with the climate process. Inference models with distinct process and data levels more appropriately reflect scientific understanding of the observations, the climate system, and the relationship between the two.

8.7.4. Temporal and spatial covariance modeling

The methods described in Section 8.6, as generally implemented, do not consider temporal structure in the data time series, nor do the geographical coordinates of the time series enter into the calculations. While various alternatives to least squares regression have been employed in paleoclimate research to account for temporal structure, such as the Cochrane–Orcutt algorithm used in Thejll and Schmith (2005) and the "prewhitening" procedure used in Cook et al. (1999), these approaches lack the flexibility and transparency of assumptions which are the hallmarks of hierarchical modeling. In terms of spatial structure, results from methods discussed in Section 8.6 are unchanged if the locations of the data time series are randomized prior to the analysis, and then unscrambled afterwards.

In contrast, the process level of a hierarchical model can explicitly model both temporal and spatial structure in the target climate process. By specifying a parametric spatiotemporal process level (Section 4), hierarchical models allow for imputation of the field at any location and for forecasting to times for which there are no observations (compare with Lee et al., 2008), while the estimated process-level parameters inform the distance and time over which such imputations are reasonably well constrained (compare with Cook et al., 1999). In addition, the data level accounts for the different temporal and spatial supports of the various data sets, so that the differing spatial and temporal covariance structures of the data sets can be accounted for, conditional on the covariance structure of the climate process itself (see Fig. 3 and Sections 4.2 & 5.2).

8.7.5. Uncertainty estimates

The natural end product of any multivariate linear regression-based method is a point estimate and uncertainty measure for each missing observation. This information cannot be readily used to quantify the uncertainty for non-linear functions of the inferred climate process, such as the probability that one interval featured higher values of the climate field than another, a greater rates of change, or higher variability. Bayesian inference results in a number of posterior draws of the climate field (conditional on the data and modeling assumptions), and such ensemble-based reconstructions are more appropriate for both characterizing natural variability and assessing the extent to which recent climate is anomalous with respect to that of the past. Bayesian inference is certainly not the only option for producing ensembles – as an example, Li et al. (2007b) produce- an ensemble using frequentist inference combined with a bootstrap.

8.7.6. Relative performance

Finally, there is no consensus as to which of the regression-based reconstruction methods discussed in Section 8.6 performs best in a global context (e.g., Lee et al., 2008; Christiansen et al., 2009, 2010; Rutherford et al., 2010; Emile-Geay et al., submitted for publication-a, submitted for publication-b). Each method requires the specification or estimation of one or more parameters that determine the strength of the regularization (e.g., the number of principal components to retain in principal component regression, or the ridge parameter in ridge-regularized RegEM), and the choice of regularization parameter(s) is not always clear. Much of the debate over the relative merits of these regularized regression methods concerns the details of the regularization scheme rather than the underlying model choices (e.g., Christiansen et al., 2009, 2010; Rutherford et al., 2010; Emile-Geay et al., submitted for publication-a, submitted for publication-b).

Hierarchical models allow for the inclusion of more scientific knowledge about both the data and underlying process, and require that the modeling assumptions at both of these level be clearly stated. Bayesian data analysis then provides a logically sound framework for performing the inference. It is our hope that a more wide-spread adoption of this approach by the paleoclimate community will produce more substantive research into the underlying science.

9. Discussion

Reconstructing past climate through space and time is a difficult endeavor. Harnessing the scientific understanding of both the climate system and the various proxies together with the tools of modern statistical science has the potential to substantially increase our understanding of past climate. Statisticians require scientific knowledge in order to construct models that are properly tailored to the particular characteristics of the paleoclimate reconstruction problem. Scientists, in turn, can benefit from the careful and efficient treatment of uncertainty offered by statisticians. By describing the paleoclimate reconstruction in terms of hierarchical statistical models, and indicating the tools available for building and fitting such models, we hope to spark greater collaborations between the climate and statistical sciences.

Hierarchical statistical modeling offers a logically consistent framework for tackling the paleoclimate reconstruction problem, and there are a number of advantages to this framework. The multilevel construction clarifies that the relationship between the data and the target process is distinct from the covariance structure of the process. Likewise, a focus on model construction clarifies the distinction between model assumptions and the techniques used to perform the inference. By compartmentalizing the reconstruction problem, hierarchical models allow each component of the model to be constructed independently of the others. As an example, scientifically realistic forward models can be developed for the commonly used proxy types independently of one another and of the model that describes the space-time structure of the target climate process. Spatial and temporal misalignment (Fig. 1) and change-of-support (Fig. 3), as well as the missing data mechanism, can be modeled in the level of the model that makes most sense given the particular analysis – either as part of the forward model (Section 5) or at the data level (Section 6). A secondary advantage of the hierarchical approach is that different components of the model can be separately tested before they are incorporated into the hierarchy (e.g., Craigmile et al., 2009), which may be especially useful when incorporating multiple proxy processes.

Hierarchical models provide a cohesive framework for propagating uncertainty through an analysis. However, these models only account for known uncertainties and like all statistical techniques, can yield poor results if the analysis model is misspecified. Model checking and validation is an important part of any statistical analysis, which was not covered in this article; see Gelman et al. (2003) for a discussion of these issues in a Bayesian context. Draper (1995) discusses how to assess and propagate model uncertainty.

Pseudoproxy experiments have been employed extensively in the paleoclimate literature (e.g., Gonzalez-Rouco et al., 2003; von Storch et al., 2004; Hegerl et al., 2007; Mann et al., 2007; Christiansen et al., 2009; Li et al., 2010; Smerdon et al., 2010; Christiansen and Ljungqvist, 2011). These experiments generally proceed by sub-sampling and corrupting the output from a climate model, using this limited data set to infer some quantity (the temperature field, for example), and then comparing results and uncertainty estimates to the original model output. The results of such experiments are useful in assessing the appropriateness of the uncertainty estimates derived from a reconstruction method, and exploring if the statistical model underlying that method is appropriate. As an example, to the extent that the climate model and pseudoproxies are realistic representations of the actual climate system and proxy data, differences between the actual and nominal coverage rates of uncertainty intervals for the estimates of the withheld climate model data could indicate that the statistical model used to infer these withheld values is not well specified (e.g., Li et al., 2010; Tingley and Huybers, 2010b).

The modeling formalism developed in the preceding sections is intentionally general, and we have gone to some length to detail just how rich the modeling choices can be. The price for such complicated models is paid in terms of the effort required to perform the inference, and we have discussed various (newer) methodological developments that may be necessary to do so. Simple space-time separable models can in certain scenarios result in excellent predictive inferences, even in cases where the underlying process is very much non-separable (e.g., Genton, 2007) – as is likely the case for climate variables (see Fig. 4). In general, fitting a more involved model will require inference on additional process-level parameters, and with a fixed amount of data, the uncertainty associated with the inference on these parameters will increase. Determining the level of complexity required at each level of the hierarchical model in order to optimally infer past climate from proxy data is an open question, which we will explore in future work.

The framework presented here can be readily applied to the reconstruction of multivariate climate processes, such as the joint temperature-precipitation process. The key challenges are, as in the univariate case, the specification of the joint space-time covariance structure of the multivariate process – which may be very complicated – and the conditional distribution of the data types given the multivariate process. While the univariate and multivariate reconstruction problems are formally equivalent, the latter is both more challenging and more scientifically defensible. At the process level, it becomes necessary to account for the interaction between the processes in both space and time. At the data level, forward models can then reflect the scientific knowledge that many natural proxies (tree, coral, and pollen observations, for example) are influenced by more than one climate variable.

More generally, we note that specific model construction choices should reflect the specific scientific question under investigation. For example, if the goal is to understand the relationship between the temperature field and estimates of green house gas, solar, and volcanic forcings, then the time series of estimated forcings should be included in the hierarchy. Inference on the process-level parameters that describe the connection between the forcings and the climate then provide estimates of the sensitivity of the climate system to changes in those forcings. Note that withholding the forcings and comparing them to an independent reconstruction prevents the consistent propagation of errors in the estimate of the link between the two. Alternatively, there may be scientific interest in which of two well-characterized and understood spatial patterns was likely dominant as a function of time – for example, if the proxies indicate that the temperature field for a particular year was more indicative of La Niña or El Niño like conditions. In this case, a process-level model that includes a mean structure similar to the form of Eq. (9), but which allows for a time dependent switching between two known spatial patterns, may be appropriate:

$$\boldsymbol{\mu}_t = \beta_t \mathbf{V}_1 + (1-\beta_t)\mathbf{V}_2, \quad \beta_t \in \{0,1\}, \quad t \in \mathcal{T}. \tag{24}$$

Inference on the time series of β_t then indicates which pattern was likely to have been dominant for each year.

Determining the extent to which climate during the post-industrial era is anomalous or extreme, either in value or rate of change, with respect to climate over the last thousand (or more) years is often a goal of paleoclimate reconstructions. Answering these question requires that both the proxy data and the methods used to reconstruct past climate accurately capture the tail behavior of the probability distribution of climate. Modeling the tail behavior of both the instrumental and proxy data using the generalized extreme value distribution (e.g., Coles, 2001) can potentially determine the extent to which the proxies capture climate extremes, and the extent to which the distribution of climate

extremes has changed through time. This is an active area of research – see Mannshardt-Shamseldin et al. (submitted for publication), which considers the extremal behavior of proxy series directly and explores possible distributional shifts in the extremes of proxy observations as a function of time and space.

Over the last decade, a vast amount of intellectual capital has been spent on developing, testing, and comparing reconstruction methods that share a common statistical model, but differ in the tools used to conduct the inference. The literature on this topic is too vast to cite in its entirety. We simply note that from 1998 (Mann et al., 1998) to 2009 (Christiansen et al., 2009) all methods proposed, studied, or used to reconstruct the surface temperature process (as opposed to some large-scale spatial average) from annually resolved proxy data fall under the umbrella of the regression-based models discussed in Section 8.6. Devoting similar resources to the development of scientifically richer models has the potential to vastly improve our understanding of past climate and address numerous societally relevant questions. It is our hope that the next decade will witness a rich debate in the paleoclimate literature over scientifically motivated forward models, and how to best model the target climate process at the process level.

There is much that climate science can learn from statistical science. In this article we have only skimmed the surface of the possible models and modeling strategies that can be employed. Space-time statistical modeling is currently an active area of research, with clear applications to the analysis of climate data in general and the reconstruction problem in particular. In turn, there is much that statisticians can learn from climate scientists with regards to both the instrumental and proxy observations, and the structure and characteristics of climate processes. We hope that this article both encourages and facilitates future collaborations between these two communities.

Acknowledgments

We thank SAMSI and the organizers of the 2010–2011 Program on Space-time modeling for Epidemiology, Climate Change, and Environmental Mapping for making this collaboration possible. The manuscript benefited from conversations with Bo Christiansen, John Haslett, Cindy Greenwood, Michael Evans, Matthew Schofield, and the participation of MPT, PFC, and EM-S in the 11th International Meeting for Statistical Climatology. We thank Noel Cressie, Julien Emile-Geay, Michael Mann, Douglas Nychka, Tapio Schneider, Eugene Wahl, and five referees for comments that improved both the content and presentation of this manuscript. MPT is supported by the National Science Foundation under award number CMG-0724828. PFC is supported by the National Science Foundation under award numbers DMS-0604963 and DMS-0906864. MH is partially supported by the US Geological Survey (USGS-CDI) and the National Science Foundation (NSF-HSD). BL is partially supported by the National Science Foundation under award number DMS-1007686. BR is partially supported by the National Science Foundation under awards numbers DMS-0906392, AGS-1003823, SU-WI-EVP10, and SUFSC08-SUFSC10-SMSCVISG0906.

References

Adam, D., West, G., 1983. Temperature and precipitation estimates through the last glacial cycle from Clear Lake, California, pollen data. Science 219, 168.
Ammann, C., Genton, M., Li, B., 2010. Technical Note: Correcting for signal attenuation from noisy proxy data in climate reconstructions. Climate of the Past 6, 273–279.
Auestad, B.H., Shumway, R.H., Tjøstheim, D., Verosub, K.L., 2008. Linear and nonlinear alignment of time series with applications to varve. Environmetrics 19, 409–427.
Banerjee, S., 2005. On geodetic distance computations in spatial modeling. Biometrics 61, 617–625.
Banerjee, S., Carlin, B.P., Gelfand, A.E., 2004. Hierarchical Modeling and Analysis for Spatial Statistics. Chapman & Hall/CRC, New York.
Barnett, T., Preisendorfer, R., 1987. Origins and levels of monthly and seasonal forecast skill for united states surface air temperatures determined by canonical correlation analysis. Monthly Weather Review 115, 1825–1850.
Bellone, E., Hughes, J., Guttorp, P., 2000. A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. Climate Research 15, 1–12.
Berger, J., 1994. An overview of robust Bayesian analysis (with discussion). Test 3, 5–124.
Berger, J., 2006. The case for objective Bayesian analysis. Bayesian Analysis 1, 385–402.
Berliner, L., Wikle, C., Cressie, N., 2000a. Long-lead prediction of Pacific SSTs via Bayesian dynamic modeling. Journal of Climate 13, 3953–3968.
Berliner, L.M., Levine, R., Shea, D., 2000b. Bayesian climate change assessment. Journal of Climate 13, 3805–3820.
Berrocal, V.J., Raftery, A.E., Gneiting, T., 2008. Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. Annals of Applied Statistics 2, 1170–1193.
Box, G., 1980. Sampling and Bayes' inference in scientific modelling and robustness. Journal of the Royal Statistical Society: Series A (Statistics in Society) 143, 383–430.
Briffa, K., Jones, P., Bartholin, T., Eckstein, D., 1992. Fennoscandian summers from AD 500: temperature changes on short and long timescales. Climate Dynamics 7, 111–119.
Brockwell, P.J., Davis, R.A., 2002. Introduction to Time Series and Forecasting, second ed. Springer, New York.
Brohan, P., Kennedy, J.J., Harris, I., Tett, S.F.B., Jones, P.D., 2006. Uncertainty estimates in regional and global observed temperature changes: a new data set from 1850. Journal of Geophysical Research 2, 99–113.
Brooks, S., Gelman, A., Jones, G., Meng, X., 2011. Handbook of Markov Chain Monte Carlo. Chapman & Hall/CRC.
Brynjarsdóttir, J., Berliner, L.M., 2011. Bayesian hierarchical modeling for paleoclimate reconstruction from geothermal data. The Annals of Applied Statistics 5, 1328–1359.
Bürger, G., 2007. Comment on "The spatial extent of 20th-century warmth in the context of the past 1200 years, " by TJ Osborn and KR Briffa. Science 316, 1844.
Caragea, P.C., 2003. Approximate likelihoods for spatial processes, Ph.D. Dissertation. Technical Report. University of North Carolina, Chapel Hill, NC. Department of Statistics.
Carlin, B., Louis, T., 2009. Bayesian Methods for Data Analysis. Chapman & Hall/CRC.
Carroll, R., Ruppert, D., 1996. The use and misuse of orthogonal regression in linear errors-in-variables models. The American Statistician 50, 1–6.
Carroll, R., Ruppert, D., Stefanski, L., Crainiceanu, C., et al., 2006. Measurement Error in Nonlinear Models: A Modern Perspective, second edition. Chapman & Hall, Boca Raton, Florida.
Carslaw, H., Jaeger, J., 1959. Conduction of Heat in Solids. Oxford University Press, USA.
Christiansen, B., 2011a. Reconstructing the NH mean temperature: Can underestimation of trends and variability be avoided? Journal of Climate 24, 674–692.
Christiansen, B., 2011b. Reply to "Comments on. by Tingley and Li. Journal of Climate (Under review; currently available online at web.dmi.dk/fsweb/solar-terrestrial/staff/boc/loc reply.pdf).
Christiansen, B., Ljungqvist, F., 2011. Reconstruction of the extra-tropical nh mean temperature over the last millennium with a method that preserves low-frequency variability. Journal of Climate 24, 6013–6034.
Christiansen, B., Schmith, T., Thejll, P., 2009. A surrogate ensemble study of climate reconstruction methods: Stochasticity and Robustness. Journal of Climate 22, 951–976.
Christiansen, B., Schmith, T., Thejll, P., Christiansen, B., 2010. Reply to Comments on "A surrogate ensemble study of climate reconstruction methods: Stochasticity and Robustness" by Rutherford et al. Journal of Climate 23, 2839–2844.
Coles, S., 2001. An Introduction to Statistical Modeling of Extreme Values. Springer, New York.
Cook, E., Meko, D., Stahle, D., Cleaveland, M., 1999. Drought reconstructions for the continental united states. Journal of Climate 12, 1145–1162.
Craigmile, P.F., Calder, C.A., Paul, R., Li, H., Cressie, N., 2009. Hierarchical model building, fitting, and checking: A behind-the-scenes look at a Bayesian analysis of arsenic exposure pathways with discussion. Bayesian Analysis 4.
Cressie, N., Johannesson, G., 2008. Fixed rank kriging for very large spatial data sets. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70, 209–226.
Cressie, N., Tingley, M.P., 2010. Discussion of "The value of multi-proxy reconstruction of past climate" by Bo Li, Dougles W. Nychka, and Caspar M. Ammann. Journal of the American Statistical Association 105, 895–900.
Cressie, N.A., 1993. Statistics for Spatial Data, second ed. John Wiley & Sons, New York.
D'Arrigo, R., Wilson, R., Liepert, B., Cherubini, P., 2008. On the `Divergence Problem' in Northern Forests: A review of the tree-ring evidence and possible causes. Global and Planetary Change 60, 289–305.
Das, B., 2000. Global covariance modeling: a deformation approach to anisotropy. Ph.D. thesis. Department of Statistics, University of Washington. Seattle, WA.
Dempster, A., Laird, N., Rubin, D., et al., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 39, 1–38.
Draper, D., 1995. Assessment and propagation of model uncertainty (with discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology) 57, 45–97.

Emile-Geay, J., Cobb, K., Mann, M., Rutherford, S., Wittenberg, A.T. Estimating tropical pacific SST variability over the past millennium. Part 1: Methodology and validation. Journal of Climate, submitted for publication-a; currently available at: http://college.usc.edu/labs/jeg/publications/.

Emile-Geay, J., Cobb, K., Mann, M., Rutherford, S., Wittenberg, A.T. Estimating tropical pacific SST variability over the past Millennium. Part 2: Reconstructions and uncertainties. Journal of Climate, submitted for publication-b; currently available at: http://college.usc.edu/labs/jeg/publications/.

Esper, J., Cook, E.R., Schweingruber, F.H., 2002. Low frequency signals in long tree-ring chronologies for reconstructing past temperature variability. Science 295, 2250–2253.

Evans, M., Reichert, B., Kaplan, A., Anchukaitis, K., Vaganov, E., Hughes, M., Cane, M., 2006. A forward modeling approach to paleoclimatic interpretation of tree-ring data. Journal of Geophysical Research 111, 3008.

Fierro, R., Golub, G., Hansen, P., OLeary, D., 1997. Regularization by truncated total least squares. SIAM Journal on Scientific Computing 18, 1223–1241.

Finley, A., Banerjee, S., Carlin, B., 2007. spBayes: an R package for univariate and multivariate hierarchical point-referenced spatial models. Journal of Statistical Software 19, 1–20.

Flegal, J., Haran, M., Jones, G., 2008. Markov chain Monte Carlo: Can we trust the third significant figure? Statistical Science 23, 250–260.

Frame, D.J., Booth, B.B.B., Kettleborough, J.A., Stainforth, D.A., Gregory, J.M., Collins, M., Allen, M.R., 2005. Constraining climate forecasts: The role of prior assumptions. Geophysical Research Letters 32, L09702.

Frost, C., Thompson, S., 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.

Fuentes, M., 2006. Testing for separability of spatial-temporal covariance functions. Journal of Statistical Planning and Inference 136, 447–466.

Fuentes, M., 2007. Approximate likelihood for large irregularly spaced spatial data. Journal of the American Statistical Association 102, 321–331.

Fuller, W., 1987. Measurement Error Models. Wiley, New York, NY.

Furrer, R., Genton, M., Nychka, D., 2006. Covariance tapering for interpolation of large spatial datasets. Journal of Computational and Graphical Statistics 15, 502–523.

Gelfand, A., Zhu, L., Carlin, B., 2001. On the change of support problem for spatio-temporal data. Biostatistics 2, 31–45.

Gelman, A., 2006. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis 1, 515–533.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2003. Bayesian Data Analysis, second edition. Chapman & Hall/CRC, Boca Raton.

Genton, M.G., 2007. Separable approximations of space-time covariance matrices. Environmetrics 18, 681–695.

Gilks, W., Gilks, W., Richardson, S., Spiegelhalter, D., 1996. Markov Chain Monte Carlo in Practice. Chapman & Hall/CRC.

Gneiting, T., Genton, M.G., Guttorp, P., 2007. Geostatistical space-time models, stationarity, separability and full symmetry. In: Isham, V., Finkelstadt, B., Härdle, W. (Eds.), Statistics of Spatio-Temporal Systems. Taylor and Francis, Boca Raton, pp. 151–176.

Golub, G., Hansen, P., O'Leary, D., 1999. Tikhonov regularization and total least squares. SIAM Journal on Matrix Analysis And Applications 21, 185–194.

Golub, G., van Loan, C., 1980. An analysis of the total least squares problem. SIAM Journal on Numerical Analysis 17, 883–893.

Gonzalez-Rouco, F., Storch, H.V., Zorita, E., 2003. Deep soil temperature as proxy for surface air-temperature in a coupled model simulation of the last thousand years. Geophysical Research Letters 30, 2116.

Guiot, J., Wu, H., Garreta, V., Hatté, C., Magny, M., 2009. A few prospective ideas on climate reconstruction: from a statistical single proxy approach towards a multi-proxy and dynamical approach. Climate of the Past 5, 571–583.

Haam, E., Huybers, P., 2010. A test for the presence of covariance between time-uncertain series of data with application to the Dongge Cave speleothem and atmospheric radiocarbon records. Paleoceanography 25, PA2209.

Hansen, J., Ruedy, R., Sato, M., Lo, K., 2010. Global surface temperature change. Reviews of Geophysics 48, RG4004.

Harris, R., 2007. Variations in air and ground temperature and the POM-SAT model: results from the Northern Hemisphere. Climate of the Past 3, 611–621.

Harris, R., Chapman, D., 2001. Mid-latitude (30°–60° N) climatic warming inferred by combining borehole temperatures with surface air temperatures. Geophysical Research Letters 28, 747–750.

Hartmann, D.L., 1994. Global Physical Climatology. Academic Press, New York.

Harvill, J., Ray, B., 2006. Functional coefficient autoregressive models for vector time series. Computational Statistics and Data Analysis 50, 3547–3566.

Haslett, J., Parnell, A., Salter-Townsend, M., 2006a. Modelling temporal uncertainty in palaeoclimate reconstructions, in: Proceedings of the 21st International Workshop on Statistical Modelling, pp. 26–37.

Haslett, J., Whiley, M., Bhattacharya, S., Salter-Townshend, M., Wilson, S., Allen, J., Huntley, B., Mitchell, F., 2006b. Bayesian palaeoclimate reconstruction. Journal of the Royal Statistical Society: Series A (Statistics in Society) 169, 395–438.

Hasselmann, K., 1993. Optimal fingerprints for the detection of time-dependent climate change. Journal of Climate 6, 1957–1971.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining. Inference and Prediction, second ed. Springer, New York.

Hegerl, G., Crowley, T., Allen, M., Hyde, W., Pollack, H., Smerdon, J., Zorita, E., 2007. Detection of human influence on a new, validated 1500-year temperature reconstruction. Journal of Climate 20, 650–666.

Higdon, D., 1998. A process-convolution approach to modelling temperatures in the North Atlantic Ocean. Environmental and Ecological Statistics 5, 173–190.

Hoerl, A., Kennard, R., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67.

Huang, H.C., Cressie, N., Gabrosek, J., 2002. Fast, resolution-consistent spatial prediction of global processes from satellite data. Journal of Computational and Graphical Statistics 11, 63–88.

van Huffel, S., Vandewalle, J., 1991. The Total Least Squares Problem: Computational Aspects and Analysis. Society for Industrial and Applied Mathematics, Philadelphia.

Hughes, M., Ammann, C., 2009. The future of the past – an earth system framework for high resolution paleoclimatology: editorial essay. Climatic Change 94, 247–259.

Ihaka, R., Gentleman, R., 1996. R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics 5, 299–314.

Jansen, E., Overpeck, J., Briffa, K., 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., Zhang, D., 2007. Palaeoclimate. In: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K., Tignor, M., Miller, H. (Eds.), Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA (Chapter 6).

Jolliffe, I., 2002. Principal Component Analysis. Springer Verlag, New York.

Jones, P., Briffa, K., Osborn, T., Lough, J., van Ommen, T., Vinther, B., Luterbacher, J., Wahl, E., Zwiers, F., Mann, M., et al., 2009. High-resolution palaeoclimatology of the last millennium: a review of current status and future prospects. The Holocene 19, 3–49.

Jones, R., 1963. Stochastic processes on a sphere. Annals of Mathematical Statistics 34, 213–218.

Kass, R., Tierney, L., Kadane, J., 1989. Approximate methods for assessing influence and sensitivity in Bayesian analysis. Biometrika 76, 663.

Kass, R., Wasserman, L., 1996. The selection of prior distributions by formal rules. Journal of the American Statistical Association 91, 1343–1370.

Kaufman, C., Schervish, M., Nychka, D., 2008. Covariance Tapering for Likelihood-Based Estimation in Large Spatial Data Sets. Journal of the American Statistical Association 103, 1545–1555.

Kaufman, D., Schneider, D., McKay, N., Ammann, C., Bradley, R., Briffa, K., Miller, G., Otto-Bliesner, B., Overpeck, J., Vinther, B., et al., 2009. Recent warming reverses long-term arctic cooling. Science 325, 1236.

Kopp, R., Simons, F., Mitrovica, J., Maloof, A., Oppenheimer, M., 2009. Probabilistic assessment of sea level during the last interglacial stage. Nature 462, 863–867.

Lahiri, S., 2003. Resampling Methods for Dependent Data. Springer Verlag, New York.

Le, N., Zidek, J., 2006. Statistical Analysis of Environmental Space-time Processes. Springer, New York.

Lee, T., Zwiers, F., Tsao, M., 2008. Evaluation of proxy-based millennial reconstruction methods. Climate Dynamics 31, 263–281.

Leith, N., Chandler, R., 2010. A framework for interpreting climate model outputs. Journal of the Royal Statistical Society: Series C (Applied Statistics) 59, 279–296.

Li, B., Eriksson, M., Srinivasan, R., Sherman, M., 2008. A geostatistical method for Texas NexRad data calibration. Environmetrics 19, 1–19.

Li, B., Genton, M., Sherman, M., 2007a. A nonparametric assessment of properties of space-time covariance functions. Journal of the American Statistical Association 102, 736–744.

Li, B., Murthi, A., Bowman, K.P., North, G.R., Genton, M.G., Sherman, M., 2009. Statistical tests of Taylor's hypothesis: an application to precipitation fields. Journal of Hydrometeorology 10, 254–265.

Li, B., Nychka, D., Ammann, C., 2007b. The 'hockey stick' and the 1990s: a statistical perspective on reconstructing hemispheric temperatures. Tellus A 59, 591–598.

Li, B., Nychka, D., Ammann, C., 2010. The value of multi-proxy reconstruction of past climate. Journal of the American Statistical Association 105, 883–911.

Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data, second ed. Wiley, New York.

Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D., 2000. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing 10, 325–337.

Luterbacher, J., Dietrich, D., Xoplaki, E., Grosjean, M., Wanner, H., 2004. European seasonal and annual temperature variability, trends, and extremes since 1500. Science 303, 1499–1503.

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

Mann, M., Bradley, R., Hughes, M., 1999. Northern hemisphere temperatures during the past millennium: inferences, uncertainties, and limitations. Geophysical Research Letters 26, 759–762.

Mann, M., Jones, P., 2003. Global surface temperatures over the past two millennia. Geophysical Research Letters 30, 1820.

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

Mann, M., Rutherford, S., Wahl, E., Ammann, C., 2007. Robustness of proxy-based climate field reconstruction methods. Journal of Geophysical Research 112, D12109.

Mann, M.E., Zhang, Z., Hughes, M.K., Bradley, R.S., Miller, S.K., Rutherford, S., Ni, F., 2008a. 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, Z., Hughes, M.K., Bradley, R.S., Miller, S.K., Rutherford, S., Ni, F., 2008b. Supporting information for "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. http://www.pnas.org/content/suppl/2008/09/02/0805721105.DCSupplemental/0805721105SI.pdf.
Mannshardt, E., Craigmile, P., Tingley, M. Statistical modeling of extreme value behavior in North American tree-ring density series, submitted for publication.
Mardia, K., Kent, J., Bibby, J., 1979. Multivariate Analysis. Academic Press, San Diego.
McShane, B.B., Wyner, A.J., 2011. A statistical analysis of multiple temperature proxies: Are reconstructions of surface temperatures over the last 1000 years reliable? Annals of Applied Statistics 5, 5–44 See also the accompanying editorial by Michael Stein, discussions, and rejoinder.
Melvin, T., Briffa, K., 2008. A "signal-free" approach to dendroclimatic standardisation. Dendrochronologia 26, 71–86.
Mitchell, M., Genton, M., Gumpertz, M., 2005. Testing for separability of space-time covariances. Environmetrics 16, 819–831.
Moberg, A., Sonechkin, D., Holmgren, K., Datsenko, N., Karlén, W., 2005. Highly variable northern hemisphere temperatures reconstructed from low-and high-resolution proxy data. Nature 433, 613–617.
Neumaier, A., Schneider, T., 2001. Estimation of parameters and eigenmodes of multivariate autoregressive models. ACM Transactions on Mathematical Software (TOMS) 27, 27–57.
NRC, 2006. Surface Temperature Reconstructions for the Last 2000 Years. The National Academies Press, Washington, D.C.
O'Hagan, A., Buck, C., Daneshkhah, A., Eiser, J., Garthwaite, P., Jenkinson, D., Oakley, J., Rakow, T., 2006. Uncertain Judgements: Eliciting Experts' Probabilities. John Wiley & Sons.
Ohlwein, C., Wahl, E.R., 2012. Review of probabilistic pollen-climate transfer methods. Quaternary Science Reviews 31, 17–29.
Osborn, T.J., Briffa, K.R., 2006. The spatial extent of 20th-century warmth in the context of the past 1200 years. Science 311, 841–844.
Rajaratnam, B., 2010. High Dimensional Multiproxy Paleoclimate Reconstructions: New Perspectives. Technical Report. Department of Statistics, Stanford University.
Ribeiro, P., Diggle, P., 2001. geoR: A package for geostatistical analysis. R News 1, 14–18.
Robert, C., Casella, G., 2004. Monte Carlo Statistical Methods. Springer Verlag, New York.
Roe, G., Baker, M., 2007. Why is climate sensitivity so unpredictable? Science 318, 629–632.
Rubin, D., 1976. Inference and missing data (with discussion). Biometrika 63, 581–592.
Rutherford, S., Mann, M., Delworth, T., Stouffer, R., 2003. Climate Field Reconstruction under Stationary and Nonstationary Forcing. Journal of Climate 16, 462–479.
Rutherford, S., Mann, M., Osborn, T., Bradley, R., Briffa, K., Hughes, M., Jones, P., 2005. Proxy-based northern hemisphere surface temperature reconstructions: sensitivity to method, predictor network, target season, and target domain. Journal of Climate 18, 2308–2329.
Rutherford, S.D., Mann, M.E., Ammann, C.M., Wahl, E.R., 2010. Comments on "A Surrogate Ensemble Study of Climate Reconstruction Methods: Stochasticity and Robustness, by Christiansen et al". Journal of Climate 23, 2832–2838.
Sahu, S., Mardia, K., 2005. Recent trends in modeling spatio-temporal data. In: Proceedings of the Special Meeting on Statistics and Environment, pp. 69–83.
Santner, T.J., Craigmile, P.F., Calder, C.A., Paul, R., 2008. Demographic and behavioral modifiers of arsenic exposure pathways: A Bayesian hierarchical analysis of NHEXAS data. Environmental Science and Technology 42, 5607–5614.
Schmidt, G.A., Mann, M.E., Rutherford, S.D., 2011. A Comment on "A Statistical Analysis of Multiple Temperature Proxies: Are Reconstructions of Surface Temperatures over the Last 1000 Years Reliable?" by McShane and Wyner. Annals of Applied Statistics 5, 65–70.
Schneider, T., 2001. Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. Journal of Climate 14, 853–871.
Schneider, T., Neumaier, A., 2001. Algorithm 808: Arfit—a matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models. ACM Transactions on Mathematical Software (TOMS) 27, 58–65.
Schofield, M., in prep. Climate reconstruction using tree-ring data.
Smerdon, J., Kaplan, A., Chang, D., 2008. On the origin of the standardization sensitivity in RegEM climate field reconstructions. Journal of Climate 21, 6710–6723.
Smerdon, J., Kaplan, A., Chang, D., 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.
Smith, B.J., Yan, J., Cowles, M.K., 2008. Unified geostatistical modeling for data fusion and spatial heteroskedasticity with R package RAMPS. Journal of Statistical Software 25, 91–110.
Song, Y., Wikle, C., Anderson, C., Lack, S., 2007. Bayesian estimation of stochastic parameterizations in a numerical weather forecasting model. Monthly Weather Review 135, 4045–4059.
Steig, E., Schneider, D., Rutherford, S., Mann, M., 2009. Warming of the antarctic ice-sheet surface since the 1957 international geophysical year. Nature 457, 459–462.
Stein, M., 2005. Statistical methods for regular monitoring data. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 667–687.
Stein, M.L., Chi, Z., Welty, L.J., 2004. Approximating likelihoods for large spatial data sets. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66, 275–296.
Stidd, C., 1953. Cube-root-normal precipitation distributions. Transactions. American Geophysical Union 34, 31–35.
von Storch, H., Zorita, E., Jones, J., Dimitriev, Y., González-Rouco, F., Tett, S., 2004. Reconstructing past climate from noisy data. Science 306, 679–682.
Stouffer, R., Hegerl, G., Tett, S., 2000. A comparison of surface air temperature variability in three 1000-yr coupled ocean-atmosphere model integrations. Journal of Climate 13, 513–537.
Tebaldi, C., Smith, R., Nychka, D., Mearns, L., 2005. Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the analysis of multimodel ensembles. Journal of Climate 18, 1524–1540.
ter Braak, C., 1995. Non-linear methods for multivariate statistical calibration and their use in palaeoecology: a comparison of inverse (k-nearest neighbours, partial least squares and weighted averaging partial least squares) and classical approaches. Chemometrics and Intelligent Laboratory Systems 28, 165–180.
Thejll, P., Schmith, T., 2005. Limitations on regression analysis due to serially correlated residuals: Application to climate reconstruction from proxies. Journal of Geophysical Research 110, D18103.
Tikhonov, A., Arsenin, V., 1977. Methods for Solving Ill-Posed Problems. Wiley, New York.
Tingley, M., Huybers, P., 2010a. A Bayesian algorithm for reconstructing climate anomalies in space and time. Part 1: Development and applications to paleoclimate reconstruction problems. Journal of Climate 23, 2759–2781.
Tingley, M., Huybers, P., 2010b. A Bayesian algorithm for reconstructing climate anomalies in space and time. Part 2: Comparison with the regularized expectation-maximization algorithm Journal of Climate 23.
Tingley, M., Li, B., 2011. Comments on "Reconstructing the NH mean temperature: can underestimation of trends and variability be avoided?" by Bo Chirstiansen. Journal of Climate (Under review; currently available online at www.martintingley.com/wp-content/uploads/2011/08/Comment_on_Christiansen.pdf).
Tolwinski-Ward, S., Evans, M., Hughes, M., Anchukaitis, K., 2011. An efficient forward model of the climate controls on interannual variation in tree-ring width. Climate Dynamics 36, 2419–2439.
Vecchia, A.V., 1988. Estimation and model identification for continuous spatial processes. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 50, 297–312.
Visser, H., Molenaar, J., 1988. Kalman filter analysis in dendroclimatology. Biometrics 44, 929–940.
Wahl, E., Schoelzel, C., Tigrek, S., 2010. On the use and value of Bayesian hierarchical modeling for paleoclimate reconstruction: a discussion of the value of multi-proxy reconstruction of past climate by Bo Li, Dougles W. Nychka, and Caspar M. Ammann. Journal of the American Statistical Association 105, 900–905.
West, M., 1997. Time series decomposition. Biometrika 84, 489.
Whittle, P., 1953. Estimation and information in stationary time series. Arkiv för Matematik 2, 423–434.
Wikle, C., Anderson, C., 2003. Climatological analysis of tornado report counts using a hierarchical Bayesian spatiotemporal model: Application of recent advances in space-time statistics to atmospheric data. Journal of Geophysical Research 108, 9005.
Wikle, C.K., Milliff, R.F., Nychka, D., Berliner, L.M., 2001. Spatiotemporal hierarchical Bayesian modeling tropical ocean surface winds. Journal of the American Statistical Association 96, 382–397.
Yule, G., 1926. Why do we sometimes get nonsense-correlations between Time-Series? – a study in sampling and the nature of time-series. Journal of the Royal Statistical Society 89, 1–63.
Zhang, Z., Mann, M., Cook, E., 2004. Alternative methods of proxy-based climate field reconstruction: application to summer drought over the conterminous united states back to ad 1700 from tree-ring data. The Holocene 14, 502.