

A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part I: Development and Applications to Paleoclimate Reconstruction Problems

MARTIN P. TINGLEY

*Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, North Carolina,
and Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts*

PETER HUYBERS

Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts

(Manuscript received 6 January 2009, in final form 30 October 2009)

ABSTRACT

Reconstructing the spatial pattern of a climate field through time from a dataset of overlapping instrumental and climate proxy time series is a nontrivial statistical problem. The need to transform the proxy observations into estimates of the climate field, and the fact that the observed time series are not uniformly distributed in space, further complicate the analysis. Current leading approaches to this problem are based on estimating the full covariance matrix between the proxy time series and instrumental time series over a “calibration” interval and then using this covariance matrix in the context of a linear regression to predict the missing instrumental values from the proxy observations for years prior to instrumental coverage.

A fundamentally different approach to this problem is formulated by specifying parametric forms for the spatial covariance and temporal evolution of the climate field, as well as “observation equations” describing the relationship between the data types and the corresponding true values of the climate field. A hierarchical Bayesian model is used to assimilate both proxy and instrumental datasets and to estimate the probability distribution of all model parameters and the climate field through time on a regular spatial grid. The output from this approach includes an estimate of the full covariance structure of the climate field and model parameters as well as diagnostics that estimate the utility of the different proxy time series.

This methodology is demonstrated using an instrumental surface temperature dataset after corrupting a number of the time series to mimic proxy observations. The results are compared to those achieved using the regularized expectation–maximization algorithm, and in these experiments the Bayesian algorithm produces reconstructions with greater skill. The assumptions underlying these two methodologies and the results of applying each to simple surrogate datasets are explored in greater detail in Part II.

1. Introduction

To put current and projected future changes of the climate system into context, it is imperative to understand the natural variability and past evolution of the climate system. Particular attention has been given in this regard to the time evolution of the surface temperature field over the last several thousand years, as this variable is of societal importance and features a relatively complete instrumental record extending back to about 1850. Given that a longer record is desirable for both

investigating the dynamics of the system and testing the output of climate models, it becomes necessary to call upon paleoclimate observations, which are noisy and sparsely distributed in space, to extend reconstructions back in time. Information about surface temperatures over the last few millennia can be derived from historical documents, and from elements of the natural world sensitive to local temperature variations, such as tree rings, ice cores, and lake floor sediment cores. For a general review of the uses of these various proxies, see NRC (2006) and Jones et al. (2009).

A common goal when analyzing paleoclimate data is to estimate, with uncertainties, the values of a field on a regular spatial grid (the target locations), at regularly spaced time intervals. For example, much attention has been paid to the problem of reconstructing annual mean

Corresponding author address: Martin P. Tingley, Department of Earth and Planetary Sciences, Harvard University, 20 Oxford St., Cambridge, MA 02138.
E-mail: tingley@fas.harvard.edu

surface temperatures on a regular grid and estimating the regional average of some or all of these grid points (see NRC 2006 and references therein). In general there are two distinct types of data used in paleoclimate reconstructions of a climate field:

- Instrumental time series exist at a number of spatial locations, but perhaps not at all target locations. These time series are assumed to be in the correct units, are often of different lengths, and might feature intermittently missing values.
- Proxy time series exist at a number of spatial locations, some but not all of which might correspond to target locations. These time series are generally longer than the instrumental time series and are not in the same units as the climate field but are assumed to contain information about the climate field; they might, like the instrumental time series, feature intermittently missing values. There may be several distinct types of proxy records, such as various measurements on tree rings, ice cores, lake sediments, corals, etc.

The proxy datasets overlap with the instrumental dataset during a calibration interval. The goal of the analysis is to assimilate all available information to estimate, with uncertainties, either the time series of spatial means, or the field values through time at the target grid locations.

The simpler problem of reconstructing the spatial mean can be approached by combining all instrumental records in a region, combining all proxy records in a region, and then linearly transforming the proxy composite into the units of the instrumental composite. A number of different methods have been used to form the proxy composite and to estimate the coefficients used in the transformation; see Jones et al. (2009) for a discussion of these “composite plus scale” variants.

Estimation of the spatial mean is complicated by the fact that neither the proxy nor the instrumental time series are expected to be uniformly distributed in space, while the field under analysis will typically display spatial covariance. The sample average across the proxy or instrumental observations available for a given year is therefore generally not the best estimate of the spatial mean for that year. The estimate of the mean should consider the spatial distribution of the available data, and this is often done in practice by weighting the proxy observations by the areas they are conjectured to represent (e.g., Mann and Jones 2003; Mann et al. 2005) or by condensing large numbers of closely clustered proxy time series, such as networks of tree ring measurements, by retaining only a few dominant modes from a principal component analysis (PCA) (e.g., Mann and Jones 2003). These procedures involve a number of steps that complicate the propagation of uncertainties. Likewise, the

standard estimate of the uncertainty—the sample standard deviation scaled by the square root of the sample size (e.g., Zar 1999)—is generally not an accurate reflection of the uncertainty in the estimate of the mean. The uncertainty in the estimate of the spatial mean is a function of the spatial covariance of the field, which determines the extent to which observations at a heterogeneous set of locations can be used to predict the field at other locations.

We are interested not only in the time evolution of the spatial mean of climate fields, like temperature, but also the spatial patterns of variability about the mean value. Inferences on the field as a whole, rather than simply the mean value, provide more complete characterizations of the climate variable, which can be compared to the output of climate models and may be useful for studying the dynamics of the climate system (e.g., Riedwyl et al. 2009). In addition, an estimate of the spatially complete field, as well as the associated uncertainties, can be used to produce estimates of the spatial mean and associated uncertainty that take into account the spatial distribution of the observations and the spatial covariance of the field.

A number of methods have been developed and used to reconstruct climate fields from overlapping proxy and instrumental datasets (e.g., Jones et al. 2009). These methods are generally based on multivariate regressions, using the overlap between the instrumental and proxy time series to establish the relationship between the two types of data, and may use the leading modes resulting from a PCA rather than the original time series. The estimated coefficients are then used to predict the values of the instrumental time series back through time using the available proxy time series (Fig. 1). Some approaches consider the linear relationships between all possible pairs of proxy and instrumental time series (e.g., Mann et al. 1998; Schneider 2001; Luterbacher et al. 2004), whereas others use only those proxies within a certain radius to predict the field at each grid point (e.g., Cook et al. 1999). While the former exploits covariances between time series of the field at widely separated locations, the latter, being based on localized regressions, does not. The extent to which each set of assumptions is correct will likely depend on the particular field being analyzed.

At the heart of these multivariate regression approaches is the estimation of the mean, through time, of each proxy and instrumental time series, and the joint covariance matrix of the instrumental and proxy datasets—a submatrix of which must be inverted to calculate the regression coefficients. Some form of regularization is often required to ensure the existence of the matrix inverse, and a solution to this problem is offered by the regularized expectation–maximization (RegEM) algorithm of Schneider (2001), which has been used in a number of

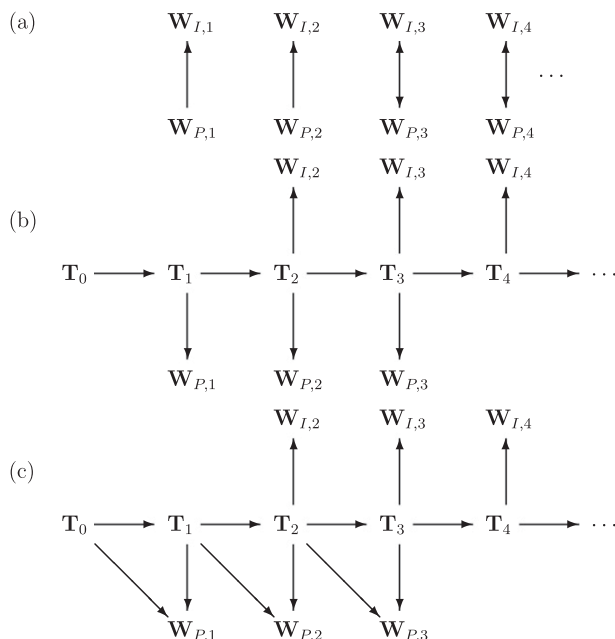


FIG. 1. Schematics of various approaches to reconstructing climate fields. (a) The RegEM algorithm models the relationship between the proxy (\mathbf{W}_P) and instrumental (\mathbf{W}_I) observations and uses this relationship to predict the instrumental values when only proxy observations are available. (b) The hidden Markov model used in BARCAST. Arrows denote the directions of conditional dependencies, in the sense that the observations $\mathbf{W}_{I,t}$ and $\mathbf{W}_{P,t}$ are determined only by the true field vector \mathbf{T}_t , which is in turn determined by the previous value of the field \mathbf{T}_{t-1} . The middle row of arrows, linking the \mathbf{T}_k , corresponds to the process level, while the top and bottom rows, linking the \mathbf{T} to the $\mathbf{W}_{I,P}$, correspond to the data level. (c) A possible generalization of the basic hidden Markov model appropriate for a proxy that integrates two years of the local value of the field.

climate field reconstruction studies (Rutherford et al. 2003; Zhang et al. 2004; Rutherford et al. 2005; Mann et al. 2007, 2008; Steig et al. 2009). There are both benefits and limitations to this methodology, which we will partly address here and in more detail in Tingley and Huybers (2010, hereafter Part II).

An alternative analysis strategy can be formulated by specifying parametric forms for the spatial covariance and temporal evolution of the field and the relationships between the data types and the field. Here we present a hierarchical Bayesian model, referred to as BARCAST for “A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time,” that is used to infer the joint distribution of the scalar parameters that define the model and the field values through time at the target locations. (A package of Matlab code that implements the algorithm is available at <ftp://ftp.ncdc.noaa.gov/pub/data/paleo/softlib/barcast/>). Multiple draws from the posterior result in spatially and temporally complete ensembles of

the climate field evolution compatible with the data and the model assumptions—an information-rich end product. Probability distributions for various statistics can be estimated from this ensemble, from simple measures like the time series of spatial means to more exotic quantities like the maximum decadal average over a specific region. The model also outputs the uncertainty in all scalar parameters, including the coefficients that transform the proxy values into the units of the field. Posterior analysis can quantify both the relative contributions of the different proxies to the field reconstruction and the extent to which the model can constrain the various parameters, while residual analysis can be used to check the validity of the model assumptions and identify particular time series that are not in agreement with the others.

Section 2 describes the technical details of BARCAST, section 3 presents an example demonstrating the functionality of BARCAST and compares results to those from the RegEM algorithm, section 4 discusses limitations of the Bayesian approach and a number of possible extensions to the basic model developed in section 2, and section 5 offers conclusions. Part II provides a detailed comparison of the assumptions and performance of BARCAST to the more established RegEM method. For ease of description we will assume the field of interest is that of annual mean surface temperatures, though the method we have developed is general and in principle applicable to the reconstruction of any climate field.

2. The formulation of BARCAST

a. Basic approach

Our approach to climate field reconstruction is based on a hierarchical Bayesian model consisting of three levels: the *process level* describes the evolution of the true surface temperatures as a multivariate autoregressive process with spatially correlated innovations; the *data level* specifies the relationships between the measurements (both proxy and instrumental) and the true field values; and the *prior level* specifies diffuse and, where possible, conjugate (e.g., Gelman et al. 2003) prior distributions for all unknown parameters to provide closure to the scheme (Fig. 1). In the language of statistics, this formulation is referred to as a discrete time, continuous state hidden Markov model (e.g., Wikle and Berliner 2006). Notation is summarized in Tables 1 and 2.

b. The model equations

1) PROCESS LEVEL

The evolution of the true temperature field sampled at a finite number of spatial locations \mathbf{T}_t is assumed to follow a multivariate first-order autoregressive process:

TABLE 1. Forms of the priors and conditional posteriors, along with brief descriptions, for the unknowns inferred by BARCAST. MV stands for multivariate, and nonstandard indicates that the conditional posterior does not follow a well-known distribution.

	Prior form	Conditional posterior	Description
\mathbf{T}_0	MV normal	MV normal	Field values for the time step prior to the first observations.
$\mathbf{T}_{k=1 \dots \kappa}$	—	MV normal	Field values at each time step for which there are observations.
α	Uniform	Truncated normal	AR(1) coefficient in the field evolution equation.
μ	Normal	Normal	Mean of \mathbf{T} .
σ^2	Inverse-gamma	Inverse-gamma	Partial sill of the spatial covariance matrix of the innovations that drive the AR(1) process.
ϕ	Log-normal	Nonstandard	Inverse range of this spatial covariance matrix.
τ_I^2	Inverse-gamma	Inverse-gamma	Error variance of instrumental observations.
τ_P^2	Inverse-gamma	Inverse-gamma	Error variance of proxy observations.
β_1	Normal	Normal	Scaling factor in the proxy observation equation.
β_0	Normal	Normal	Additive constant in the proxy observation equation.

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

where μ is the mean of the process, α is the AR(1) coefficient, $\mathbf{1}$ is a vector of ones, and the subscript t indexes the year. A more general model could be formulated by replacing the scalar α with a matrix, which would allow the elements of \mathbf{T}_t to have different autoregressive parameters, and to display cross dependencies. Here and below we make the simplest assumptions that we consider reasonable, and we discuss possible extensions in section 4. The innovations ϵ_t are assumed to be independent and identically distributed (iid) normal draws, $\epsilon_t \sim N(0, \Sigma)$, with spatial covariance structure given by

$$\Sigma_{ij} = \sigma^2 \exp(-\phi |\mathbf{x}_i - \mathbf{x}_j|), \quad (2)$$

where $|\mathbf{x}_i - \mathbf{x}_j|$ is the distance between the i th and j th elements of the field vector \mathbf{T} . Note that this formulation of the spatial covariance intentionally excludes a nugget effect (e.g., Banerjee et al. 2004); the reason for this is addressed below. The implications and limitations of assuming an exponentially decaying spatial covariance structure will be addressed in detail below, but for now we note that the Climate Research Unit (CRU) annual mean instrumental temperature data (Brohan et al. 2006)

does seem to exhibit plausible exponential decay of correlation with separation, at least for separations smaller than about 4000 km (Fig. 2). The saturation of correlation at positive values at length scales longer than 4000 km is likely the result of trends in the CRU dataset—indeed, if data are detrended first the correlation decays to values indistinguishable from zero.

2) DATA LEVEL

It is useful to decompose the vector \mathbf{T} , at each year, into three subvectors:

$$\mathbf{T} = \begin{pmatrix} \mathbf{T}_I \\ \mathbf{T}_P \\ \mathbf{T}_R \end{pmatrix}, \quad (3)$$

where \mathbf{T}_I and \mathbf{T}_P are the true temperatures at locations for which there are instrumental or proxy observations, respectively. If there is both a proxy and instrumental observation at the same location, then the true field value for that location appears in both \mathbf{T}_I and \mathbf{T}_P . The values for \mathbf{T}_R are the true temperatures at the target locations where there are no observations. For the examples presented below, we select these target locations to be the remaining nodes of a uniform grid.

TABLE 2. Descriptions of other variables appearing frequently in the model equations.

	Description
\mathbf{W}_t	Observations of a subset of \mathbf{T}_t , subdivided into $\mathbf{W}_{I,t}$ and $\mathbf{W}_{P,t}$, where the additional subscript I or P indicates instrumental and proxy observations, respectively.
κ	Number of years for which there are (any) observations; time runs from 0, the year prior to the first observations, to κ .
$N_{I,t}$ and $N_{P,t}$	The number of instrumental and proxy observations at year t ; without the time index, these refer to the total number of locations for which there are either instrumental or proxy observations.
$M_I \equiv \sum_{k=1}^{\kappa} N_{I,k}$	Total number of instrumental observations; M_P is the equivalent measure for the proxy data.
N_A	Total number of locations at which the field is estimated.
Θ	A vector consisting of the eight scalars that define the model.

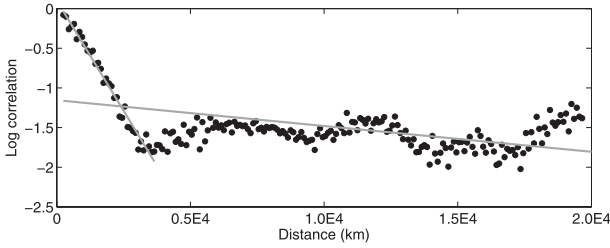


FIG. 2. Log correlation as a function of separation for annual mean temperature anomalies, using the global CRU dataset. Results are shown for the 305 locations over land with at least 8 monthly values a year for at least 100 years. Distances were grouped into 100-km bins, and the log of the median correlation in each bin is plotted. Linear fits to log correlation as a function of separation are to the entire dataset and to only those distance bins less than 3700 km. The fit to the entire dataset corresponds to an e -folding length scale of 30 000 km, while that to distances less than 3700 km corresponds to a length scale of 1800 km.

The instrumental observations at each year, $\mathbf{W}_{I,t}$, are assumed to be noisy versions of the true temperatures at the corresponding locations:

$$\mathbf{W}_{I,t} = \mathbf{T}_{I,t} + \mathbf{e}_{I,t}. \quad (4)$$

The noise terms are assumed to be iid multivariate normal draws, $\mathbf{e}_{I,t} \sim N(0, \tau_I^2 \mathbf{I})$, where \mathbf{I} is the identity matrix. Note that the instrumental temperature observations are subject to systematic errors (Brohan et al. 2006) that are not dealt with here.

The proxy observations are assumed to have an unknown, statistically linear relationship to the true temperatures at the corresponding locations. This motivates the regression equation:

$$\mathbf{W}_{P,t} = \beta_1 \mathbf{T}_{P,t} + \beta_o \mathbf{1} + \mathbf{e}_{P,t}. \quad (5)$$

The noise terms are once more assumed to be iid normal draws, $\mathbf{e}_{P,t} \sim N(0, \tau_P^2 \mathbf{I})$. A standard regression model that seeks to predict the field values would generally rewrite this equation to isolate $\mathbf{T}_{P,t}$ on the left-hand side. Such a formulation is equivalent to Eq. (5) in the sense that the parameters of one model can be written in terms of the parameters of the other. From the Bayesian perspective, however, it is simpler to describe the form of the observations conditional on the unknown value of $\mathbf{T}_{P,t}$, as this quantity appears in the expression for the posterior of $\mathbf{T}_{P,t}$. In most cases the assumption of linearity will be a gross simplification of the relationship between the proxies and the climate field of interest—tree ring growth, for example, has a complicated and highly nonlinear relationship with local climate variables (e.g., Evans et al. 2006). This possibly poor assumption is common to all regression-based reconstruction

approaches, and in some situations it may be useful to transform the proxy data prior to the analysis to bring the data into better agreement with the assumptions.

No nugget effect (e.g., Banerjee et al. 2004) is included in the spatial covariance matrix Σ , as it would be redundant given the observational error variances τ_I^2 and τ_P^2 . A nugget effect is usually included in a parameterized spatial covariance matrix to inflate the covariance at separation zero to account for variation, including observational error, on length scales that cannot be resolved by the data (Banerjee et al. 2004), but the two τ^2 parameters already account for this. Inference on a nugget in the spatial covariance matrix would therefore be ill conditioned, as multiple parameters would model the same phenomenon.

We can represent the observation equations at each year, taking into account the missing data structure, as

$$\mathbf{W}_t = \begin{pmatrix} \mathbf{H}_{I,t} \\ \beta_1 \cdot \mathbf{H}_{P,t} \end{pmatrix} \mathbf{T}_t + \begin{pmatrix} \mathbf{e}_{I,t} \\ \mathbf{e}_{P,t} + \beta_o \mathbf{1} \end{pmatrix}, \quad (6)$$

where $\mathbf{H}_{I,t}$ and $\mathbf{H}_{P,t}$ are selection matrices of zeros and ones that pick out, at each year, the elements of \mathbf{T}_t corresponding to locations for which there are observations.

The observations for any given year, conditional on the true field vector and parameters, are multivariate normal:

$$\mathbf{W}_t | \mathbf{T}_t, \boldsymbol{\Theta} \sim N(\mathbf{H}_t \mathbf{T}_t + \mathbf{B}_t, \boldsymbol{\Xi}_t), \quad (7)$$

where $\boldsymbol{\Theta}$ is a vector composed of the eight scalar parameters (Table 1) and the following notation has been used to simplify the equations:

$$\mathbf{H}_t = \begin{pmatrix} \mathbf{H}_{I,t} \\ \beta_1 \cdot \mathbf{H}_{P,t} \end{pmatrix}, \quad \mathbf{B}_t = \begin{pmatrix} \mathbf{H}_{I,t} \\ \mathbf{H}_{P,t} \end{pmatrix} \begin{pmatrix} \mathbf{0}_{N_I} \\ \beta_o \mathbf{1}_{N_P} \end{pmatrix},$$

$$\boldsymbol{\Xi}_t = \begin{pmatrix} \tau_I^2 \cdot \mathbf{I}_{N_I} & \mathbf{0} \\ \mathbf{0} & \tau_P^2 \cdot \mathbf{I}_{N_P} \end{pmatrix}. \quad (8)$$

We suppress \mathbf{T}_R , the third element of \mathbf{T} , in expressions for $\mathbf{W}_t | \mathbf{T}_t, \boldsymbol{\Theta}$ as the field is never observed at these locations.

3) PRIOR LEVEL AND DRAWING FROM THE POSTERIOR

To close the analysis scheme, priors must be specified for the eight scalar parameters and the climate field for the first year in the analysis. Our approach is to use weakly informative but proper prior distributions, and show that the information provided by the data overwhelms the prior. Where possible, we have used conditionally conjugate priors to facilitate computations (Table 1). Methods

of choosing the hyperparameters for the prior distributions are discussed in appendix A.

The probability of the data, given the true field values and all parameters, can be factored to give

$$P(\mathbf{W}_1, \dots, \mathbf{W}_\kappa | \mathbf{T}_1, \dots, \mathbf{T}_\kappa, \Theta) = \prod_{k=1}^{\kappa} P(\mathbf{W}_k | \mathbf{T}_k, \Theta). \quad (9)$$

Applying Bayes' rule and rearranging results in

$$P(\mathbf{T}_0, \dots, \mathbf{T}_\kappa, \Theta | \mathbf{W}_1, \dots, \mathbf{W}_\kappa) \propto P(\mathbf{T}_0) \cdot P(\Theta) \cdot \prod_{k=1}^{\kappa} P(\mathbf{W}_k | \mathbf{T}_k, \tau_I^2, \tau_P^2, \mu, \beta_0, \beta_1) P(\mathbf{T}_k | \mathbf{T}_{k-1}, \sigma^2, \phi, \alpha). \quad (10)$$

Samples are drawn from this posterior using a Gibbs sampler with a single Metropolis step (e.g., Gelman et al. 2003) used to update ϕ (the inverse spatial range parameter). For all parameters other than ϕ , the priors are conditionally conjugate so samples can be drawn directly from the full conditional posteriors. The forms and parameters of these full conditional posteriors, as well as the details of the Metropolis step used to update ϕ , are described in appendix B.

To speed up the convergence of the Gibbs sampler, the three variance parameters (σ^2 , τ_I^2 , τ_P^2) can be truncated to exclude very large values. As the value of the true underlying space–time field used to initialize the sampler can disagree considerably with the information provided by the data, the draws of the variance parameters can initially inflate to extremely high values. In our experience, samples of the variance parameters eventually converge to more reasonable values, as the samples of the field come into better agreement with the data, but this process can take many thousands of iterations. If an upper bound is placed on the values of the variance parameters, we find that convergence is much faster, as the algorithm cannot initially account for all discrepancy between the estimate of the true field and the data by setting the variance parameters to high values. This is equivalent to setting the priors to truncated inverse-gamma distributions (see appendix A). Upper bounds are set to be far in excess of any sample after the algorithm has converged, so this approach does not artificially reduce the posterior uncertainty in these parameters. To further increase the speed of convergence, we have found it useful to initially run the Gibbs sampler in a reduced mode that only updates values of the underlying field, and not the scalar parameters. This ensures that the estimates of the field are in rough agreement with the data and initial parameter values, so when the full version of the algorithm is applied, the variance parameters more readily converge.

The speed of the algorithm can also be increased by exploiting the fact that the matrix inverse needed to sample from the conditional posterior of \mathbf{T}_k is a function only of the pattern of missing data for that year. As a result, the number of matrices that must be inverted to sample from each of the \mathbf{T}_k in turn is given by the number of unique patterns of missing data, which can be much smaller than the number of years in the reconstruction.

While the computational demands of the Bayesian approach are larger than for other comparable methods, they are not prohibitive. Using an ordinary desktop computer, each iteration of BARCAST for the experiments discussed below takes about three seconds, so that producing the 2200 draws used in each analysis requires about 2 hours of computer time. In contrast, RegEM with one ridge regression per missing observation takes about 12 minutes to converge for each example.

c. Standardization of the proxy time series

If the proxy time series are standardized prior to the analysis to have means of zero and standard deviations of one, then β_1 and β_0 can be solved for in terms of the other parameters to give

$$\beta_0 = -\beta_1 \mu \quad \text{and} \quad \beta_1 = \left[\frac{(1 - \tau_P^2)(1 - \alpha^2)}{\sigma^2} \right]^{1/2}. \quad (11)$$

These expressions follow from taking the expectation and variance of both sides of Eq. (5), respectively, and the fact that each element of \mathbf{T} is AR(1) in time, with variance $\sigma^2/(1 - \alpha^2)$ (e.g., Brockwell and Davis 1991).

The probability model could reasonably be simplified in this specific case by setting β_0 and β_1 to these functions of the other parameters. We introduce β_0 and β_1 as distinct parameters for several reasons. First, the inclusion of β_0 and β_1 makes the model applicable in cases where the proxy time series are not standardized. Second, standardization involves estimating population parameters from data. We anticipate that the estimates of the population mean and standard deviation will be imperfect, and including these two parameters is one way of accounting for and propagating the uncertainty in these estimates. Third, the conditional posteriors are simplified by the introduction of these parameters, which facilitates the implementation of the Gibbs sampler used to draw from the posterior. Fourth, we anticipate that the linear relationship between the proxies and field assumed by the model is imperfect. Inclusion of these two parameters, even in cases where the proxy time series have been standardized, gives the model more flexibility to account for imperfections in the assumed relationship. In addition, we do not want the inferred mean field value μ to be corrupted by nonlinearities in the field–proxy

relationship, which could happen if we fixed the β parameters as in Eq. (11). In short, the analysis model includes a mean and temporal standard deviation value for the field, as well as a location and scale value that relates the proxies to the field, and we expect there to be a relationship between these quantities if the proxies have been standardized and the linearity assumption is correct. Note that even if the proxies have been standardized, the parameters β_1 and β_0 are not redundant, in the sense that inference on these parameters is well conditioned. This is in contrast to the situation discussed above with regards to a nugget effect and the τ^2 parameters.

More generally, the manner in which the proxy time series are standardized prior to the analysis can have significant impacts on the results of reconstructions (Mann et al. 2007; Lee et al. 2008; Smerdon and Kaplan 2007; Smerdon et al. 2008). While this issue is not the focus of the current work, the importance of the issue and a possible treatment offered by BARCAST warrants a few words. Broadly speaking, we distinguish two different approaches to standardizing the proxy time series:

- 1) Each proxy time series is standardized by removing the sample mean and scaling by the sample standard deviation of that time series.
- 2) All proxies are standardized by removing a common mean and scaling by a common standard deviation.

Similarly, we distinguish between two different types of analysis that assume a local relationship between the proxy time series and the instrumental or true field time series:

- 1) Each true field or instrumental time series is regressed onto the nearest standardized proxy time series, with a different set of coefficients estimated for each regression.
- 2) The true field or instrumental time series are regressed onto the standardized proxy time series using a common set of coefficients.

If the first standardization approach is combined with the first analysis approach, then the standardization is irrelevant, in the sense that the reconstruction would be the same if the proxies were not standardized. The underlying assumption in this case is that each proxy is linearly related to the local field value, with a different relationship for each proxy time series. The analysis requires the estimation of two regression parameters and an error variance for each proxy time series [cf. Eq. (5)].

If the second standardization approach is combined with the second analysis approach, then standardization is likewise irrelevant. The underlying assumption is that there is a linear relationship between each proxy and the true local field value and that the relationship is the same

for all proxies. This analysis requires the estimation of two regression parameters plus an error variance, which are assumed to be the same for each proxy time series, and is the approach taken when applying the version of BARCAST described above [Eq. (5)].

Standardization influences the reconstruction when the first standardization approach is combined with the second analysis approach. The reconstruction then requires the estimation of two parameters per record (standardization), plus two additional regression parameters and an error variance common to all proxy records. The logic behind this approach is that, if the standardized proxy time series all have the same error variances, then the linear relationship with the local true field should be the same for each. This assumption, however, only holds if all proxy time series are standardized using the population values of the time series means and standard deviations (i.e., calculated from time series of infinite length). As the standardization is conducted using sample estimates of the mean and standard deviation, rather than the population quantities, even if all (standardized) error variances are the same, the linear transformation relating the proxies to the true field will be slightly different for each proxy time series. The error introduced by the estimation of the standardizing coefficients for each time series is not propagated through the analysis, and experiments we have performed with surrogate data show that the resulting credible intervals for field values reconstructed using BARCAST tend to be too narrow. We will return to this issue in section 4, which discusses shortcomings and extensions.

d. Connections with other statistical techniques

Applied to one year of instrumental data, BARCAST reduces to a Bayesian implementation of the standard spatial technique of *kriging* (e.g., Banerjee et al. 2004), which is used to predict a spatial field from observations at a discrete set of locations. The value of BARCAST is in the inclusion of multiple types of data, each having a different relationship with the underlying field, and in the treatment of the time dimension.

If all scalar parameters are specified a priori, then the BARCAST estimates of the mean and variance of the field at each year are equivalent to those from the Kalman smoother (e.g., Kalman 1960; Wikle and Berliner 2006). The advantages of BARCAST include the simultaneous estimation of both the field and the scalar parameters that define the model, and the resulting ensemble of draws of the space time field that are consistent with both the data and the modeling assumptions. BARCAST offers a cohesive framework for estimating all unknowns, produces a richer end product, and results in uncertainty

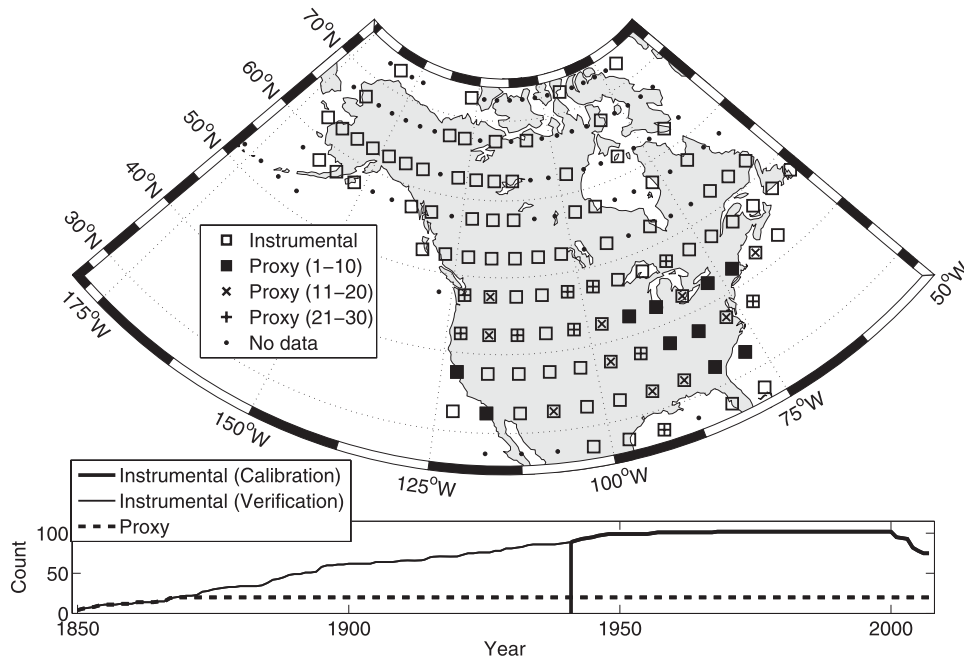


FIG. 3. (top) Locations of the data time series used to reconstruct North American surface temperatures. The squares indicate the locations of the CRU time series, while symbols inside the squares indicate the locations of the longer CRU time series that are used in the construction of pseudoproxies in the nine experiments. Solid black marks the locations of the ten longest CRU time series; the x's mark the 11th through 20th and the plus signs the 21st through 30th longest time series. The small black dots mark the remainder of the grid locations where temperatures are estimated. (bottom) The number of pseudoproxies and instrumental records available at each year for the medium dataset; the CRU values before 1941 are withheld from the analysis and are used to test the reconstructions.

estimates for the field that take into account the uncertainties in the scalar parameters.

3. Example: BARCAST

We demonstrate the functionality of BARCAST by analyzing a number of simple example datasets based on the Climate Research Unit's surface temperature compilation for North America (Brohan et al. 2006). The CRU data are a gridded product with a resolution of 5° longitude by 5° latitude that features many locations without data (Fig. 3). The temperature values at each location are annual mean anomalies, in $^\circ\text{C}$, from the 1961–90 mean. Each example dataset is formed by converting a number of the longest annual mean CRU temperature anomaly time series into pseudoproxies by specifying the values of β_1 , β_0 , and τ_P^2 in Eq. (5). The values of the regression coefficients are the same for each experiment, $\beta_1 = 2$ and $\beta_0 = 1$, while the number of pseudoproxies and the value of τ_P^2 vary between the experiments (Tables 3 and 4). We consider values of τ_P^2 that correspond to proxy signal-to-noise ratios (SNR), in terms of standard deviations, of 1, $\frac{1}{2}$, and $\frac{1}{3}$ and use 30, 20, or 10 pseudoproxy time series, for a total of nine experiments.

For ease of reference, the experiment with the largest SNR and greatest number of pseudoproxy time series will be referred to as the “easy” experiment, while those with the middle and smallest value of each quantity will be referred to as the “medium” and “hard” experiments, respectively. The SNR values are in line with those used in other assessments of reconstruction techniques (e.g., von Storch et al. 2009; Lee et al. 2008; Mann et al. 2007).

In each experiment, the CRU values after 1940 (the calibration period) form the instrumental dataset, while the CRU values available in the 1850–1940 interval are withheld from the analysis and used to test the reconstructions (Fig. 3). The goal of each analysis is to estimate the temperature field at all nodes of the grid, even those for which no observations are available, at all years in the 1850–2007 interval. As BARCAST includes a spatial model, it is possible to estimate the temperatures at an arbitrarily fine spatial resolution. There seems little point, however, in making estimates at a scale finer than that of the original data.

We also analyze each of the example datasets using the RegEM algorithm, with one ridge regression per missing value providing the regularization, and the variance inflation factor set to one (Schneider 2001); Part II will

TABLE 3. Ninety percent credible intervals for the scalar parameters inferred from BARCAST for each of the nine experiments. In each experiment, β_1 is set to 2 and β_0 is set to 1, while the values of τ_P^2 used to generate the pseudoproxies are listed in the second column. For these three parameters, credible intervals that do not contain the specified values are shown in *italics*. The experiments are arranged in the same order as in Table 4, which lists the number of pseudoproxy time series and the SNR for each experiment.

Name	τ_P^2	α	μ	σ^2	$\phi \times 10\,000$	$\tau_I^2 \times 100$	τ_P^2	β_1	β_0
Easy	2.75	(0.42, 0.47)	(−0.17, 0.10)	(0.67, 0.78)	(3.03, 3.60)	(1.15, 1.56)	(2.66, 2.89)	(2.05, 2.21)	(0.92, 1.05)
	10.0	(0.39, 0.43)	(−0.13, 0.13)	(0.67, 0.80)	(2.97, 3.53)	(1.13, 1.52)	(10.7, 11.5)	(1.85, 2.17)	(0.77, 1.02)
	21.6	(0.39, 0.43)	(−0.18, 0.12)	(0.69, 0.88)	(2.62, 3.46)	(1.15, 1.52)	(23.5, 25.3)	(1.72, 2.21)	(0.81, 1.15)
Medium	2.75	(0.42, 0.46)	(−0.12, 0.14)	(0.64, 0.78)	(3.03, 3.79)	(1.11, 1.53)	(2.38, 2.65)	(2.11, 2.30)	(0.83, 0.98)
	10.0	(0.39, 0.43)	(−0.14, 0.15)	(0.71, 0.86)	(2.71, 3.34)	(1.16, 1.51)	(10.0, 11.0)	(1.82, 2.21)	(0.79, 1.07)
	21.6	(0.39, 0.43)	(−0.15, 0.17)	(0.76, 0.92)	(2.52, 3.11)	(1.18, 1.56)	(22.2, 24.3)	(1.41, 1.94)	(0.83, 1.24)
	2.75	(0.40, 0.45)	(−0.20, 0.06)	(0.67, 0.85)	(2.73, 3.50)	(1.18, 1.57)	(2.39, 2.79)	(1.90, 2.17)	(0.99, 1.19)
Hard	10.0	(0.38, 0.43)	(−0.08, 0.21)	(0.69, 0.91)	(2.53, 3.41)	(1.20, 1.56)	(9.07, 10.3)	(1.67, 2.16)	(0.55, 0.91)
	21.6	(0.39, 0.43)	(−0.12, 0.20)	(0.70, 0.94)	(2.45, 3.32)	(1.17, 1.60)	(19.9, 22.5)	(1.51, 2.16)	(0.42, 0.94)

provide a more in-depth comparison of BARCAST and RegEM. We do not standardize the pseudoproxy or instrumental datasets prior to the analysis with either method, and the effects of various standardization procedures on BARCAST and RegEM will not be explored in this work.

The focus of this section will be the output produced by applying BARCAST to the medium dataset (Tables 3 and 4), while results from the other experiments are included to show that the conclusions are robust to the particulars of the experiments. The basic result of applying BARCAST to each dataset is an ensemble of posterior draws of the space–time field and scalar parameters, consistent with the data and the model assumptions; the results below are based on 2000 such draws. These results do not represent a complete vetting of BARCAST, but rather demonstrate the utility of the new algorithm in a variety of reasonable scenarios. Future work with climate model data and real proxy data will be required to fully assess the strengths and weaknesses of this new approach.

a. Fields and time series

We show the BARCAST and RegEM field estimates for two representative years of the medium dataset (Figs. 4 and 5, respectively): 1888, for which only pseudoproxy observations are available, and 1988, for which both types of observation are available. The BARCAST field estimate at each location is the median of the 2000 draws from the posterior, while the corresponding uncertainty is the distance between the 5th and 95th percentiles of the posterior draws. For both years, BARCAST identifies the temperature estimates in regions far from any available data as the most uncertain. For 1888, the BARCAST field estimate decays smoothly toward the mean value (and the uncertainty increases) as one moves away from the two concentrated areas of observations—this is a result of the assumed covariance structure. There is more structure and less uncertainty in the BARCAST field estimate for 1988 than for 1888, as there are more observations to constrain the estimates.

TABLE 4. Descriptions of the three experiments used to demonstrate BARCAST and compare it to RegEM, along with statistics used to test the reconstructions. “Number” refers to the number of pseudoproxy time series used in each experiment, while SNR gives the signal-to-noise ratio of the pseudoproxies in terms of standard deviations. “Average r^2 ” and “average CE” refer to the means of the r^2 and CE values, respectively, calculated for the withheld CRU time series in each experiment (cf. Fig. 11), while “coverage rate” refers to the fraction of the withheld CRU values covered by the 90% credible (BARCAST) or confidence (RegEM) intervals. The rightmost column shows the correlation between the spatial mean time series estimated from BARCAST and RegEM during the 1850–1940 testing interval.

Name	Number	SNR	τ_P^2	Average r^2		Average CE		Coverage Rate		Correlation (1850–1940)
				BARCAST	RegEM	BARCAST	RegEM	BARCAST	RegEM	
Easy	30	1	2.75	0.47	0.42	0.26	0.09	0.91	0.76	0.97
	30	1/2	10.0	0.29	0.22	0.00	−0.21	0.89	0.78	0.90
	30	1/3	21.6	0.25	0.17	−0.08	−0.27	0.90	0.76	0.65
Medium	20	1	2.75	0.38	0.34	0.10	−0.02	0.89	0.75	0.93
	20	1/2	10.0	0.27	0.21	−0.07	−0.25	0.89	0.78	0.84
	20	1/3	21.6	0.17	0.14	−0.28	−0.44	0.90	0.78	0.61
	10	1	2.75	0.29	0.27	−0.01	−0.04	0.90	0.78	0.87
	10	1/2	10.0	0.20	0.14	−0.27	−0.34	0.89	0.76	0.76
Hard	10	1/3	21.6	0.14	0.07	−0.37	−0.44	0.90	0.78	0.64

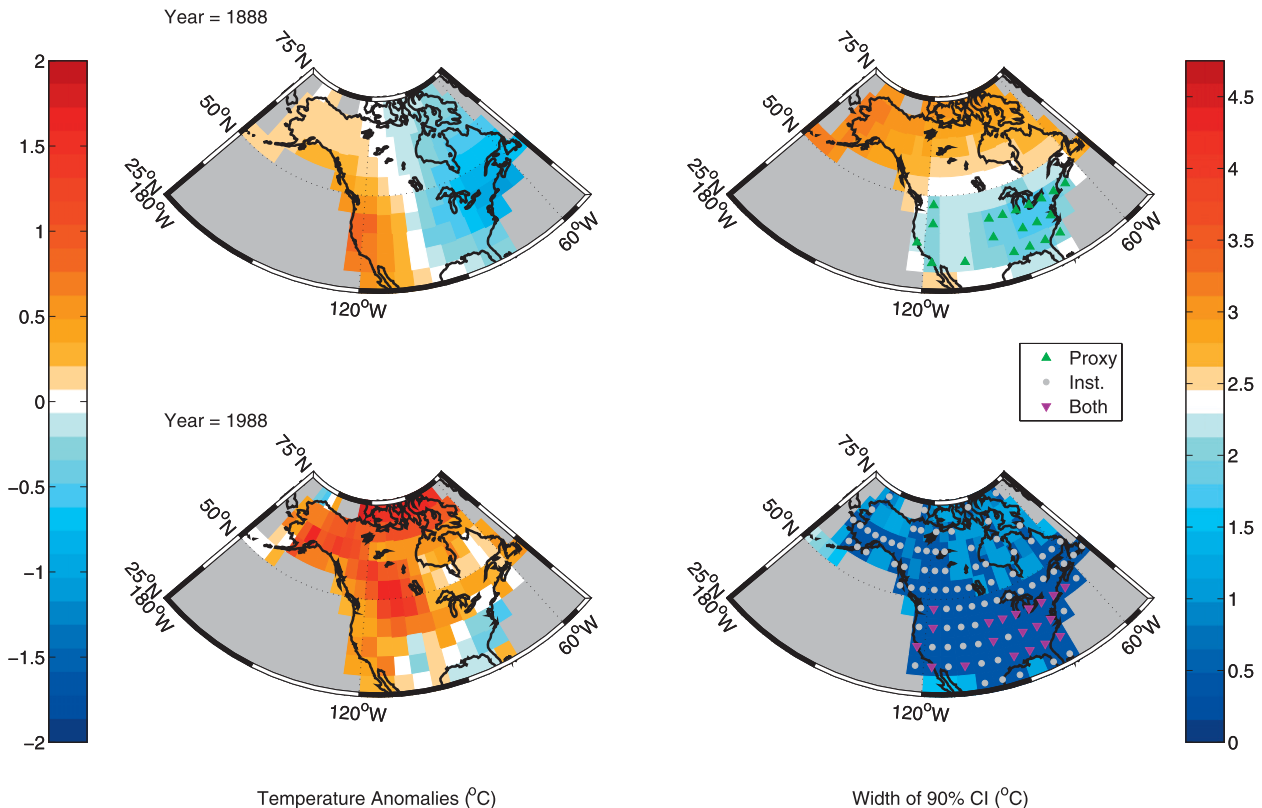


FIG. 4. (left) BARCAST estimates of the temperature anomaly field in North America for two different years, and (right) the corresponding uncertainty, for the medium experiment. The estimates are the medians of the posterior draws, and the uncertainties the widths of the 90% credible intervals. In the uncertainty panels, gray dots, green triangles, and purple triangles indicate locations where there are, respectively, instrumental, proxy, or both types of observation available for that year.

RegEM as currently implemented is designed to infer missing values in an incomplete dataset (Schneider 2001), and so does not impute the field at locations where there are no observations (Fig. 5; see Part II for a discussion of this issue). The widths of the 90% uncertainty intervals for the estimates of the field at each location are the standard error estimates provided by RegEM, multiplied by 2.71 (the distance between the 5th and 95th percentiles of the standard normal distribution). For 1888, the RegEM uncertainty tends to grow as the distance from the proxy observations increases, which is consistent with the use of an exponentially decaying spatial covariance structure in BARCAST. There are instrumental observations for 1988 at each location where there are *any* instrumental observations. As RegEM does not consider errors in the instrumental observations (see Part II), the RegEM temperature estimate at each location for 1988 is simply the original CRU value, and there is no uncertainty in these estimates (Fig. 5). The 1988 CRU temperature field is visually similar to that estimated using BARCAST, which is dominated by the instrumental observations, whereas in 1888 the features estimated by

RegEM have smaller amplitudes than those estimated by BARCAST.

The posterior draws from BARCAST provide estimates, with uncertainty, of the temperature time series at each location (Fig. 6). At locations and times where there are instrumental observations, the reconstructed temperatures are in close (but not perfect) agreement with the instrumental values. At these times and locations, the uncertainty is very small—according to the algorithm, the instrumental observations are excellent estimates of the true field. Going back in time, the uncertainty at these locations increases rapidly as the instrumental observations end in 1941. Prior to 1941, the uncertainty is largest at locations and times for which no proxy observations are available, and there are noticeable deviations between the estimates from BARCAST and the withheld CRU values. The uncertainty in the estimation of the field values, calculated from multiple draws from the joint posterior, accounts for the uncertainty in all other parameters of the model, but does not account for errors in the *structure* of the model. Our results are conditional on the assumptions we have made

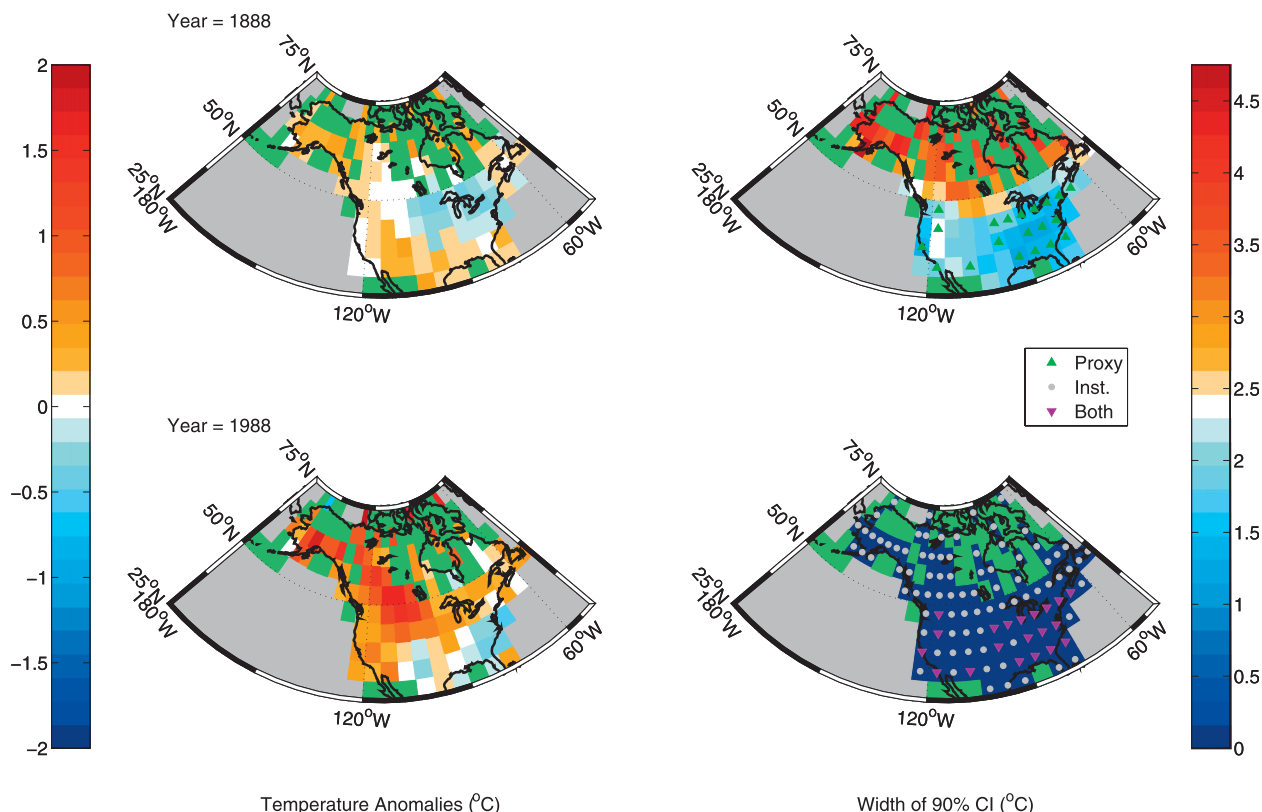


FIG. 5. As in Fig. 4, but using RegEM to estimate the field and uncertainties, with one ridge regression per missing value providing the regularization. In all panels, the green shading indicates locations where RegEM does not predict the field because there are no instrumental time series. The color scales for the estimates of the field values and the associated uncertainties are the same as those in Fig. 4.

about the covariance matrix, AR(1) temporal evolution, and observation equations. The consistency between the calculated uncertainties and the deviations between the reconstructed and withheld time series will be addressed below.

Estimates of the time series of spatial means, both raw and smoothed in time, can be calculated from the posterior draws of the field produced by BARCAST (Fig. 7), where the weighting of each grid box is proportional to the area of the land it contains. To estimate the smoothed time series of spatial means and the associated uncertainty, we smooth the mean time series calculated from each posterior draw from BARCAST, and then take, at each year, the 5th, 50th, and 95th percentiles of the resulting distribution. In general, the posterior distribution of any function of the space–time field can be estimated simply by calculating the quantity for each draw from the posterior distribution of the space–time field. We also show the corresponding results from analyzing the medium dataset with RegEM (Fig. 7). In the case of RegEM, calculating the uncertainty in the smoothed time series is a nontrivial task. We are not

aware that these difficulties have been addressed elsewhere, and we do not attempt to calculate a statistically appropriate confidence interval estimate. As a rough indication of the uncertainty, we show the smoothed uncertainty envelope from the raw time series, which is biased wide (Fig. 7). For years when all instrumental observations are available, there is no reported uncertainty in the RegEM estimate of the spatial mean, as no missing values need to be imputed. During the 1940–2007 calibration interval, the RegEM and BARCAST estimates are in close agreement save for the last few years, which feature an increasing number of missing instrumental observations. Prior to 1940, the shapes of the average time series estimated using each method are similar—the correlation between the two is 0.84—but BARCAST infers a larger amplitude for the variations, which is most apparent in the smoothed time series (Fig. 7). As evidenced by the correlations between the block average time series, the agreement between the two analysis methods is greatest for the easy experiment and generally decreases as the SNR and number of proxy time series decrease (Table 4).

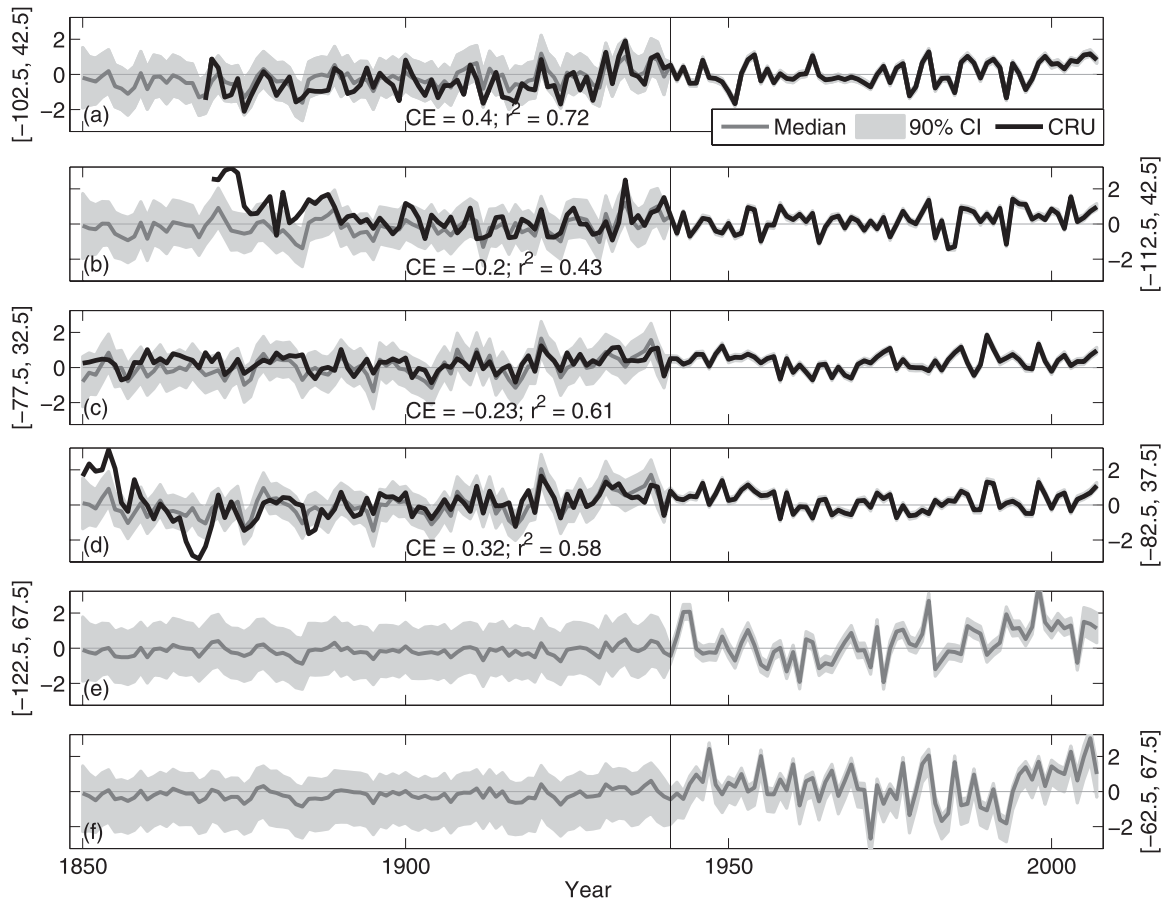


FIG. 6. BARCAST estimates of temperature anomaly (in $^{\circ}\text{C}$) time series at selected locations, for the medium experiment. The light gray fill shows the 90% credible intervals from BARCAST, the medium gray lines the medians from BARCAST, and the black lines the unaltered CRU values, when and where they are available. Locations are given on the y axes as longitude–latitude pairs. (a),(b) Locations where there are CRU observations only; (c),(d) locations where there are both CRU observations and pseudoproxy observations; (e),(f) locations without data. In (a)–(d), the CE and correlation statistics are shown for the testing interval, the time prior to 1941 during which the CRU observations are available but withheld from the analysis.

b. Scalar parameters and convergence for BARCAST

Histograms of the priors and posteriors of the eight scalar parameters inferred by BARCAST in each of the named experiments show that the priors have virtually no influence on the analysis—the posteriors are dominated by the information supplied by the data (Fig. 8). The posterior distributions of the four parameters that define the space–time structure of the field (α , μ , σ^2 , and ϕ) vary somewhat between the experiments (Fig. 8 and Table 3) but display considerable overlap. The three experiments with the lowest SNR value generally estimate higher values for α . With smaller proxy observational errors in the longer pseudoproxy time series, BARCAST infers a greater degree of temporal autocorrelation for the underlying field. The posterior distributions for σ^2 and ϕ shift toward higher and lower values, respectively, as both the SNR and the number of proxy records decrease,

while the posterior distributions of μ are largely unchanged between experiments (Fig. 8 and Table 3).

BARCAST infers a spatial correlation length scale, calculated as $1/\phi$, of about 3300 km, which is larger than the 1700 and 1500 km estimated by Hansen and Lebedeff (1987) and Mann and Park (1993), respectively. The estimates from those studies are in line with the 1800 km derived from the global CRU dataset (Fig. 2). Note that our domain contains no ocean, which would introduce sharp boundaries, and that the length-scale estimate is derived largely from data after 1940. The secular increase in surface temperatures since that time could contribute to the estimates of the length scale being larger than in other studies.

The posterior draws of ϕ and σ^2 are negatively correlated ($r = -0.93$; Fig. 9). Given the spatial covariance form [Eq. (2)] and time series at only two locations, a curve in (ϕ, σ^2) space results in a modeled covariance

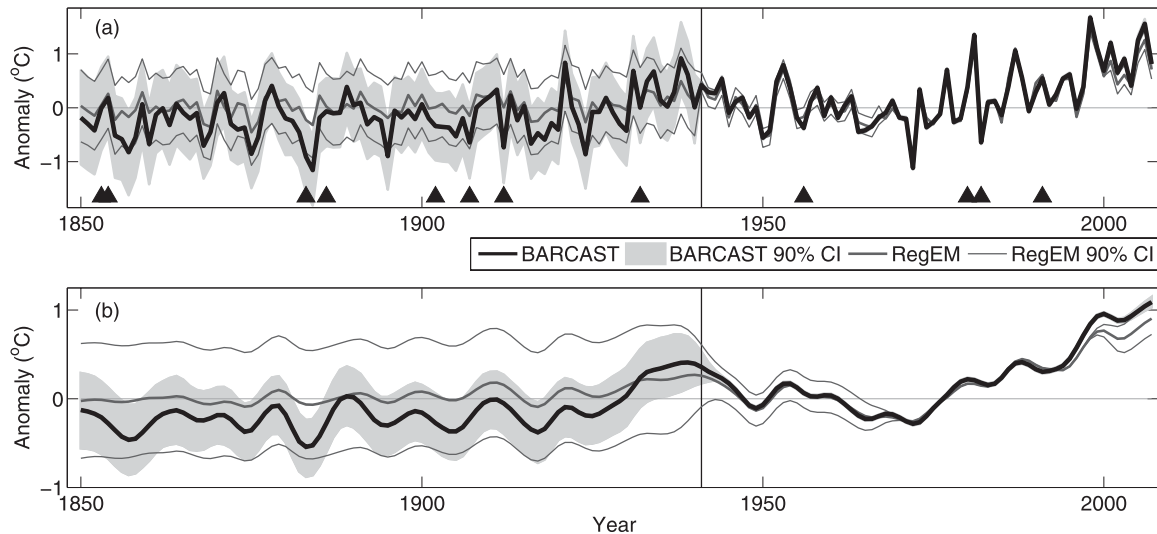


FIG. 7. (a) Reconstructed spatial average temperature anomaly (in $^{\circ}\text{C}$) for North America, with uncertainty, for the medium experiment. The thick black line and light gray fill are, respectively, the median and 90% credible intervals from BARCAST. The medium and thin black lines are, respectively, the mean and 90% uncertainty estimated using RegEM, where one ridge regression per missing value provides the regularization. The black triangles mark volcanic eruptions with a volcanic explosivity index of at least 5 (Simkin and Siebert 1994). (b) Estimates, with uncertainties, of the time series of the spatial mean smoothed by a 9-point Hanning window. In the case of RegEM, the uncertainty bounds are simply those from (a) smoothed by a 9-point Hanning window.

that exactly matches the sample covariance between the two time series. With time series at additional locations, the match cannot be perfect, but there is still a trade-off between the two parameters, which results in the posterior draws being correlated. This raises the concern that inference on these two parameters will be ill defined, in the sense that the posterior draws will converge toward this curve and then will wander in its vicinity over significant ranges of the two parameters. In practice, however, we find that the posterior distributions of these parameters are well constrained, having high probability over a narrow range of each (Figs. 8 and 9).

The posterior estimates of the instrumental observational error variance τ_I^2 are virtually identical for all experiments (Fig. 8 and Table 3), which is to be expected as each experiment involves the same set of instrumental observations.

The three parameters that link the pseudoproxies to the true values (β_1 , β_0 , and τ_P^2) are specified in the construction of the pseudoproxy time series, and these true values can be compared to the posterior estimates from BARCAST (Fig. 8 and Table 3). For each of these three parameters, the posterior credible intervals are narrowest for the easy experiment and become wider as the SNR and the number of proxies decrease, indicating that with more and higher quality data BARCAST can better constrain these parameters. We estimate a 90% credible interval for each of these three parameters in each of the nine experiments, and only 16 of these 27 intervals contain

the true value of the parameter (Table 3). While fewer than 90% of the credible intervals contain the specified values—an indication that the BARCAST model assumptions are imperfect—the posterior distributions are generally peaked within $\pm 15\%$ of the values used to construct the data. Note that the estimates of τ_P^2 should be biased high by $\beta_1^2 \tau_I^2$ relative to the specified values (Table 3) as the pseudoproxies are constructed by adding noise to instrumental observations, rather than the true field values. The value of $\beta_1^2 \tau_I^2$ is, however, at least two orders of magnitude smaller than the values of τ_P^2 , so this bias cannot account for the low coverage rate of the 90% intervals.

We are reassured that even in the case of the hard experiment, which involves 10 pseudoproxies with signal-to-noise ratios of $1/3$, and 65 years of overlap with the CRU data, the posterior distributions from BARCAST for the parameters linking the proxy observations to the field are narrow relative to the priors. As the overlap between the two types of data will in general be longer in real applications than in these experiments, the parameters linking the proxies to the underlying field should be well constrained in many practical circumstances.

For fixed values of the SNR, the 90% credible intervals for τ_P^2 generally become narrower and shift toward higher values as the number of proxy records increases, indicating that, with more proxy time series, BARCAST ascribes more of the variance in these time series to observational noise (Table 3). For a fixed number of

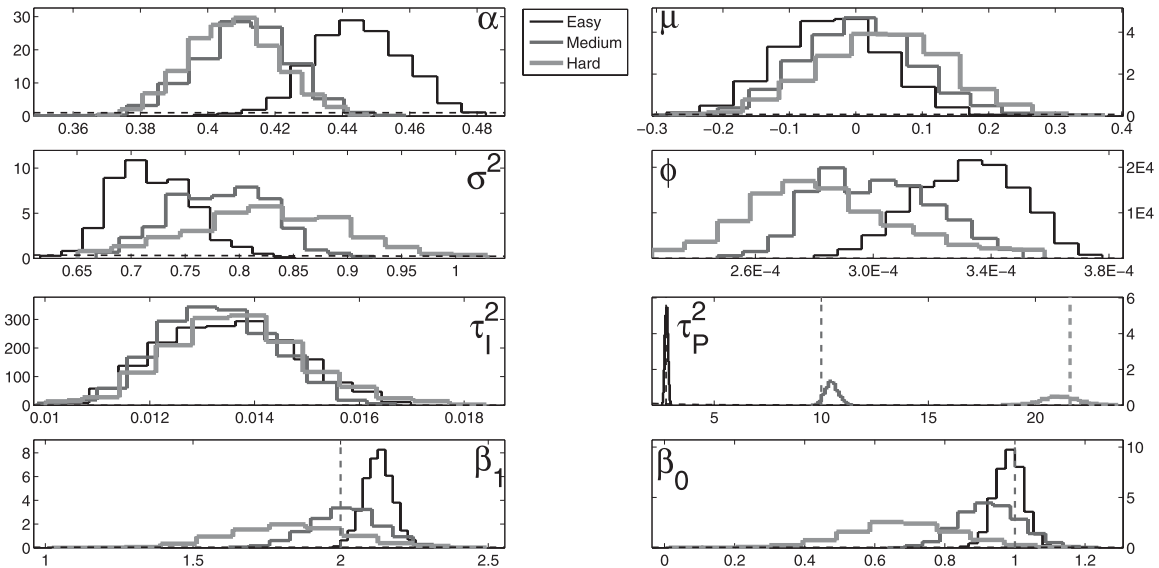


FIG. 8. Posterior histograms of the eight scalar parameters estimated by BARCAST (see Table 1), for each of the three named experiments. Priors are shown in thin dotted gray but are in most cases not discernible from zero. In the last three panels (τ_P^2 , β_1 and β_0), the values used to construct the pseudoproxies are indicated with vertical dashed lines.

proxy records, the 90% credible intervals for β_1 become narrower and shift toward higher values as the SNR increases, indicating that with smaller observational errors in the proxy time series, BARCAST infers a stronger relationship between the proxies and the underlying field. The intervals for β_1 are less sensitive to changes in the number of proxy records for a fixed value of the SNR (Table 3). The widths of the 90% credible intervals for β_0 increase as both the SNR and the number of proxy records decrease, but there is no clear pattern in the locations of these intervals.

In each experiment with BARCAST, the first 200 iterations of the Gibbs sampler are withheld from the posterior analysis, as this number appears sufficient to allow the algorithm to converge to the correct area of probability space (Fig. 9). We find that, in these examples and in practical applications, there is generally sufficient data that the algorithm readily and clearly converges to the correct area of probability space, provided that the values of the variance parameters are not initially allowed to inflate to very high values. For a general discussion concerning the convergence of Markov chain-based sampling algorithms, see Gelman et al. (2003).

c. Residual quantities

To check if the dataset is in agreement with the model assumed by BARCAST, or if more complexity is needed to account for structures in the data, we can examine several residual quantities (Fig. 10). This is akin to checking that the residuals from a simple linear regression are independent and normally distributed. From the

observation equations [Eqs. (4) and (5)], we estimate the time series of instrumental and proxy observational errors for each posterior draw, leading to both point estimates of these quantities and estimates of the associated uncertainties. By assumption, the observational error sequences should be iid normal draws, with common variance for all proxy error series and all instrumental error series, respectively. If one proxy record was biased in some way, or displayed considerably larger observational error than the others, this residual analysis should show that the observational error estimates for that proxy time series are inconsistent with those for the other proxies. As the pseudoproxies currently under analysis were constructed to be consistent with the modeling assumptions [Eq. (5)], such a feature is not observed (Fig. 10).

From the field evolution equation [Eq. (1)], we estimate the multivariate time series of innovations driving the AR(1) process for each posterior draw (Fig. 10). The uncertainty in the estimates of the innovations depends strongly on the amount of available data, being smallest at times and places where instrumental observations are available, and largest where no observations are available. By assumption, the innovation vectors for each year are iid samples from a multivariate normal distribution with mean zero and exponentially decaying spatial covariance. If the simple AR(1) model cannot account for the structures in the datasets, we would expect to see patterns arising in the innovation time series. For example, if the innovations for years immediately following major volcanic eruptions were uniformly negative, we would conclude that volcanic forcing violates the

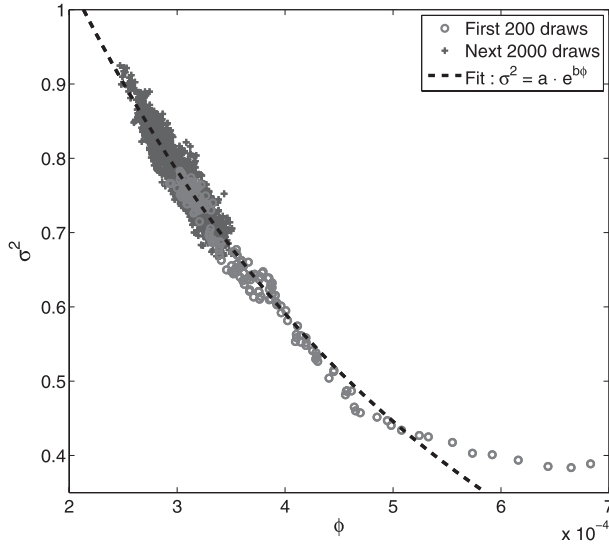


FIG. 9. Scatterplot of the draws of ϕ and σ^2 from the medium experiment, differentiating between the first 200 and the subsequent 2000, which are used in the analysis. Also shown is a fit to the later 2000 points based on the relationship between ϕ and σ^2 in Eq. (2).

AR(1) assumption [Eq. (1)]. Similarly, if the twentieth-century warming is incompatible with the AR(1) assumption, then more of the innovations over this interval should be positive than would be expected by chance alone. While this phenomenon is not observed in these experiments, we find in more realistic applications that more innovations are positive over the twentieth century than would be expected by chance alone (Tingley 2009).

The field estimates produced by BARCAST are influenced by both the model assumptions and the available data. The assumption of a stationary AR(1) process is almost certainly a simplification, and the system's response to volcanism and the increases in atmospheric greenhouse gas concentrations over the twentieth century are obvious violations of this assumption. In realistic applications, then, we expect that the modeling assumptions will not be strictly met by the data under analysis, and signatures of volcanism or twentieth-century warming in the innovations can be interpreted in two ways: first, as indicating that the model assumptions are not met by the data and thus the model should be modified (see below); second, as indicating that the field estimates are capturing important and known features of the system that violate the assumption of stationarity. In other words, signatures of nonstationarity in the residuals indicate that the field estimates produced by BARCAST are dominated by the data, rather than the model assumptions, and are robust to departures from these modeling assumptions.

d. Assessing the reconstructions

To assess the agreement between each reconstruction and the withheld CRU values we utilize both the r^2 coefficient of determination (e.g., Zar 1999) and the coefficient of efficiency (CE) statistic (see, e.g., Rutherford et al. 2005 and references therein):

$$CE = 1 - \frac{\sum_{i=1}^N (\hat{y}_i - y_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}. \quad (12)$$

If each estimate \hat{y}_i is set to the mean \bar{y}_i of that variable over the testing interval, then the CE is zero. A positive value indicates that the reconstruction contains information about the variation of the true values about the mean. If the actual and reconstructed time series have similar shapes, then r^2 will be large, but if the amplitudes of the features or the mean values are different, then the CE will be small or negative.

The use of the r^2 and CE statistics comes with the caveat that we are comparing the reconstructed values to the withheld CRU data, despite the fact that BARCAST explicitly includes a measurement error for the CRU observations. This error, however, is extremely small, as evidenced by the posterior distribution of τ_P^2 and the excellent agreement between the CRU and reconstructed values at times and places where CRU observations are used in the reconstruction (Figs. 6 and 8). To provide intuition with regards to these statistics, we indicate the values for the representative time series discussed above for the BARCAST analysis of the medium dataset (Fig. 6).

For each of the named experiments (Table 4), and for analysis with both BARCAST and RegEM, we plot both the r^2 and the CE statistics as a function of location (Fig. 11). These maps are incomplete, as the statistics are only calculated at those locations where there are at least 10 years of CRU observations during the 1850–1940 testing interval. Both the CE and r^2 values are strongly influenced by the number of nearby time series of observations. In the case of BARCAST, this is consistent with the exponential form specified for the spatial covariance, which anticipated predictions being more precise at locations close to observations. The similar spatial patterns seen in the maps of the CE and r^2 statistics from RegEM and BARCAST supports the use of the simple spatial covariance form specified for BARCAST.

As expected, regardless of the analysis technique (BARCAST or RegEM) or statistic (r^2 , CE), the agreement between the reconstruction and the withheld values is greatest for the easy experiment, decreases as both the SNR and the number of proxy records decreases, and is

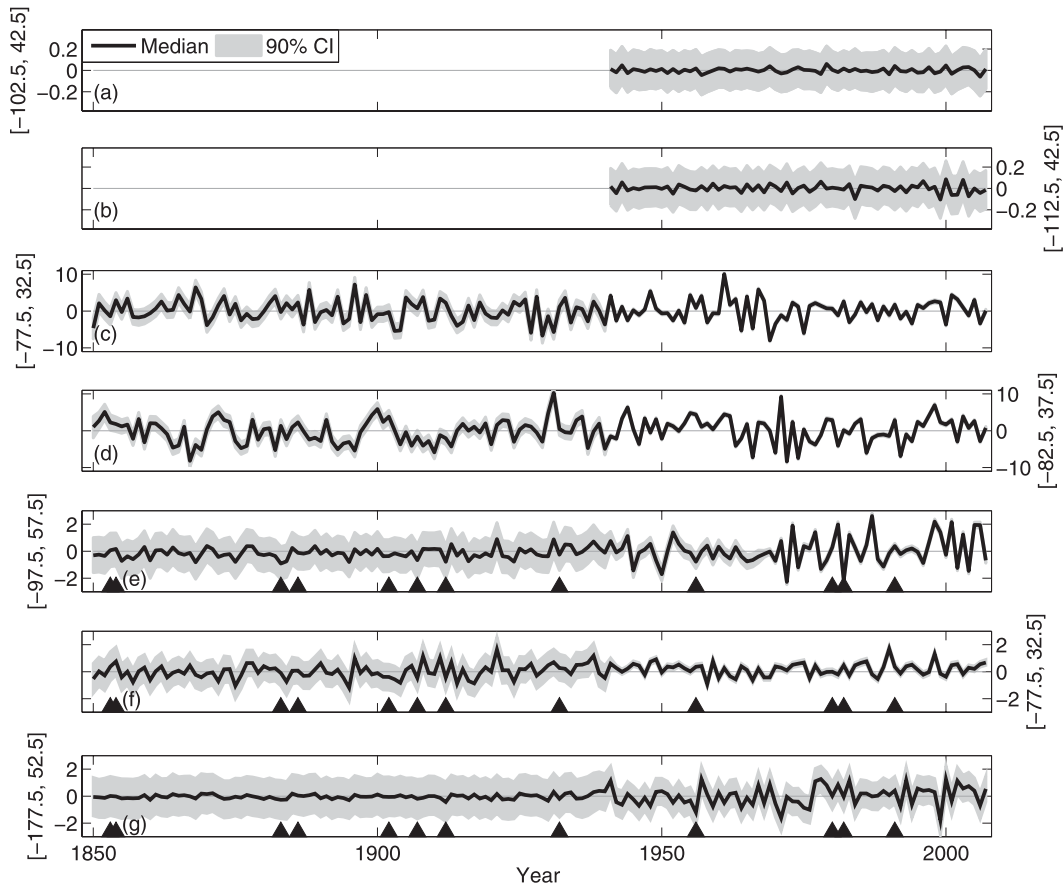


FIG. 10. Estimates of the observational error sequences and innovation time series at selected locations, for the medium experiment. The light gray fill shows the 90% credible intervals from BARCAST, and the black lines the medians. Locations are given on the y axes as latitude–longitude pairs. (a),(b) Instrumental observational error sequences (in $^{\circ}\text{C}$) at two locations; (c),(d) proxy observational error sequences (in proxy units) at two locations; (e)–(g) innovation time series (in $^{\circ}\text{C}$) at locations where there are, respectively, only instrumental observations, both instrumental and proxy observations, and no observations. The black triangles in (e)–(g) mark volcanic eruptions with a volcanic explosivity index of at least 5 (Simkin and Siebert 1994).

smallest for the hard experiment (Table 4 and Fig. 11). In addition, the CE values are uniformly smaller than the corresponding r^2 values, indicating that both algorithms are better at inferring the correct shape of the variability than they are at inferring the correct amplitude or mean value (see Christiansen et al. 2009 for a general discussion of this issue). For each experiment, the average r^2 and CE values for BARCAST are higher than those for RegEM, indicating that BARCAST produces better estimates of both the shape and the amplitude of the withheld CRU time series (Table 4; see also Fig. 7) in these particular experiments.

Finally, both BARCAST and RegEM produce uncertainty intervals associated with the estimates of the withheld CRU surface temperature observations, and we calculate the fraction of the withheld CRU values that fall within the 90% uncertainty intervals for each experiment and analysis method (Table 4). Ideally, 90%

of the withheld values should fall within the 90% uncertainty intervals, while the extents to which the coverage rates differ from 90% are indications of errors in the estimated confidence or credible intervals. In each experiment, the 90% credible intervals produced using BARCAST cover nearly 90% of the withheld values, but those produced using RegEM cover only about 77% of the withheld values, indicating that the BARCAST uncertainty intervals have the correct coverage rate, while those from RegEM are too narrow. The coverage rates do not appear sensitive to either the SNR or the number of proxy records. In the case of BARCAST, the credible intervals for missing instrumental values are constructed by adding white noise to each draw of the field, with variance given by the corresponding draw of τ_i^2 , and then taking the 5th and 95th percentiles of the resulting distributions. The distinction between uncertainty intervals for missing instrumental observations

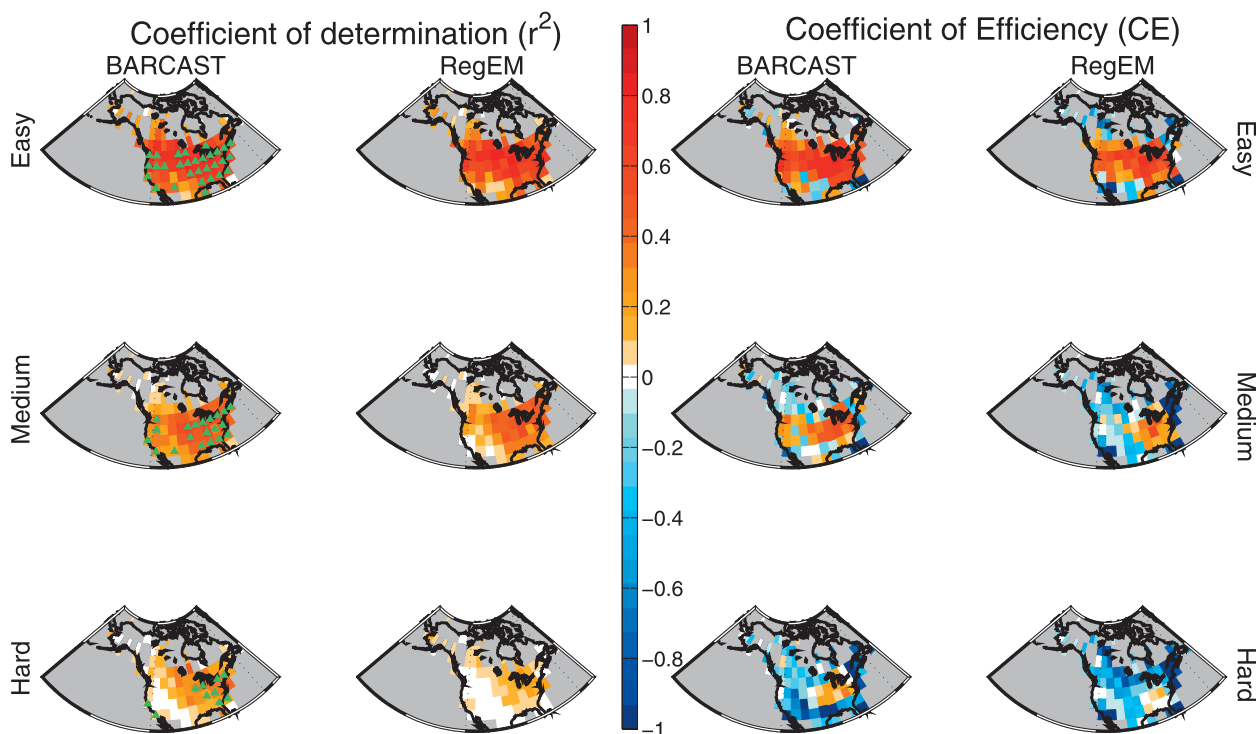


FIG. 11. Measures of skill for the three named experiments, using both BARCAST and RegEM to perform the reconstructions. At each location where there are at least 10 CRU observations in the pre-1941 testing interval, both the r^2 and CE between the reconstructed time series and the CRU time series are plotted. To ensure a consistent color scheme, where warm colors indicate positive values of the statistics, all CE values less than or equal to -1 are plotted using the same color. Locations of the pseudoproxy time series used in each experiment are indicated by green triangles in the leftmost column of figures.

and those for the underlying true values will be discussed in Part II.

The RegEM algorithm requires the specification of a variance inflation factor (Schneider 2001), set to a default of one in these experiments. Given the low coverage rate of the 90% confidence intervals, the RegEM variance inflation factor apparently ought to be set to a higher value, where that value would be determined by repeated numerical experiments (Schneider 2001). We leave the variance inflation factor at one to stress the need to accurately estimate this additional parameter when using RegEM. This and related points are addressed in more detail in Part II.

4. Shortcomings and extensions

For the sake of simplicity, we have described and demonstrated the simplest form of BARCAST that can be directly applied to actual paleoclimate problems. This basic analysis scheme can be extended and modified in a number of potentially useful ways, and doing so will involve modifying the conditional posteriors derived in appendix B. There are also shortcomings and limitations to BARCAST, some of which are particular to the

Bayesian approach to reconstructing climate fields and some of which apply more broadly. We review a number of existing shortcomings and possible extensions below.

a. Shortcomings

1) COMPUTATIONAL REQUIREMENTS

One of the drawbacks of the Bayesian method is that it is computationally intensive. A minimum of several hundred draws from the posterior are necessary (e.g., Gelman et al. 2003), and for each posterior draw and each unique pattern of missing data, a matrix with dimension given by the number of spatial locations used in the analysis must be inverted. Although BARCAST is slower than RegEM, the required computation is small relative to many earth science problems, comparable to running a coarse-resolution general circulation model. Parallel processing can be exploited in several ways to increase the speed of the analysis. It is possible, for example, to run one chain to convergence, and then use the output from this chain to initialize a number of other chains. Alternatively, large-scale reconstructions can be performed region by region, retaining all available observations that lie within a certain number of e -folding length scales,

given by $1/\phi$, of the boundaries of each region (cf. Cook et al. 1999).

2) TEMPORAL VARIANCE

Climate reconstruction approaches based on least squares regression, which minimize the error sum of squares, result in reconstructions (of a field or a spatial average) with lower temporal variance over the proxy interval than over the calibration interval (NRC 2006). This is a feature of all ordinary least squares regression models, including BARCAST, and is not necessarily a flaw. The regression equation predicts the mean value of the response, given the predictors, so does not take into account the variability of the response about that mean value. Method-of-moment-type approaches avoid this shortcoming by scaling the mean and standard deviation of the proxy part of the reconstruction to match those of the instrumental over the calibration interval (NRC 2006). Such approaches, however, do not minimize the error sum of squares and so are not optimal from the standpoint of minimizing prediction error (e.g., Casella and Berger 2002). In addition to providing a best estimate of the field evolution, and an estimate of the associated uncertainty, BARCAST results in an ensemble of draws of the space-time field, each consistent with the data and the model assumption, that have, on average, the correct temporal variance; this issue is investigated in Part II.

3) DISTRIBUTION OF ERRORS

Both the instrumental and proxy observation equations [Eqs. (4) and (5)] assume that the errors are iid normal draws, which is unlikely to be true in practice. In particular, errors in the proxies could well be correlated in both space and time. A number of studies have explored the sensitivity of various climate reconstruction techniques to the type of proxy noise (e.g., von Storch et al. 2009), and similar exercises will need to be conducted for BARCAST.

4) STANDARDIZATION

The standardization applied to the data prior to analysis, which can be thought of as the estimation of three additional parameters for either the proxy dataset as a whole, or for each proxy time series, can affect the skill of field reconstructions (e.g., NRC 2006; Mann et al. 2007; Smerdon and Kaplan 2007; Lee et al. 2008; Smerdon et al. 2008). In general, the uncertainty introduced by the error in estimating the mean and standard deviation of the proxy dataset, or each time series, is not propagated through the analysis, leading to confidence intervals that are too narrow. This issue is discussed in more detail below.

b. Extensions

1) MULTIPLE OBSERVATION EQUATIONS

The current model assumes that all proxy records share the same observation equation [Eq. (5)]. To account for differences between proxy types, say tree ring width and tree ring density time series, a separate observation equation can be specified for each (see Tingley 2009 for an example application). This introduces a triplet of parameters (β_0 , β_1 , τ_P^2) for each proxy type, for which we must specify priors and derive conditional posteriors.

2) PROXIES THAT AVERAGE OVER TIME OR SPACE

Some climate proxies, such as pollen assemblages, contain information about the field integrated or averaged over time and/or space, and the proxy observation equation can be generalized to reflect these relationships. As a simple example, suppose that a proxy observation is assumed to have a statistically linear relationship with the local mean temperature over the previous two years (Fig. 1). We can then specify the observation equation for this proxy type as

$$\mathbf{W}_{P,t} = \sum_{k=0}^1 \beta_{k+1} \mathbf{T}_{P,t-k} + \beta_0 \mathbf{1} + \mathbf{e}_{P,t}. \quad (13)$$

The resolution of the proxy time series—the spacing in time between subsequent data points—is not an issue, as BARCAST is constructed to handle incomplete datasets. The observation equation can likewise be modified to account for a proxy that reflects the underlying field averaged over some spatial domain.

3) GENERALIZING THE SPATIAL COVARIANCE STRUCTURE

The exponential covariance function is a special case of the more general Matérn class of covariance functions, with the smoothness parameter set to 0.5 (Banerjee et al. 2004). Realizations from such a process are continuous, but not differentiable. If this is insufficiently smooth to accurately model the underlying field, the exponential covariance can be replaced with a Matérn covariance function with a larger smoothness parameter.

We have assumed an isotropic covariance function for the spatial field. In the context of modeling climate fields, it might be reasonable to assume that covariance decays at different rates in the zonal and meridional directions, which leads to a more generalized covariance function,

$$\Sigma_{ij} = \sigma^2 \exp(-\phi_{\text{lon}}|x_i - x_j| - \phi_{\text{lat}}|y_i - y_j|), \quad (14)$$

where $|x_i - x_j|$ and $|y_i - y_j|$ are the separations in the zonal and meridional directions, respectively, of the i th and j th locations. Similar modifications can be used to

describe different rates of decay between locations over land, locations over sea, and locations that straddle a coastline.

4) INCORPORATING CLIMATE FORCINGS

Reconstructions of such climate forcings as volcanism, solar variability, and greenhouse gas concentrations can be added into the model, at the process level, by replacing the temperature evolution equation [Eq. (1)] with

$$\begin{aligned} \mathbf{T}_t &= \mu \mathbf{1} - a_1 V_t - a_2 S_t - a_3 G_t \\ &= \alpha (\mathbf{T}_{t-1} - \mu \mathbf{1} - a_1 V_{t-1} - a_2 S_{t-1} - a_3 G_{t-1}) + \epsilon_t, \end{aligned} \quad (15)$$

where V , S , and G are the time series of volcanic, solar, and greenhouse gas forcings, respectively. If these forcings are important, then a formulation that excludes them will result in posterior estimates of the innovation time series that are not iid normal draws. Including these forcings should reduce or eliminate any structures in the innovations. In addition, posterior estimates of the a_i provide estimates of the sensitivity of the climate system to changes in the forcings. If we include these forcings, assume that the instrumental observational error variance is zero, and consider co-located instrumental and proxy composite time series, then the modeling assumptions reduce to those assumed by Lee et al. (2008). The reconstruction method employed by Lee et al. (2008) uses a Kalman filter to estimate the instrumental composite when only the proxy composite is available and makes use of a separate analysis to estimate the various parameters that specify the model.

5) PREDICTIONS

As BARCAST includes a temporal model, it is possible to forecast the field for years after the last, or before the first, observations. The skill of such a forecast is set by the size of α , which gives the lag-one correlation. In the examples presented above, α is about 0.4, which corresponds to an r^2 of about 0.16 for a one-step-ahead prediction, indicating that the potential for forecasting is weak. Including time series of the forcing into the analysis would increase the forecasting skill as the mean value to which the forecast regresses is then time dependent.

6) HIERARCHY OF REGRESSION COEFFICIENTS

As discussed earlier, the standardization applied to the proxy time series prior to the analysis can influence the results of reconstructions.

If each proxy time series of a particular type is individually standardized to have mean zero and standard deviation one, the regression coefficients that locally link the true field time series to these proxy time series

should be similar, but not identical. An intermediate between estimating a common set of regression coefficients for each proxy time series of a particular type, or a different set for each, can be constructed within the Bayesian framework by assuming that the regression parameters and observational error variances for each proxy time series are, respectively, draws from a common distribution. Building a hierarchy on these parameters results in estimates of the regression parameters that are weighted means of those resulting from separate estimates of the coefficients for each proxy time series and the estimate of one common set of coefficients for all proxy time series (e.g., Gelman et al. 2003). This approach is akin to including the standardization step directly into the analysis and thus facilitates the propagation of uncertainties.

5. Conclusions

We present a new methodological framework for reconstructing the temporal evolution of climate fields from incomplete data. The Bayesian scheme provides a rich output, including the full covariance uncertainty structure associated with the estimates of the true field values through time and the scalars that parameterize the model. The posterior draws of the space–time field can be used to estimate the posterior distribution of any function of the field, from simple measures like the spatial mean to more complex measures such as the maximum value for a given year. While the median of the posterior draws of the time series of spatial means (or the time series at a given location) will have, on average, smaller temporal variance than the corresponding quantity calculated from the true field, the members of the ensemble of posteriors draws have, on average, the correct variance. By specifying a simple parametric form for the spatial covariance matrix, BARCAST can predict the field at locations where there are no observations, while the specified autocorrelation means that the estimates of the field for a given year are influenced by observations from neighboring years. The model assumptions can also be checked a posteriori, similar to the residual analysis of a standard regression.

Demonstrations of the new method using pseudo-proxy datasets constructed from the CRU temperature compilation show that BARCAST produces reasonable results from reasonable data. Three measures of skill—the r^2 , CE, and percentage of withheld values covered by 90% uncertainty intervals—indicate that BARCAST outperforms the RegEM algorithm in these experiments. Part II presents more theoretical comparisons between these two algorithms, which complement the practical comparisons presented here.

The methodology we have developed is immediately applicable to any number of paleoclimate reconstruction problems. One of the advantages of the hierarchical Bayesian analysis is the ease with which it can be extended and generalized within a cohesive conceptual framework. For the purposes of illustrating the approach, we have made a simple set of reasonable assumptions and indicated a number of possibly useful extensions.

Acknowledgments. The content and presentation of this manuscript benefited from discussions with E. Butler, A. Dempster, B. Farrell, A. Rhines, T. Schneider, J. Smerdon, S. Wofsy, and C. Wunsch and from the comments of four anonymous reviewers. Simulations were run on the Odyssey cluster supported by the FAS Research Computing Group at Harvard University. Funding for this work was provided by NSF Grant ATM-0902374.

APPENDIX A

Choosing the Hyperparameters for the Priors

To provide closure to the hierarchical model, we must place prior distributions on all elements of Θ , as well as on \mathbf{T}_0 , the vector of true temperature values for the year before the first observations. Our approach is to put independent and weakly informative proper priors on all parameters and to show that the posteriors are dominated by the data. Where possible, the priors are selected to be conditionally conjugate. The use of uninformative, improper priors in a model as complex as this is difficult, as there is then no guarantee the posterior will be proper, and checking is an onerous procedure (Gelman et al. 2003). Below we explain our choices for the hyperparameters, which tend toward those that are simplest, and provide the values used for the examples in section 3.

- $\mathbf{T}_0 \sim N(\mu_o, \Sigma_o)$. The simplest approach is to set $\mu_o = 0$ and $\Sigma_o = \sigma_o^2 \cdot \mathbf{I}$, and then specify the prior standard deviation σ_o as some multiple (2 in the case of the examples) of the standard deviation estimated from all available instrumental data. Technically, this approach involves a double use of the instrumental data, which is used to set the hyperparameters and to estimate the elements of Θ and \mathbf{T} . This is not a concern, however, as the data are only used to ensure that the prior variance for \mathbf{T}_0 is similar to but larger than the variance estimated from all instrumental data, and thus the prior provides a weak but reasonable constraint on the posterior estimate of \mathbf{T}_0 .
- $\alpha \sim \text{uniform}(a_0, a_1)$. The simplest choice is to set $a_0 = 0$ and $a_1 = 1$, which is done in the examples.

- $\mu \sim N(\mu_0, \sigma_0^2)$. We set the prior mean μ_0 to the mean of all instrumental observations, and then set the prior standard deviation, σ_0 , to a high value (5 in the examples). The use of the data to set μ_0 is not an issue because of the corresponding high value of σ_0 . Our approach is to use the data to ensure that the prior is centered near the right area, but to then use a very large prior variance to ensure that the posterior is dominated by the data.
- $\sigma^2 \sim \text{inverse-gamma}(\lambda_T, \nu_T)$. In other words, $\mathbf{P}(\sigma^2) \propto (\sigma^2)^{-(\lambda_T+1)} \cdot \exp(-\nu_T/\sigma^2)$. This prior is conditionally conjugate and corresponds to $2\lambda_T$ prior observations with an average squared deviation of ν_T/λ_T (Gelman et al. 2003). Provided that λ_T is small and ν_T reasonable (they are both set to one-half in the examples), the prior has little influence on the posterior.
- $\phi \sim \text{log-normal}(\mu_\phi, \sigma_\phi^2)$, so that $\mathbf{P}(\phi) \propto (1/\phi) \cdot \exp(-(\ln \phi - \mu_\phi)/2\sigma_\phi^2)$. For many large-scale climate reconstruction problems, the spatial correlation distance is expected to be, broadly speaking, somewhere between 10 and 1000 km, so the log of the inverse distance should be between about -7 and -2.3 ; a low information prior can be set accordingly. In the examples we set $\mu_\phi = -4.65$ and $\sigma_\phi^2 = 1.2$. This prior is not conditionally conjugate.
- $\tau_I^2 \sim \text{inverse-gamma}(\lambda_I, \nu_I)$ and $\tau_P^2 \sim \text{inverse-gamma}(\lambda_P, \nu_P)$. These are treated in the same way as the prior for σ^2 , by setting the prior parameters to correspond to a very small number of prior observations with a reasonable sum of squares, and showing that the data overwhelms the prior. In the examples, we set all four of these parameters to one-half.
- $\beta \sim N(\eta_1, \delta_1^2)$. Following Eq. (11), we set the prior mean to

$$\eta_1 = \left[\frac{(1 - \tau_P^2)(1 - \alpha^2)}{\sigma^2} \right]^{-1/2},$$

using the prior modes of τ_P^2 , α , and σ^2 , and then set the prior standard deviation δ_1 to something suitably high (8 in the examples). If each proxy time series is assumed a priori to be strictly positively correlated with the local value of the underlying field, then the prior for β_1 can be set to a truncated normal to reflect this information. Our strategy is to use a broader prior, which puts nonzero weight on the whole real line. If the resulting posterior has substantial weight less than zero, we would interpret this as an indication that the proxies cannot reliably constrain the true field values.

- $\beta_0 \sim N(\eta_0, \delta_0^2)$ following Eq. (11), we set the prior mean to the negative of the product of the prior means for μ and β_1 , and then set the prior standard deviation δ_0 to something suitably high (8 in the examples).

APPENDIX B

Conditional Posteriors

We detail the full conditional posterior of each unknown (Table 1), using the notation $X|\cdot$ to indicate the conditional distribution of X given all other applicable aspects of the model (parameters and data).

The conditional posterior of \mathbf{T}_0 is multivariate normal:

$$\mathbf{T}_0|\cdot \sim N(\Psi_0 \mathbf{V}_0, \Psi_0), \quad \text{where} \quad (\text{B1})$$

$$\mathbf{V}_0 = \Sigma^{-1}[\alpha \mathbf{T}_1 - \alpha(1 - \alpha)\mu \mathbf{1}] + \Sigma_o^{-1}\mu_o, \quad \text{and} \quad (\text{B2})$$

$$\Psi_0 = (\alpha^2 \Sigma^{-1} + \Sigma_o^{-1})^{-1}. \quad (\text{B3})$$

The conditional posteriors of the \mathbf{T}_k , $0 < k < \kappa$ are multivariate normal:

$$\mathbf{T}_k|\cdot \sim N(\Psi_k \mathbf{V}_k, \Psi_k), \quad \text{where} \quad (\text{B4})$$

$$\mathbf{V}_k = \mathbf{H}_k^T \Xi_k^{-1}(\mathbf{W}_k - \mathbf{B}_k) + \Sigma^{-1}[\alpha(\mathbf{T}_{k+1} + \mathbf{T}_{k-1}) + (1 - \alpha)^2 \mu \mathbf{1}], \quad \text{and} \quad (\text{B5})$$

$$\Psi_k = [\mathbf{H}_k^T \Xi_k^{-1} \mathbf{H}_k + (1 + \alpha^2) \Sigma^{-1}]^{-1}. \quad (\text{B6})$$

The conditional posterior of \mathbf{T}_κ is multivariate normal:

$$\mathbf{T}_\kappa|\cdot \sim N(\Psi_\kappa \mathbf{V}_\kappa, \Psi_\kappa), \quad \text{where} \quad (\text{B7})$$

$$\mathbf{V}_\kappa = \mathbf{H}_\kappa^T \Xi_\kappa^{-1}(\mathbf{W}_\kappa - \mathbf{B}_\kappa) + \Sigma^{-1}[\alpha \mathbf{T}_{\kappa-1} + (1 - \alpha) \mu \mathbf{1}], \quad (\text{B8})$$

and

$$\Psi_\kappa = (\mathbf{H}_\kappa^T \Xi_\kappa^{-1} \mathbf{H}_\kappa + \Sigma^{-1})^{-1}. \quad (\text{B9})$$

The conditional posterior of α is a truncated normal, where the bounds in subscripted square brackets indicate the region of nonzero probability resulting from the uniform prior (see appendix A):

$$\alpha|\cdot \sim N_{[a_0, a_1]}(\Psi_\alpha V_\alpha, \Psi_\alpha), \quad \text{where} \quad (\text{B10})$$

$$V_\alpha = \sum_{k=1}^{\kappa} (\mathbf{T}_{k-1} - \mu \mathbf{1})^T \Sigma^{-1} (\mathbf{T}_k - \mu \mathbf{1}), \quad \text{and} \quad (\text{B11})$$

$$\Psi_\alpha = \left[\sum_{k=1}^{\kappa} (\mathbf{T}_{k-1} - \mu \mathbf{1})^T \Sigma^{-1} (\mathbf{T}_{k-1} - \mu \mathbf{1}) \right]^{-1}. \quad (\text{B12})$$

The conditional posterior of μ is normal:

$$\mu|\cdot \sim N(\Psi_\mu V_\mu, \Psi_\mu), \quad \text{where} \quad (\text{B13})$$

$$V_\mu = \frac{\mu_0}{\sigma_0^2} + (1 - \alpha) \mathbf{1}^T \Sigma^{-1} \left[\sum_{k=1}^{\kappa} (\mathbf{T}_k - \alpha \mathbf{T}_{k-1}) \right], \quad \text{and} \quad (\text{B14})$$

$$\Psi_\mu = \left(\frac{1}{\sigma_0^2} + \kappa(1 - \alpha)^2 \mathbf{1}^T \Sigma^{-1} \mathbf{1} \right)^{-1}. \quad (\text{B15})$$

The conditional posterior of σ^2 is inverse-gamma:

$$\sigma^2|\cdot \sim \text{inverse-gamma} \left(\lambda_T + \frac{N_A \kappa}{2}, \nu_T + \frac{1}{2} \sum_{k=1}^{\kappa} \Delta \mathbf{T}_{k,k-1}^T \mathbf{R}^{-1} \Delta \mathbf{T}_{k,k-1} \right), \quad \text{where} \quad (\text{B16})$$

$$\Delta \mathbf{T}_{k,k-1} = (\mathbf{T}_k - \alpha \mathbf{T}_{k-1} - (1 - \alpha) \mu \mathbf{1}), \quad \text{and} \quad (\text{B17})$$

$$\mathbf{R}_{ij} = \exp(-\phi |x_i - x_j|). \quad (\text{B18})$$

The posterior of ϕ does not follow any standard distributional form:

$$P(\phi|\cdot) \propto P(\phi) \cdot |\mathbf{R}|^{-\kappa/2} \cdot \exp \left(-\frac{1}{2\sigma^2} \sum_{k=1}^{\kappa} \Delta \mathbf{T}_{k,k-1}^T \mathbf{R}^{-1} \Delta \mathbf{T}_{k,k-1} \right), \quad (\text{B19})$$

where $P(\phi)$ is the prior for ϕ , and \mathbf{R} is a function of ϕ . We place a log-normal prior on ϕ , and, after suitable transformation, sample the posterior distribution of $\log(\phi)$ using a Metropolis step. Transforming to $\log(\phi)$ facilitates the Metropolis step (Gelman et al. 2003), as $\log(\phi)$ has non-zero probability on the real line, so we can use a symmetric jumping distribution. The posterior for $\Phi \equiv \log(\phi)$ is

$$P(\Phi|\cdot) \propto |\mathbf{R}|^{-\kappa/2} \cdot \exp \left(\frac{-(\Phi - \mu_\phi)^2}{2\sigma_\phi^2} - \frac{1}{2\sigma^2} \sum_{k=1}^{\kappa} \Delta \mathbf{T}_{k,k-1}^T \mathbf{R}^{-1} \Delta \mathbf{T}_{k,k-1} \right). \quad (\text{B20})$$

We use a normal jumping distribution for the Metropolis step to sample from the posterior, $\Phi^*|\Phi_{t-1} \sim N(\Phi_{t-1}, \sigma_{\Phi, MH}^2)$, adjusting the jumping variance parameter $\sigma_{\Phi, MH}^2$ so the acceptance rate is about 40% (Gelman et al. 2003), and exponentiate to transform back to ϕ .

The conditional posterior of τ_I^2 is inverse-gamma:

$$\tau_I^2|\cdot \sim \text{inverse-gamma} \left(\lambda_I + \frac{M_I}{2}, \nu_I + \frac{1}{2} \sum_{k=1}^{\kappa} \mathbf{r}_{I,k}^T \mathbf{r}_{I,k} \right), \quad \text{where} \quad (\text{B21})$$

$$\mathbf{r}_{I,k} = \mathbf{W}_{I,k} - [\mathbf{H}_k \mathbf{T}_k + \mathbf{B}_k]_I. \quad (\text{B22})$$

The subscript I denotes the subset of the bracketed vector corresponding to the instrumental observations. Simply put, $\mathbf{r}_{I,k}$ is the vector of residuals between the instrumental observations and the true temperatures at time t , and the posterior distribution for τ_I^2 depends on the sum of the squares of these residuals, which (properly scaled) is an estimate of the instrumental observational error variance.

The conditional posterior of τ_P^2 is also inverse-gamma, and similar to that for τ_I^2 :

$$\tau_P^2 | \cdot \sim \text{inverse-gamma} \left(\lambda_P + \frac{M_P}{2}, \nu_P + \frac{1}{2} \sum_{k=1}^K \mathbf{r}_{P,k}^T \mathbf{r}_{P,k} \right), \quad (\text{B23})$$

where

$$\mathbf{r}_{P,k} = \mathbf{W}_{P,k} - [\mathbf{H}_k \mathbf{T}_k + \mathbf{B}_k]_P. \quad (\text{B24})$$

The subscript P denotes the subset of the bracketed vector corresponding to the proxy observations. Similar to $\mathbf{r}_{I,k}$, $\mathbf{r}_{P,k}$ is the vector of residuals between the proxy observations and the transformed true temperatures at time t , and the posterior distribution for τ_P^2 depends on the sum of the squares of these residuals, which (properly scaled) is an estimate of the proxy observational error variance.

The conditional posterior of β_1 , is normal:

$$\beta_1 | \cdot \sim N(\Psi_{\beta_1} V_{\beta_0}, \Psi_{\beta_1}), \quad \text{where} \quad (\text{B25})$$

$$V_{\beta_1} = \frac{\eta_1}{\delta_1^2} + \frac{1}{\tau_P^2} \sum_{k=1}^K \mathbf{T}_k^T (\mathbf{W}_{P,k} - \beta_0 \mathbf{1}_{N_{P,k}}), \quad \text{and} \quad (\text{B26})$$

$$\Psi_{\beta_1} = \left(\frac{1}{\delta_1^2} + \frac{1}{\tau_P^2} \sum_{k=1}^K \mathbf{T}_{P,k}^T \mathbf{T}_{P,k} \right)^{-1}. \quad (\text{B27})$$

The conditional posterior of β_0 is normal:

$$\beta_0 | \cdot \sim N(\Psi_{\beta_0} V_{\beta_0}, \Psi_{\beta_0}), \quad \text{where} \quad (\text{B28})$$

$$V_{\beta_0} = \frac{\eta_0}{\delta_0^2} + \frac{1}{\tau_P^2} \sum_{k=1}^K \mathbf{1}_{N_{P,k}}^T (\mathbf{W}_{P,k} - \beta_1 \mathbf{H}_{P,k} \mathbf{T}_k), \quad \text{and} \quad (\text{B29})$$

$$\Psi_{\beta_0} = \left(\frac{1}{\delta_0^2} + \frac{M_P}{\tau_P^2} \right)^{-1}. \quad (\text{B30})$$

REFERENCES

- Banerjee, S., B. P. Carlin, and A. E. Gelfand, 2004: *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall, 472 pp.
- Brockwell, P., and R. Davis, 1991: *Time Series: Theory and Methods*. 2nd ed. Springer, 577 pp.
- Brohan, P., J. J. Kennedy, I. Harris, S. F. B. Tett, and P. D. Jones, 2006: Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *J. Geophys. Res.*, **111**, D12106, doi:10.1029/2005JD006548.
- Casella, G., and R. Berger, 2002: *Statistical Inference*. Thomson Learning Pacific Grove, 660 pp.
- Christiansen, B., T. Schmith, and P. Thejll, 2009: A surrogate ensemble study of climate reconstruction methods: Stochasticity and robustness. *J. Climate*, **22**, 951–976.
- Cook, E., D. Meko, D. Stahle, and M. Cleaveland, 1999: Drought reconstructions for the continental United States. *J. Climate*, **12**, 1145–1162.
- Evans, M., B. Reichert, A. Kaplan, K. Anchukaitis, E. Vaganov, M. Hughes, and M. Cane, 2006: A forward modeling approach to paleoclimatic interpretation of tree-ring data. *J. Geophys. Res.*, **111**, G03008, doi:10.1029/2006JG000166.
- Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, 2003: *Bayesian Data Analysis*. 2nd ed. Chapman & Hall/CRC, 668 pp.
- Hansen, J., and S. Lebedeff, 1987: Global trends of measured surface air temperature. *J. Geophys. Res.*, **92** (D11), 13 345–13 372.
- Jones, P., and Coauthors, 2009: High-resolution palaeoclimatology of the last millennium: A review of current status and future prospects. *Holocene*, **19**, 3–49, doi:10.1177/0959683608098952.
- Kalman, R., 1960: A new approach to linear filtering and prediction problems. *J. Basic Eng.*, **82**, 35–45.
- Lee, T., F. Zwiers, and M. Tsao, 2008: Evaluation of proxy-based millennial reconstruction methods. *Climate Dyn.*, **31**, 263–281.
- Luterbacher, J., D. Dietrich, E. Xoplaki, M. Grosjean, and H. Wanner, 2004: European seasonal and annual temperature variability, trends, and extremes since 1500. *Science*, **303**, 1499–1503.
- Mann, M., and J. Park, 1993: Spatial correlations of interdecadal variation in global surface temperatures. *Geophys. Res. Lett.*, **20**, 1055–1058.
- , and P. Jones, 2003: Global surface temperatures over the past two millennia. *Geophys. Res. Lett.*, **30**, 1820, doi:10.1029/2003GL017814.
- , R. Bradley, and M. Hughes, 1998: Global-scale temperature patterns and climate forcing over the past six centuries. *Nature*, **392**, 779–787.
- , S. Rutherford, E. Wahl, and C. Ammann, 2005: Testing the fidelity of methods used in proxy-based reconstructions of past climate. *J. Climate*, **18**, 4097–4107.
- , —, —, and —, 2007: Robustness of proxy-based climate field reconstruction methods. *J. Geophys. Res.*, **112**, D12109, doi:10.1029/2006JD008272.
- , Z. Zhang, M. Hughes, R. Bradley, S. Miller, S. Rutherford, and F. Ni, 2008: Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia. *Proc. Natl. Acad. Sci. USA*, **105**, 13 252–13 257, doi:10.1073/pnas.0805721105.
- NRC, 2006: *Surface Temperature Reconstructions for the Last 2,000 Years*. The National Academies Press, 145 pp.
- Riedwyl, N., M. Küttel, J. Luterbacher, and H. Wanner, 2009: Comparison of climate field reconstruction techniques: Application to Europe. *Climate Dyn.*, **32**, 381–395.

- Rutherford, S., M. Mann, T. Delworth, and R. Stouffer, 2003: Climate field reconstruction under stationary and nonstationary forcing. *J. Climate*, **16**, 462–479.
- , —, T. Osborn, R. Bradley, K. Briffa, M. Hughes, and P. Jones, 2005: Proxy-based Northern Hemisphere surface temperature reconstructions: Sensitivity to method, predictor network, target season, and target domain. *J. Climate*, **18**, 2308–2329.
- Schneider, T., 2001: Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *J. Climate*, **14**, 853–871.
- Simkin, T., and L. Siebert, 1994: *Volcanoes of the World: A Regional Directory, Gazetteer, and Chronology of Volcanism during the Last 10,000 Years*. 2nd ed. Geoscience Press, 349 pp. [Available online at ftp://ftp.ncdc.noaa.gov/pub/data/paleo/treering/reconstructions/n_hem_temp/nhemtemp_data.txt.]
- Smerdon, J., and A. Kaplan, 2007: Comments on testing the fidelity of methods used in proxy-based reconstructions of past climate: The role of the standardization interval. *J. Climate*, **20**, 5666–5670.
- , —, and D. Chang, 2008: On the origin of the standardization sensitivity in RegEM climate field reconstructions. *J. Climate*, **21**, 6710–6723.
- Steig, E., D. Schneider, S. Rutherford, M. Mann, J. Comiso, and D. Shindell, 2009: Warming of the Antarctic ice-sheet surface since the 1957 International Geophysical Year. *Nature*, **457**, 459–462.
- Tingley, M. P., 2009: A Bayesian approach to reconstructing space–time climate fields from proxy and instrumental time series, applied to 600 years of Northern Hemisphere surface temperature data. Ph.D. dissertation, Harvard University, 230 pp.
- , and P. Huybers, 2010: A Bayesian algorithm for reconstructing climate anomalies in space and time. Part II: Comparison with the regularized expectation–maximization algorithm. *J. Climate*, **23**, 2782–2800.
- von Storch, H., E. Zorita, and F. González-Rouco, 2009: Assessment of three temperature reconstruction methods in the virtual reality of a climate simulation. *Int. J. Earth Sci.*, **98**, 67–82.
- Wikle, C. K., and L. M. Berliner, 2006: A Bayesian tutorial for data assimilation. *Physica D*, **230**, 1–16.
- Zar, J. H., 1999: *Biostatistical Analysis*. 4th ed. Pearson Education, 929 pp.
- Zhang, Z., M. Mann, and E. Cook, 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. *Holocene*, **14**, 502–516.