Adv. Stat. Clim. Meteorol. Oceanogr., 3, 1–16, 2017
www.adv-stat-clim-meteorol-oceanogr.net/3/1/2017/
doi:10.5194/ascmo-3-1-2017

ASCMO
Open Access

Reconstruction of spatio-temporal temperature from sparse historical records using robust probabilistic principal component regression

John Tipton[1], **Mevin Hooten**[2,3,1], **and Simon Goring**[4]

[1]Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA
[2]U.S. Geological Survey, Colorado Cooperative Fish and Wildlife Research Unit, Fort Collins, CO 80523, USA
[3]Department of Fish, Wildlife, and Conservation Biology, Colorado State University, Fort Collins, CO 80523, USA
[4]Department of Geography, University of Wisconsin, Madison, WI 53706, USA

Correspondence to: John Tipton (jtipton25@gmail.com)

Received: 30 June 2016 – Revised: 21 October 2016 – Accepted: 21 October 2016 – Published: 27 January 2017

Abstract. Scientific records of temperature and precipitation have been kept for several hundred years, but for many areas, only a shorter record exists. To understand climate change, there is a need for rigorous statistical reconstructions of the paleoclimate using proxy data. Paleoclimate proxy data are often sparse, noisy, indirect measurements of the climate process of interest, making each proxy uniquely challenging to model statistically. We reconstruct spatially explicit temperature surfaces from sparse and noisy measurements recorded at historical United States military forts and other observer stations from 1820 to 1894. One common method for reconstructing the paleoclimate from proxy data is principal component regression (PCR). With PCR, one learns a statistical relationship between the paleoclimate proxy data and a set of climate observations that are used as patterns for potential reconstruction scenarios. We explore PCR in a Bayesian hierarchical framework, extending classical PCR in a variety of ways. First, we model the latent principal components probabilistically, accounting for measurement error in the observational data. Next, we extend our method to better accommodate outliers that occur in the proxy data. Finally, we explore alternatives to the truncation of lower-order principal components using different regularization techniques. One fundamental challenge in paleoclimate reconstruction efforts is the lack of out-of-sample data for predictive validation. Cross-validation is of potential value, but is computationally expensive and potentially sensitive to outliers in sparse data scenarios. To overcome the limitations that a lack of out-of-sample records presents, we test our methods using a simulation study, applying proper scoring rules including a computationally efficient approximation to leave-one-out cross-validation using the log score to validate model performance. The result of our analysis is a spatially explicit reconstruction of spatio-temporal temperature from a very sparse historical record.

1 Introduction

There is a need for accurate estimates of paleoclimate, especially temperature and precipitation, to better understand how climate has changed in the past. Scientific measurements of temperature and precipitation have been recorded for several hundred years, and in many locations for a much shorter time. Because of long-standing interest in weather, there are a vast number of anecdotal, nonscientific records of weather. However, many reconstructions of paleoclimate using compiled historical records are not amenable to direct statistical analysis because they consist of imprecise measurements of weather reported in letters, newspapers, books, and other documents (Bell and Ogilvie, 1978; Ogilvie, 1984; Kastellet et al., 1998; Brázdil et al., 2006). The large quantity of historical weather records, combined with appropriate statistical models, has the potential to facilitate the extension of scientific understanding of climate further back in time.

Thus, there is a need for a statistical framework that can model historical data compiled from a variety of disparate sources by leveraging climate data from the recent past.

Historical observer weather data are often unreliable, sparse both temporally and spatially, and noisy because these data were recorded before widespread adoption of scientific measurement standards. As a result, historical observer weather data have not been widely used for rigorous statistical reconstructions of climate because these challenges make it difficult to create generic statistical approaches for analysis. Historical observer climate data can occur at hourly, daily, or monthly timescales, and the current-era analog data used to train statistical models can also vary in temporal resolution. Therefore, there is often a change of temporal support between the historical observer and current-era analog data that must be accounted for (Gotway and Young, 2002).

Another complication is that the true target one wishes to predict (the historical, unobserved climate) is never available to evaluate model predictive performance. Moreover, the historical observer data are often of unknown or of varying reliability and are typically sparse, sometimes involving only a few locations per year. The consequences of such data characteristics for evaluating model performance are underexplored; hence, we explore methods to validate historical observer-era model predictions under these sparse data scenarios.

We used spatially and temporally sparse historical observer measurements of temperature recorded at United States (US) military forts and other historical observer stations to reconstruct spatially explicit maps of mean midday July temperature by leveraging modern spatially explicit current-era analog data to impute missing spatial structure. We perform the reconstruction within a model framework that accounts for uncertainty in current-era data products and uncertainty in parameter estimation, and properly evaluates predictive skill. We test eight model specifications using a simulation study, generate predictions for mid-day July temperature at approximately 20 000 locations for each year in 1820–1894 with associated uncertainties, and evaluate model performance using a computationally efficient approximation to leave-one-out cross-validation.

2 Data

We used two datasets we refer to as the *historical observer dataset* and the *current-era analog* dataset. The historical observer dataset consists of temperature records from 1820 to 1894 at US forts in the Upper Midwestern US as well as nonmilitary observer stations. These data were compiled as part of the Climate Database Modernization Program (Andsager et al., 2004; CDMP 19th Century Forts and Voluntary Observers Database Build Project: http://www.isws.illinois.edu/atmos/clirecord.asp; CDMP, 2016). At the observer stations, measurements were recorded with time and date; however, the timing of measurements varied among and within individual observer stations and was often temporally imprecise (“daily min”, “daily max”, “mid-day”, etc.).

Protocols varied across the observer stations through space and time, leading to many irregularities in the historical observer data. Temperature measurements were obtained by a variety of methods: some records report daily minimum and maximum temperatures, others report hourly measurements, and sometimes there are days or weeks with missing measurements. In addition, the number and locations of the observer stations change through time, containing between 1 and 234 locations per year; this variation is due to historical events, including the Civil War and the westward expansion of the US in the late 19th century. Most years have only a few observations and, in general, the number of observer locations per year increases through time. Therefore, the model must align the temporal and spatial scales of the two data sources to reconstruct continuous temperature fields across the Upper Midwest. An example of 4 years of historical data is shown in Fig. 1a.

Because the historical observer data are spatially sparse, traditional spatial statistical methods, such as Kriging, are not applicable, as these methods require larger sample sizes to produce reasonable predictive surfaces. Thus, we used the current-era analog data to provide spatial structure for the reconstruction. For the current-era analog data, we used the Parameter-elevation Relationships on Independent Slopes Model (PRISM) monthly mean mid-day temperature surfaces created by interpolation of the US Historical Climate Network (USHCN) data over the period 1895–2010 (PRISM Climate Group, Oregon State University, 2016). The PRISM data include 115 years of mean mid-day July temperatures resolved to an 800 m × 800 m grid, resulting in almost 20 000 spatial locations of interest in the study region (Fig. 1b). Unlike the historical observer data, the PRISM data are compiled from the USHCN and consist of commonly used model interpolated temperature records. Other data products are available, including high-quality data from satellite measurements; however, we used PRISM for the current-era analog data due to the longer temporal coverage that provides the model with more examples of the spatial structure of midday July temperature. Because PRISM is a data product and not raw data, we account for potential measurement errors in the current-era analog data using our modeling framework.

2.1 Temporal change of support

To enable statistical learning about climate in the historical observer period, we aligned the two data sources to common spatial and temporal scales. We assigned each historical observer station to the closest grid cell in the current-era analog data, thus accounting for any potential spatial misalignment. Because the grid we aligned to is very fine scale (800 m × 800 m grid cells) and temperature surfaces are generally smooth over this spatial resolution, we assume any er-

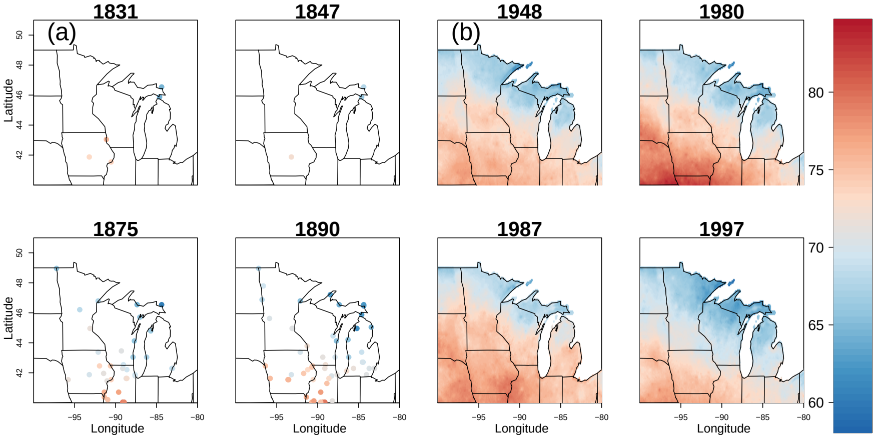

Figure 1. Four years of the historical observer temperature data **(a)** and the current-era analog temperature data **(b)**.

rors induced by the spatial alignment are negligible relative to other sources of noise in the data and ignore potential effects of spatial misalignment. Aligning the data sources in time was more complicated because the historical observer station data are highly irregular, whereas the current-era analog data are monthly mean mid-day temperatures. We modeled the historical observer period mean mid-day July temperature using cyclic cubic splines that are highly flexible, able to accommodate the irregular nature of the historical data, and constrained to reconstruct diurnal patterns (Wood, 2006). We focused on the month of July because the annual temperature curve peaks in July and thus there is little/no seasonal change in temperature that needs to be accounted for when computing a monthly average. The methodology could be applied to other months, but the calibration in Eq. (1) would need to account for seasonal trend.

We define our models using the following notation. Scalars are denoted by lowercase letters, vectors are bold lowercase letters, and matrices are bold uppercase letters. Fixed values, like data, are generally represented by Latin letters and parameters are written in Greek letters. Using this notation, the linear mixed model for estimating daily historical observer mean mid-day July temperature is

$$\widetilde{y}_{itj}(s) = l_i\beta + \boldsymbol{b}(s)'\boldsymbol{\alpha} + \eta_{it} + \eta_i + \eta_t + \varepsilon_{itj}(s), \tag{1}$$

where $\widetilde{y}_{itj}(s)$ is the raw historical observer temperature observation at location i, year t, day j, and hour s. The covariate l_i is the latitude at location i and gives rise to a spatially varying intercept for temperature parameterized by the coefficient β. The vector $\boldsymbol{b}(s)$ is a cyclic cubic spline basis expansion of order 4 over the 24 h daily cycle with coefficients $\boldsymbol{\alpha}$ that account for the diurnal pattern in temperature. The random effects η_i, η_t, and η_{it} adjust the model fit with varying intercepts for location i, year t, and the interaction between location and year. The model is completed by the inclusion of independent, uncorrelated Gaussian error $\varepsilon_{itj}(s)$, giving rise to interpolated daily temperature curves for July at each observer station location i and year t. From the daily temperature curves, we estimated mean mid-day July temperature by first predicting $y_{it}(\widetilde{s}) = l_i\hat{\beta} + \mathbf{B}(\widetilde{s})'\hat{\boldsymbol{\alpha}} + \hat{\eta}_{it} + \hat{\eta}_i + \hat{\eta}_t$ at 1 min intervals ($\widetilde{s} = \{0, 0 + \frac{1}{60}, \ldots, 23 + \frac{59}{60}\}$) for each fort location and year. We estimated the mean mid-day temperature using the same formula as the current-era analog data, $y_{it} = \left(\min_{\widetilde{s}} y_{it}(\widetilde{s}) + \max_{\widetilde{s}} y_{it}(\widetilde{s})\right)/2$, aligning the sparse, irregular historical observer data to the monthly timescale of the current-era analog data.

To facilitate parameter estimation in the presence of sparse data, the calibration model borrows strength among days, sites, and years within the historical observer data for the month of July, reducing the influence of measurement error and improving prediction of the mid-day diurnal temperature curve. By borrowing strength, the calibration model produced a mean mid-day estimate that has less variability than the raw historical observer data. We fit the calibration model to the historical data using `R` package `mgcv` (Wood, 2011) and refer to the pre-processed mid-day estimates y_{it} as the historical observer data in what follows. We justify the loss of information induced by using the calibration model predictions instead of the raw historical observer data because the linear mixed model explained approximately 70 % of the variability in the data ($R^2 = 0.69$) and provided a mechanism for changing temporal support by integrating uncertainty over the within-month mid-day temperature.

2.2 Modeling outline

After aligning the two data sources to a common temporal scale, we constructed a modeling framework to perform our reconstruction. One method commonly used for the reconstruction of paleoclimate is principal component regression (PCR), often called empirical orthogonal function (EOF) regression in the paleoclimate literature (Preisendorfer, 1988). The use of PCR for the statistical reconstruction of climate has a long tradition, dating back to Lorenz (1956). In PCR reconstructions, the climate proxy observations are regressed on a set of patterns created from direct observations of the climate process. After learning about the regression parameters, the model is used to predict climate at the unobserved locations.

To build our spatio-temporal predictive model, we used traditional principal component regression (PCR) as well as probabilistic principal component regression (pPCR) that assumes the empirical principal components are a noisy measure of the true, latent principal components (Tipping and Bishop, 1999). We explore the temporal PCR and pPCR models in a Bayesian hierarchical framework using regularization methods to select important principal components for each year's reconstruction. Within this framework, we assign hierarchical pooling priors to improve parameter estimation for years with few observations by borrowing strength from years with many observations (Gelman and Hill, 2006). We also develop robust, Student's t specifications of PCR and pPCR models that accommodate potentially outlying measurements of mid-day July temperature in the historical observer data that may have arisen from the non-standardized data collection.

We introduce traditional PCR in Sect. 3.1 within a temporal framework that allows for flexibility among years while borrowing strength among years to improve estimation in years with few observations and define the probabilistic extension (pPCR) of PCR that accounts for measurement error in Sect. 3.2. In Sect. 3.3, we introduce the robust specification of our PCR and pPCR models that better accommodate outlying observations, and in Sect. 3.4, we show how to improve computation by integrating out the latent principal components in the pPCR model. We describe three scoring rules to validate model performance in Sect. 4, and describe a simulation study in Sect. 5 where we evaluate predictive performance in a synthetic data scenario. In Sect. 6, we apply our models to reconstruct historical mean mid-day July temperature in the Upper Midwestern US, choosing the model that performs best based on scoring rules.

3 Model statement

3.1 Principal component regression

A common statistical approach for reconstruction of the historical climate using current-era analog data is to regress the partially observed historical observer data onto the current-era analog observations. For a given reconstruction year, define the regression model

$$y_{it} = \mu_t + \boldsymbol{x}_i' \boldsymbol{\alpha}_t + \epsilon_{it}, \tag{2}$$

where y_{it} is a historical observer period observation (preprocessed using calibration model Eq. 1) of mean mid-day July temperature at location i for year t. The vector $\boldsymbol{y}_t$ consists of the n_t historical observer period observations of the temperature field for year t, where we observe only n_t out of the n locations, with the number and locations of observations changing through time (Figs. 1a and 7b). The columns of the $n \times d$ matrix $\mathbf{X}$ contain d replicates of the current-era analog temperature surfaces at the n locations in the domain of interest, forming a basis set of patterns for the regression, where $\boldsymbol{x}_i'$ represents the ith row of $\mathbf{X}$. A greater number of replicates of current-era analog temperature surfaces d gives a larger set of potential spatial patterns that can be used to learn about spatial patterns in the historical observer data. The d-dimensional vector of regression coefficients $\boldsymbol{\alpha}_t$ link the historical observer data $\boldsymbol{y}_t$ with the set of climate patterns $\mathbf{X}$ in the current-era analog data for each year t, allowing for climate fields that are linear combinations of observed current-era analogs, up to uncorrelated model error. Thus, we can model temperatures that are warmer or cooler than the current-era analog period, but patterns that are not linear combinations of the current-era analogs are not accommodated in the model, necessitating use of a sufficiently long temporal record of current-era analogs. The uncorrelated model error ϵ_{it} is assumed to be independent and identically distributed Gaussian with variance τ_t^2, prior $\tau_t \sim \log N(\mu_\tau, \sigma_\tau^2)$, and vague hyperpriors $\mu_\tau \sim \mathrm{N}(0,1)$ and $\sigma_\tau \sim U(0,1)$. Because the likelihood is unaffected if μ_t is integrated out, we assume that the data $\boldsymbol{y}_t$ are centered and assume $\mu_t = 0$ (i.e., anomalies).

In Eq. (2), the columns in $\mathbf{X}$ are highly multicollinear. Multicollinearity inflates the coefficient estimate variance and, in cases of severe multicollinearity, the least squares solution is nearly singular, causing algorithm instability and unreliable estimation. One could use this model to estimate the dynamics influencing a given year's temperature surface by interpreting the estimated regression coefficients, but because we are interested in prediction of the dependent variable $\boldsymbol{y}_t$ and less interested in interpretation of the regression coefficients, we manipulate the form of $\mathbf{X}$ to improve statistical learning. We begin by computing the singular value decomposition (SVD) of $\mathbf{X} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}'$, where the columns of $\mathbf{U}$ are the left singular vectors of $\mathbf{X}$, the diagonal matrix $\boldsymbol{\Lambda}$ has the singular values in descending order on the diagonal, and the columns of $\mathbf{V}$ are the right singular vectors. The PCR model using the SVD is

$$\begin{aligned} y_{it} &= \boldsymbol{u}_i' \boldsymbol{\Lambda} \mathbf{V}' \boldsymbol{\alpha}_t + \epsilon_{it} \\ &= \boldsymbol{u}_i' \boldsymbol{\Lambda}^{\frac{1}{2}} \boldsymbol{\beta}_t + \epsilon_{it} \\ &= \boldsymbol{z}_i' \boldsymbol{\beta}_t + \epsilon_{it}, \end{aligned} \tag{3}$$

where $\boldsymbol{u}_i'$ is the ith row of $\mathbf{U}$. If the regression coefficient is given the prior $\boldsymbol{\alpha}_t \sim \mathrm{N}(\mathbf{0}, \sigma_{\alpha t}^2 \mathbf{I})$, then $\boldsymbol{\beta}_t = \boldsymbol{\Lambda}^{\frac{1}{2}} \mathbf{V}' \boldsymbol{\alpha}_t \sim N(\mathbf{0}, \sigma_{\beta t}^2 \boldsymbol{\Lambda})$ where $\sigma_{\alpha t}^2 = \sigma_{\beta t}^2$. In this model, the columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{X}'\mathbf{X}$, the diagonal elements of $\boldsymbol{\Lambda}$ are the eigenvalues of $\mathbf{X}'\mathbf{X}$, and $\boldsymbol{z}_i' = \boldsymbol{u}_i' \boldsymbol{\Lambda}^{\frac{1}{2}}$ is the scaled principal component at location i. In applications of PCR, one often performs dimension reduction by retaining only the first p eigenvectors of $\mathbf{U}$ in the $n \times p$ matrix $\mathbf{U}_p$ and the first p eigenvalues in the $p \times p$ matrix $\mathbf{\Lambda}_p$. After truncation, the truncated PCR design matrix is $\mathbf{Z}_p = \mathbf{U}_p \mathbf{\Lambda}_p^{\frac{1}{2}}$. Typically, p is chosen by cross-validation or by choosing the smallest p so that the proportion of variability explained in the model is a large value. Because truncation removes the highest-frequency eigenvectors, the truncation implies a prior that shrinks the regression coefficients and provides an implicit regularization on the model (Hastie et al., 2005). Although truncation of lower-order principal components disregards small-scale variability and therefore can only reduce the theoretical minimum prediction error, the truncated model is often more computationally stable than Eq. (2) and can improve prediction in practice.

Preexisting research suggests that truncation of the trailing principal components is not always appropriate because the higher-frequency components are often important predictors (Hadi and Ling, 1998; Jolliffe, 1982). In our paleoclimate reconstruction method, inclusion of lower-order principal components is important, especially if there are climate signals that are slowly varying or show up occasionally (i.e., every decade or century). If these uncommon processes appear in the lesser eigenvectors of the current-era analog data (which is likely because such processes are not the primary contributors to the annual-scale variability in climate), these signals would be discarded by truncation as high-frequency noise. Allowing the important principal components in the regression to vary with time, the model is capable of detecting the changes in temperature. For example, the irregular but periodic cycles of the Pacific Decadal Oscillation and the Atlantic Multidecadal Oscillation likely do not explain a large portion of the variability in the temperature records. Therefore, these and other similar climate signals could be removed through the truncation of the principal components. Ideally, one chooses the truncation p to be as large as computationally possible and then performs a variable selection or regularization method to select important principal components. Wang (2012) approached the problem of choosing the important principal components through the Bayesian model selection technique known as stochastic search variable selection (SSVS; George and McCulloch, 1993; see Hooten and Hobbs, 2015, for a review). The SSVS variable selection assumes the hierarchical prior on the jth regression coefficient at time t

$$\beta_{tj} \sim \begin{cases} N(0, \sigma^2_{\beta_t} \lambda_{p_j}) & \text{if } \xi_{tj} = 1, \\ N(0, \sigma^2_{\beta_t} \kappa_j^{-1} \lambda_{p_j}) & \text{if } \xi_{tj} = 0, \end{cases} \tag{4}$$

where λ_{p_j} is the jth diagonal element of $\mathbf{\Lambda}_p$. The SSVS prior for the pPCR model is similar, but does not include the λ_{p_j} terms. The variables ξ_{tj} are indicators of the importance of the jth latent principal component in the regression for year t and have independent Bernoulli(0.5) priors. The regression coefficient variance $\sigma^2_{\beta_t}$ could be assigned a prior if desired, but the shrinkage value $\kappa_j > 1$ must be fixed. A large κ_j produces a mixture distribution of a broad, relatively uninformative prior with large variance $\sigma^2_{\beta_t}$ (the "slab") and a highly informative prior at a small neighborhood around zero (the "spike") that provides shrinkage by truncating less important principal components using probabilistic learning, thus reducing the chance of omitting important principal components while avoiding the computationally expensive task of exploring all 2^p possible model configurations. We set $\kappa_j = 1000$ and pool across years by assuming the hierarchical model with prior $\sigma_{\beta_t} \sim \log N(\mu_{\sigma_\beta}, \sigma^2_{\sigma_\beta})$ and vague hyperpriors $\mu_{\sigma_\beta} \sim N(0, 1)$ and $\sigma_{\sigma_\beta} \sim \mathrm{U}(0, 1)$.

An alternative to variable selection methods like SSVS is penalized regression (Hastie et al., 2005). Common forms of penalized regression include ridge regression (Tikhonov or L_2 shrinkage; Hoerl and Kennard, 1970), where one minimizes

$$\sum_{i=1}^{n} (y_{it} - z_i' \boldsymbol{\beta}_t)^2 + \gamma_t \sum_{j=1}^{p} \beta_{tj}^2$$

with respect to $\boldsymbol{\beta}_t$, and the least angle subset selection operator (LASSO or L_1 shrinkage; Tibshirani, 1996), which minimizes

$$\sum_{i=1}^{n} (y_{it} - z_i' \boldsymbol{\beta}_t)^2 + \gamma_t \sum_{j=1}^{p} |\beta_{tj}|$$

with respect to $\boldsymbol{\beta}_t$ given the penalty term γ_t. The L_2 penalty shrinks the coefficients non-linearly toward zero and the L_1 penalty shrinks large coefficients linearly, but in a way that the coefficients can equal zero exactly. When viewed from this perspective, the LASSO can be viewed as a compromise between regularization and variable selection methods because, as the coefficients in the LASSO model approach zero, there is nonzero probability that the LASSO will shrink the covariate estimates to zero, thereby removing that variable from the model (Efron et al., 2004). We apply both SSVS and LASSO shrinkage methods to explore the empirical consequences of the choice of regularizer. One drawback to regularization methods is the need to estimate the penalty parameter γ_t. Often, the optimal γ_t is determined by cross-validation using predictive skill. In the Bayesian framework, the shrinkage can be estimated by cross-validation or by assigning a prior distribution and performing a fully Bayesian inference (Park and Casella, 2008; Hooten and Hobbs, 2015).

The L_2 penalty implies the prior $\boldsymbol{\beta}_t \sim N(\mathbf{0}, \gamma_t \mathbf{I})$ and the L_1 LASSO penalty assigns a Laplace (double exponential) prior $\boldsymbol{\beta}_t \sim \prod_{j=1}^{d} \frac{\gamma_t}{2\sqrt{\tau_t^2}} \exp\{-\frac{\gamma_t |\beta_{tj}|}{\sqrt{\tau_t^2}}\}$. The LASSO penalty can also be specified using the more computationally efficient hierarchical-scale mixture of Gaussian distributions with exponential mixing distribution by assigning the hierarchical prior

$$\boldsymbol{\beta}_t \sim N(\mathbf{0}, \tau_t^2 \mathbf{D}_{\gamma_t}),$$

$$\gamma_{tj} \sim \text{Exp}\left(\frac{\lambda^2}{2}\right),$$

where $\mathbf{D}_{\gamma_t} = \text{diag}(\gamma_{t1}, \ldots, \gamma_{tp})$ (Park and Casella, 2008). We hierarchically pool the error standard deviation by assigning the hyperprior $\tau_t \sim \log N(\mu_\tau, \sigma_\tau^2)$ with vague hyperparameters $\mu_\tau \sim N(0, 1)$ and $\sigma_\tau \sim U(0, 1)$, where learning across years is achieved by updating μ_τ and σ_τ. To perform a fully Bayesian regularization that properly accounts for parameter uncertainty, we assign the hyperpriors $\lambda_t^2 \sim \text{Gamma}(\alpha_\lambda, \beta_\lambda)$ to allow for differential regularization through time. We assign the hierarchical pooling prior by modeling the parameters α_λ and β_λ, reparameterizing the Gamma distribution using its mean $\mu_\lambda = \frac{\alpha_\lambda}{\beta_\lambda}$ and variance $\sigma_\lambda^2 = \frac{\alpha_\lambda}{\beta_\lambda^2}$ and assigning the vague hyperpriors $\mu_\lambda \sim \text{logN}(0, 1)$ and $\sigma_\lambda \sim \text{U}(0, 1)$.

3.2 Probabilistic principal component regression

PCR assumes the data $\mathbf{X}$, and therefore the principal components derived from $\mathbf{X}$ are observed without measurement error. The current-era analog data are model interpolated, and, therefore, the principal components have unaccounted for measurement error that violates the assumptions of traditional PCR. Hence, the eigenvectors in $\mathbf{U}$ can be thought of as estimates of the true eigenvectors under an appropriate probabilistic model. As a remedy, probabilistic principal component models assume the data matrix $\mathbf{X}$ is a noisy measurement of the true process (Tipping and Bishop, 1999). Letting $\boldsymbol{x}_i'$ be the ith row of $\mathbf{X}$, the model for the noisy observations is

$$\boldsymbol{x}_i = \boldsymbol{m} + \mathbf{K}\boldsymbol{z}_i + \boldsymbol{\eta}_i, \tag{5}$$

where $\boldsymbol{m}$ is the d-vector of means, $\mathbf{K}$ is a $d \times p$ rotation matrix, $\boldsymbol{z}_i$ is a p-vector that represents the latent eigenvectors of the process of interest, and $\boldsymbol{\eta}_i$ is zero mean, independent Gaussian error with variance σ^2. Note that $\boldsymbol{m}$ can be integrated out of Eq. (5) without changing the likelihood; thus, we assume that the data $\mathbf{X}$ have centered rows and set $\boldsymbol{m} = \mathbf{0}$ (i.e., anomalies). Because principal component vectors are orthonormal, we complete the principal component model specification by assigning independent priors $\boldsymbol{z}_i \sim N(\mathbf{0}, \mathbf{I})$, for $i = 1, \ldots, p$. A more general model is the factor analysis model, where the error term $\boldsymbol{\eta}$ has a generic diagonal covariance matrix $\boldsymbol{\Sigma}$ (Tipping and Bishop, 1999). Thus, the probabilistic principal component model can be viewed as a special case of factor analysis where the error term $\boldsymbol{\eta}$ is constrained to be diagonal with variance σ^2.

Tipping and Bishop (1999) showed the maximum likelihood estimate (MLE) of the rotation matrix $\mathbf{K}$ with p components under the pPCR model is

$$\hat{\mathbf{K}} = \mathbf{U}_p\left(\boldsymbol{\Lambda}_p - \bar{\lambda}\mathbf{I}_p\right)^{\frac{1}{2}}\mathbf{R},$$

where $\mathbf{U}_p$ is a $d \times p$ matrix with the first p columns containing the leading eigenvectors, $\boldsymbol{\Lambda}_p$ is a $p \times p$ diagonal matrix with the associated eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p$ of $\mathbf{X}'\mathbf{X}$ on the diagonal, the matrix $\mathbf{I}_p$ is the $p \times p$ identity matrix, $\bar{\lambda} = \frac{\sum_{j=p+1}^{d} \lambda_j}{d-p}$ is the average variance contribution for the truncated eigenvectors, and $\mathbf{R}$ is an arbitrary orthogonal rotation matrix (which we set to be $\mathbf{I}_p$). We set $\mathbf{K}$ at the MLE and rewrite Eq. (5) as

$$\boldsymbol{x}_i' = \hat{\mathbf{K}}\boldsymbol{z}_i + \boldsymbol{\eta}_i. \tag{6}$$

After accounting for the measurement uncertainty in our predictor matrix $\mathbf{X}$ by estimating the unknowns $\mathbf{Z}$ and $\boldsymbol{\eta}$ in Eq. (6), we link the historical observer data and current-era analog data by regressing $\boldsymbol{y}_t$ onto the latent eigenvectors $\mathbf{Z}$

$$y_{it} = \boldsymbol{z}_i'\boldsymbol{\beta}_t + \epsilon_{it} \tag{7}$$

and estimate the unknown regression coefficients $\boldsymbol{\beta}_t$ in Eq. (7).

3.3 Robust regression

The historical observer data were collected using nonstandard methods; thus, there is likely more variability in the data than can be explained by assuming a Gaussian error distribution. We propose extending Eq. (7) to a model that is robust to outliers. The robust pPCR data model using the Student's t distribution

$$y_{it} \sim \text{t}(\boldsymbol{z}_i'\boldsymbol{\beta}_t, \tau_t^2, \nu_t)$$

is a model that better accommodates outliers in the data. The parameter ν_t is the degrees of freedom of the Student's t distribution. A common choice of prior for the degrees of freedom ν_t is to model the inverse degrees of freedom with a $U(0, 0.5)$ distribution. We generalize this prior to pool across years, assigning the inverse degrees of freedom the prior $\frac{1}{\nu_t} \sim \text{Beta}(\alpha_\nu, \beta_\nu, 0, 0.5)$ where the four-parameter $\text{Beta}(\alpha, \beta, L, U)$ prior is a $\text{Beta}(\alpha, \beta)$ prior scaled to the interval $[L, U]$. To hierarchically pool the prior model, we reparameterize $\alpha_\nu = \mu_\nu \eta_\nu$ and $\beta_\nu = (1 - \mu_\nu)\eta_\nu$ for $\mu_\nu \in [0, 1]$ and $\eta_\nu \in [0, \infty)$. We complete the model statement by assigning the hyperpriors $\mu_\nu \sim \text{Beta}(5, 5)$ and $\eta_\nu \sim \text{Gamma}(10, 0.1)$. Although these priors appear informative, when reparameterized, the prior specification is similar to the commonly used vague $\text{Gamma}(2, 0.1)$ prior on the degrees of freedom ν_t (Juárez and Steel, 2010). To regularize the robust data model, we modify the LASSO prior for the regression coefficients using the variance of the Student's t distribution, resulting in the prior

$$\boldsymbol{\beta}_t \sim N\left(\mathbf{0}, \tau_t^2 \frac{\nu_t}{\nu_t - 2}\mathbf{D}_{\gamma_t}\right),$$

where the parameters have the same priors as introduced previously.

3.4 Posterior distribution

The latent principal components z_i are high dimensional (approximately 20 000-dimensional for $i = 1, \ldots, p$). We aim to avoid the computational burden of sampling this parameter. Therefore, we wish to integrate out the latent principal components

$$\int [y_{it} | z_i, \boldsymbol{\beta}_t, \tau_t^2, \nu_t][\boldsymbol{x}_i | z_i, \sigma^2][z_i] \, dz_i, \tag{8}$$

but this integral is not analytically tractable. We could attempt to numerically integrate out z_i, but at great computational cost. Instead, we write our Student's t data model as a scale mixture where $y_{it} \sim \mathrm{N}(z_i' \boldsymbol{\beta}_t, \upsilon_{it}^2)$, $\upsilon_{it}^2 \sim$ inv-$\chi^2(\nu_t, \tau_t^2)$, and $[y_{it} | z_i, \boldsymbol{\beta}_t, \tau_t^2, \nu_t] = \int [y_{it} | z_i, \boldsymbol{\beta}_t, \upsilon_{it}^2][\upsilon_{it}^2 | \tau_t^2, \nu_t] \, d\upsilon_{it}^2$. Then, we write the integral Eq. (8) as

$$\begin{aligned}
&\int [y_{it} | z_i, \boldsymbol{\beta}_t, \tau_t^2, \nu_t][\boldsymbol{x}_i | z_i, \sigma^2][z_i] \, dz_i \\
&\quad = \int \left(\int [y_{it} | z_i, \boldsymbol{\beta}_t, \upsilon_{it}^2][\upsilon_{it}^2 | \tau_t^2, \nu_t] \, d\upsilon_{it}^2 \right) [\boldsymbol{x}_i | z_i, \sigma^2][z_i] \, dz_i \\
&\quad = \int \left(\int [y_{it} | z_i, \boldsymbol{\beta}_t, \upsilon_{it}^2][\boldsymbol{x}_i | z_i, \sigma^2][z_i] \, dz_i \right) [\upsilon_{it}^2 | \tau_t^2, \nu_t] \, d\upsilon_{it}^2 \\
&\quad = \int [y_{it} | \mu_{y_{it}}, \sigma_{y_{it}}^2][\upsilon_{it}^2 | \tau_t^2, \nu_t] \, d\upsilon_{it}^2, \qquad (9)
\end{aligned}$$

where the integral in Eq. (9) is evaluated by Markov chain Monte Carlo (MCMC), first sampling $\upsilon_{it}^2 \sim$ inv-$\chi^2(\nu_t, \tau_t^2)$, then evaluating the density $[y_{it} | \mu_{y_{it}}, \sigma_{y_{it}}^2]$. These modifications result in the integrated data model (see the Supplement for details)

$$y_{it} \sim \mathrm{N}(\mu_{y_{it}}, \sigma_{y_{it}}^2), \tag{10}$$

where $\mu_{y_{it}} = \frac{\boldsymbol{x}_i' \hat{\mathbf{K}} \mathbf{M}_p^{-1} \boldsymbol{\beta}_t}{\sigma^2}$ and $\sigma_{y_{it}}^2 = \upsilon_{it}^2 + \boldsymbol{\beta}_t' \mathbf{M}_p^{-1} \boldsymbol{\beta}_t$ with $\mathbf{M}_p = \frac{\boldsymbol{\Lambda}_p}{\sigma^2} + \mathbf{I}_p$, a diagonal matrix that can be inverted efficiently. Integration results in significant computational savings because we avoid sampling p vectors of length n (approximately 20 000 each). The cost of not sampling the latent principal components $\mathbf{Z}$ is loss of conjugacy for the regression coefficients $\boldsymbol{\beta}_t$ in the MCMC algorithm. The posterior distribution (for the robust pPCA model with LASSO regularization) from which we sample using MCMC is

$$\begin{aligned}
&\prod_{t=1}^{T} \prod_{i \in \mathcal{H}_t} \left[\boldsymbol{\beta}_t, \upsilon_{it}^2, \tau_t, \mu_\tau, \sigma_\tau, \nu_t, \mu_\nu, \eta_\nu, \sigma, \boldsymbol{\gamma}_t, \lambda_t^2, \mu_\lambda, \sigma_\lambda \middle| y_{it}, \mathbf{X} \right] \propto \\
&\quad \prod_{t=1}^{T} \left(\prod_{i \in \mathcal{H}_t} \left[y_{it} \middle| \boldsymbol{\beta}_t, \sigma, \upsilon_{it}^2 \right] \left[\upsilon_{it} \middle| \nu_t, \tau_t \right] \right) \left[\boldsymbol{\beta}_t \middle| \boldsymbol{\gamma}_t, \tau_t, \nu_t \right] \left[\tau_t \middle| \mu_\tau, \sigma_\tau \right] \\
&\qquad \times [\mu_\tau][\sigma_\tau][\sigma] \left[\boldsymbol{\gamma}_t \middle| \lambda_t^2 \right] \left[\lambda_t^2 \middle| \mu_\lambda, \eta_\lambda \right] [\mu_\lambda][\eta_\lambda],
\end{aligned}$$

where $\mathcal{H}_t$ is the set of locations where there are observations for year t. We fit our models using `JAGS` (Plummer, 2003) within the `R` computing environment (R Core Team, 2016).

For each of the eight candidate models, we fit four parallel chains with random initial conditions, running 20 000 iterations per chain and discarding the first 10 000 iterations as burn-in. Fitting all eight models and the associated post-processing took approximately 18 h on a 2014 dual-core 2.6 GHz MacBook Pro with 8 GB RAM. We thinned our chains every 10 iterations to reduce post-processing time, resulting in a total of 4000 samples and evaluated model convergence using the $\hat{R}$ statistic (Gelman and Rubin, 1992). We chose vague hyperpriors throughout; the ability to estimate temperature surfaces from these priors implies the results are not highly sensitive to the prior values. A choice of stronger hyperprior values could improve inference, but very strong hyperpriors could also dominate the influence of the data in the posterior estimates. Preliminary analyses not shown in this paper indicated little sensitivity to reasonable prior choices.

4 Scoring rules

To evaluate model performance, we apply scoring rules to the estimated posterior predictive distributions. A highly desirable property of a scoring rule is propriety (Gneiting, 2011). A scoring rule is proper if the expected score of the optimal prediction is less than or equal to the expected score of any other prediction (Bernardo and Smith, 2009). Hence, a proper scoring rule, on average, chooses the best prediction from a set of candidate predictions (Gneiting et al., 2007). Often, paleoclimate reconstructions evaluate predictive performance by holding out some of the training set data for use in cross-validation, using skill scores like the coefficient of efficiency (CE) and relative efficiency (RE) (Cook et al., 1994; Rutherford et al., 2005; Tingley and Huybers, 2010a, b). Although these scoring rules are common in the paleoclimate reconstruction community, Gneiting and Raftery (2007) suggest that scoring rules like CE and RE are improper in general. Because CE and RE are improper, it is possible that the optimal prediction can, on average, have a worse score than a sub-optimal prediction, leading to incorrect inference. Therefore, we focus on three proper scoring rules: mean square prediction error (MSPE), the continuous ranked probability score (CRPS), and a computationally efficient approximation to leave-one-out cross-validation (LOO) using the log score. In general, MSPE is not proper, but because our data models are Gaussian and Student's t, MSPE is proper for predictions of the posterior mean in this case.

The use of MSPE as a scoring rule implies an L^2 loss function on the posterior distribution; therefore, our predictions are the posterior predictive means

$$\mathrm{E}(\widetilde{\boldsymbol{y}}_t | \boldsymbol{y}_t) = \int \widetilde{\boldsymbol{y}}_t [\widetilde{\boldsymbol{y}}_t | \boldsymbol{y}_t] \, d\widetilde{\boldsymbol{y}}_t,$$

where $[\widetilde{\boldsymbol{y}}_t | \boldsymbol{y}_t] = \int [\widetilde{\boldsymbol{y}}_t | \boldsymbol{\theta}_t][\boldsymbol{\theta}_t | \boldsymbol{y}_t] \, d\boldsymbol{\theta}_t$ is the posterior predictive distribution for model parameters $\boldsymbol{\theta}_t$. Given out-of-sample observations $\boldsymbol{y}_{\mathrm{oos},t}$, MSPE is

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{n-n_t}\sum_{i\notin\mathcal{H}_t}\left(\mathrm{E}(\widetilde{y}_{it}|\boldsymbol{y}_t)-y_{\mathrm{oos},it}\right)^2,$$

where $n-n_t$ is the number of out-of-sample locations for year t and $\mathcal{H}_t$ is the set of observed locations in the historical observer data. Because MSPE uses the posterior predictive mean (a point prediction) instead of the full posterior distribution, MSPE ignores much of the information in the posterior distribution gained by performing Bayesian inference. Therefore, MSPE is not an ideal scoring rule for a probabilistic prediction, such as a posterior predictive distribution, even when MSPE is proper. For example, consider two models that give rise to posterior predictive distributions with the same posterior predictive mean but different posterior predictive variances. In this case, it is obvious that the predictive distribution that has better predictive coverage should be preferred, but MSPE would score the two models identically, demonstrating how MSPE loses information by collapsing the posterior distribution into a point estimate.

An alternative to MSPE is the CRPS scoring rule. CRPS is proper, utilizes the full posterior predictive distribution, and allows for a direct comparison of point predictions and probabilistic predictions (Gneiting and Raftery, 2007). CRPS resolves the issue presented in the previously described scenario by including the width of the predictive distribution in the evaluation of the score. Several recent papers presenting climate reconstructions have made use of the CRPS for these reasons (Barboza et al., 2014; Werner and Tingley, 2015; Tipton et al., 2016). Given a prediction with the cumulative distribution function, F_{it}, at location i and time t, and out-of-sample observations $\boldsymbol{y}_{\mathrm{oos},t}$, the CRPS is defined as

$$\mathrm{CRPS}(\{F_{it}\}_{t=1}^{T},\boldsymbol{y}_{\mathrm{oos},t})=\\ -\sum_{t=1}^{T}\sum_{i\notin\mathcal{H}_t}\int_{-\infty}^{\infty}\left(F_{it}(y)-I_{\{y\geq y_{\mathrm{oos},it}\}}\right)^2\mathrm{d}y. \tag{11}$$

Gneiting and Raftery (2007) show that Eq. (11) can be written alternatively as

$$\mathrm{CRPS}(\{F_{it}\}_{t=1}^{T},\boldsymbol{y}_{\mathrm{oos}})= \\ \sum_{t=1}^{T}\frac{1}{n-n_t}\sum_{i\notin\mathcal{H}_t}\left(E_{F_{it}}\left|y_{it}-y_{\mathrm{oos},it}\right|-\frac{1}{2}E_{F_{it}}\left|y_{it}-y_{it}^{*}\right|\right), \tag{12}$$

where y_{it} and y_{it}^{*} are independent copies of a random variable with distribution function F_{it} and the expectation E is with respect to the probability density induced by F_{it}. The first expectation in Eq. (12) measures calibration (the absolute error of the prediction relative to the out-of-sample value) and the second expectation rewards predictions that are precise (i.e., narrow prediction intervals).

We can estimate the CRPS after obtaining posterior samples $\widetilde{\boldsymbol{y}}_t^{(k)}$ from the posterior predictive distribution $\left[\widetilde{\boldsymbol{y}}_t^{(k)}|\boldsymbol{y}_t\right]$ at each post burn-in iteration k. Then, Eq. (12) is approximated by

$$\widehat{\mathrm{CRPS}}(\{\hat{F}_{it}\}_{t=1}^{T},\boldsymbol{y}_{\mathrm{oos}})= \\ \sum_{t=1}^{T}\left(\frac{1}{n-n_t}\sum_{i\notin\mathcal{H}_t}\left(\frac{1}{K}\sum_{k=1}^{K}\left|\widetilde{y}_{it}^{(k)}-y_{\mathrm{oos},it}\right|-\right.\right. \\ \left.\left.\frac{1}{2K^2}\sum_{k=1}^{K}\sum_{\ell=1}^{K}\left|\widetilde{y}_{it}^{(k)}-\widetilde{y}_{it}^{(\ell)}\right|\right)\right). \tag{13}$$

A major disadvantage of both MSPE and CRPS is the need for out-of-sample validation data. For our simulation study, MSPE and CRPS are straightforward to calculate because we simulated the out-of-sample validation data; in practical paleoclimate reconstructions, there are no out-of-sample data. Therefore, MSPE and CRPS must be approximated using cross-validation methods, although these methods are computationally costly and time consuming to implement.

An alternative is to use the approximate leave-one-out cross-validation method (LOO; Vehtari et al., 2016b). LOO uses a proper scoring rule, the log score, to evaluate predictive skill (Geisser and Eddy, 1979; Gneiting and Raftery, 2007; Hooten and Hobbs, 2015). We estimate the leave-one-out log pointwise predictive density

$$\mathrm{lpd}_{\mathrm{loo}}=\sum_{t=1}^{T}\sum_{i\in\mathcal{H}_t}\log[y_{it}|\boldsymbol{y}_{(i)t}]= \\ \sum_{t=1}^{T}\sum_{i\in\mathcal{H}_t}\log\int[y_{it}|\boldsymbol{\theta}_t][\boldsymbol{\theta}_t|\boldsymbol{y}_{(i)t}]\,d\boldsymbol{\theta}_t, \tag{14}$$

where $\boldsymbol{y}_{(i)t}$ are the data $\boldsymbol{y}_t$ at time t without the ith location. One can calculate Eq. (14) directly by cross-validation at a high computational cost, or one can approximate Eq. (14) using importance sampling from post burn-in posterior samples using the full data as described in Vehtari et al. (2016b). Importance ratios with high variance can cause the estimate in Eq. (14) to be highly unstable and unreliable, and are therefore of practical concern. To test for the presence of large variance of the importance ratios, Koopman et al. (2009) proposed fitting the generalized Pareto distribution to the upper tail of importance ratios and examining the empirical estimates of the tail shape parameter ξ. If the estimated tail parameter $\hat{\xi}_{it}$ is less than $1/2$, the variance of the importance ratios is finite and the importance ratios approximating the log posterior score holding out y_{it} can be used directly to approximate LOO. If the estimated tail parameter is $1/2<\hat{\xi}_{it}<1$, the variance of the importance ratios is infinite but the mean of the importance ratios exists. Hence, Vehtari and Gelman (2015) propose using smoothed importance ratios. If the estimated tail parameter $\hat{\xi}_{it}>1$, this suggests that the mean and variance of the importance ratios do not exist but that the variance of the smoothed importance ratios is finite, but large, and the use of LOO is sensitive to the held-out observation. Using the smoothed importance weights $w_{it}^{(k)}$, we obtain the Pareto-smoothed importance sampling approximation

$$\widehat{\text{elpd}}_{\text{PSIS}} = \sum_{t=1}^{T} \sum_{i \in \mathcal{H}_t} \log\left(\frac{\sum_{k=1}^{K} w_{it}^{(k)} [y_{it} | \boldsymbol{\theta}_t^{(k)}]}{\sum_{k=1}^{K} w_{it}^{(k)}} \right). \qquad (15)$$

We use the deviance scale and set $\widehat{\text{LOO}} = -2\widehat{\text{elpd}}_{\text{PSIS}}$ to make LOO a negatively oriented score (the best model is the one with the lowest score), implementing our score using `R` package loo (Vehtari et al., 2016a).

5 Simulation

With paleoclimate data, it is difficult to verify the predictive ability of models using cross-validation. With only a handful of observations in the historical observer data available for each year, cross-validation techniques could be highly biased due to the effects of unusual observations in small sample sizes. This is important because we expect noisy and potentially outlying observations in the historical observer data due to the data collection procedures. Additionally, the high dimensionality of the field we aim to reconstruct and the use of computationally intensive MCMC estimation make cross-validation costly. Instead, we conducted a simulation study to explore the different models for the historical observer station data and evaluate model performance using the scoring rules above. Although we do not simulate from the model that is used for estimation, the simulated data represent a reasonable approximation to mid-day July temperature, providing an environment for model testing and exploration of empirical performance.

We simulate mid-day July temperature in one spatial dimension (we extend to two dimensions using the real data), allowing for faster computation and easier graphical exploration of the spatio-temporal process. We simulate $T = 50$ realizations of a latent surface from the model

$$\boldsymbol{s}_t = \mathbf{W}\boldsymbol{\beta}_t + \boldsymbol{\eta}_t,$$

where the matrix $\mathbf{W}$ represents fixed influences on climate, such as latitude, elevation, and other covariates that explain much of the temperature surface as well as time-varying components that represent slowly varying global-scale climate processes. To construct patterns that might be seen in climate observations, we simulate temporally varying regression coefficients at different periodicities to represent global-scale climate processes like the Pacific Decadal Oscillation or the Atlantic Multidecadal Oscillation. We do not claim our simulation behaves like any climatological process, only that this example facilitates exploration of complicated patterns potentially seen in climatological data.

We include a spatially correlated random effect $\boldsymbol{\eta}_t \sim \text{N}\left(\mathbf{0}, \sigma^2_{\eta_t}\mathbf{R}(\phi)\right)$ that smooths the patterns, generating realizations of a one-dimensional climate field (Fig. 2b). A common choice for the form of $\mathbf{R}(\phi)$ is the Matérn class of correlation functions. For our simulation, we use the exponential correlation function, a member of the Matérn family. In the exponential correlation function, the i, jth element $R_{ij}(\phi) = \exp\left(-d_{ij}\phi\right)$, where d_{ij} represents the Euclidean distance between the ith and jth spatial locations and ϕ is the spatial range parameter.

To create observations that match the temporal irregularities and spatial clustering behavior in the historical observer data, we sample the one-dimensional spatial field using weighted probabilities that generate clustered observations in space, storing the simulated temperature observations at the n_t locations in the vector $\boldsymbol{y}_t$. Using this sampling design, we generate noisy realizations for the simulated historical observer data for simulated years $t = 1, \ldots, 25$ (Fig. 2a) using

$$y_{it} = s_{it} + \widetilde{\epsilon}_{it}, \qquad (16)$$

where $\widetilde{\epsilon}_{it}$ is independent Student's t error with variance $\widetilde{\sigma}^2 = 1.5$ and degrees of freedom $\nu = 10$ that represent uncertainty in historical temperature measurements. To generate the set of simulated current-era analog patterns $\mathbf{X}$ that will be used in our regression model, we define for the simulated current-era analog years $t = 26, \ldots, 50$

$$\boldsymbol{x}_t = \boldsymbol{s}_t + \ddot{\boldsymbol{\epsilon}}_t, \qquad (17)$$

where, by adding uncorrelated, independent Gaussian noise $\ddot{\boldsymbol{\epsilon}}_t$ with variance $\ddot{\sigma}^2 = 0.75$, we account for the measurement error of the temperature process during the current-era analog period. The measurements used in the model for the current-era analog period are from PRISM model interpolated data and have measurement error, where this measurement error should be less than that of the historical observer data; hence, $\widetilde{\sigma}^2 > \ddot{\sigma}^2$. We then combine the simulated values into the noisy pattern matrix $\mathbf{X} \equiv (\boldsymbol{x}_{26}, \ldots, \boldsymbol{x}_{50})$ that is used in our model framework. The latent temperature patterns $\mathbf{S} \equiv (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{25})$ are the unobserved target for our reconstruction. Note that, in the real data, $\mathbf{S}$ are unavailable; therefore, our scoring rules can use only the noisy observations $\boldsymbol{y}_t$, limiting the ability to improve reconstruction skill. Figure 2a shows the simulated historical observer period noisy temperature realizations $\boldsymbol{y}_t$ and Fig. 2b shows the simulated historical observer period true, latent temperature field that is the target of our reconstruction, with the x axis representing spatial location. The noisy principal components derived from the simulated current-era analog data are plotted in Fig. 2c with the simulated latent principal components in Fig. 2d. Comparing Fig. 2c and d shows that using the noisy principal components is analogous to the errors-in-covariates framework, where noisy observations of the current-era analog covariates (in our case the principal components) can lead to bias in the regression coefficients, inflated residual variance, and a reduction in prediction skill (Carroll et al., 2006; Fuller, 2009; Buonaccorsi, 2010).

We compare the performance of each model specification using MSPE, CRPS, and LOO scoring rules in our simulation

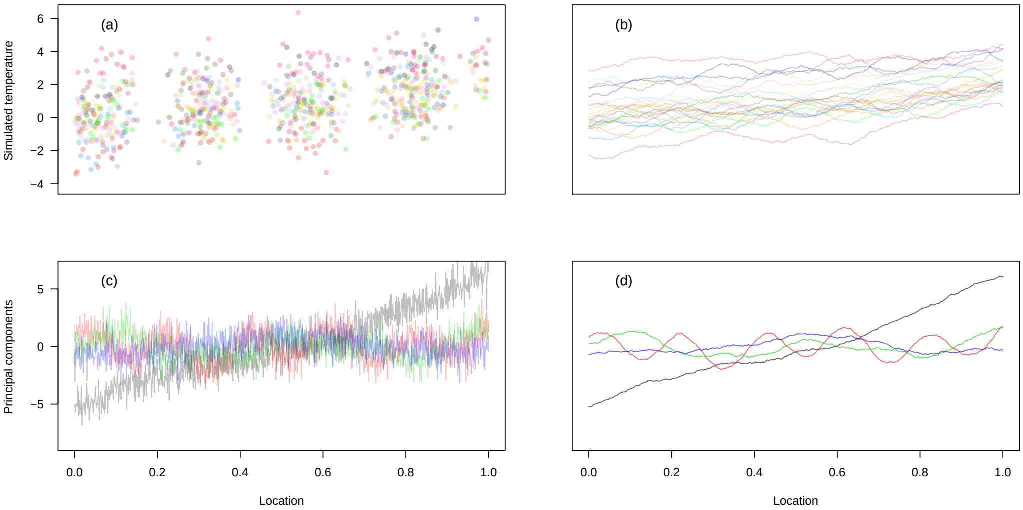

Figure 2. Simulation study showing observed noisy historical observer data **(a)**, the simulated true latent climate process we aim to predict **(b)**, the first four noisy principal components (PCs) estimated from the historical observer data **(c)**, and the first four simulated true latent PCs **(d)** that show the effect of measurement error when compared to **(c)**. Each year of the historical observer period in the simulation study in **(a)** and **(b)** is assigned a different color and the historical observer period observations **(a)** are clustered in space, changing in sample size through time, and noisier than the latent temperature in **(b)**. Both the noisy **(c)** and latent **(d)** PCs increase in variability as the number of the component increases, but the latent PCs are smoother. The first PC is in black, the second PC is in blue, the third PC is in green, and the fourth PC is in red.

Table 1. Simulation experiment scores. Smaller values indicate better model performance.

	MSPE		CRPS		LOO	
Model	Gaussian	Robust	Gaussian	Robust	Gaussian	Robust
SSVS PCR	0.379	0.380	199	194	2917	2901
SSVS pPCR	0.403	0.404	206	205	2920	2919
LASSO PCR	0.508	0.456	217	205	2970	2934
LASSO pPCR	0.398	0.397	206	205	2918	2918

where the best model is the one with the smallest score. We fit the PCR and pPCR models using SSVS and LASSO regularization with both the Gaussian and robust Student's t data models, comparing eight models with the results displayed in Table 1. Across scores, the models perform similarly, with no consistently best model, although the robust models generally have lower scores than the Gaussian models. The LOO Pareto tail parameter estimates for the robust models show less evidence of misspecification than the traditional Gaussian data models (figure not shown) and the pPCR models have consistently smaller tail parameter estimates than the PCR models, suggesting that the least misspecified models are robust pPCR.

Predicted versus simulated temperatures for the robust PCR and robust pPCR models using LASSO regularization are shown in Fig. 3 with years represented using different colors. Because the predictions for each year cluster around the 45° line, this shows the models reconstruct the annual differences accurately, which are the main features of the simulated data and of primary interest in understanding climate change. However, the models fail to reconstruct much of the within-year spatial variability (the humps and valleys within a year in Fig. 2b), which is unsurprising given the small sample sizes. Despite having similar predictive scores, there are visible differences in predictions between the two models. The robust PCR predictions predict some of the spatial structure of the mean (the point clouds are generally centered on the 45° line), but tend to have unreasonably small predictive standard deviations (not shown). The robust pPCR predictions estimate a spatially averaged annual mean in years with small sample sizes, but predict little spatial structure. Instead, robust pPCR produces predictive standard deviation estimates that account for uncertainty in years with very little data.

6 Observer station data reconstruction

After exploring the model framework using a simulation study, we applied our models to the historical observer data.

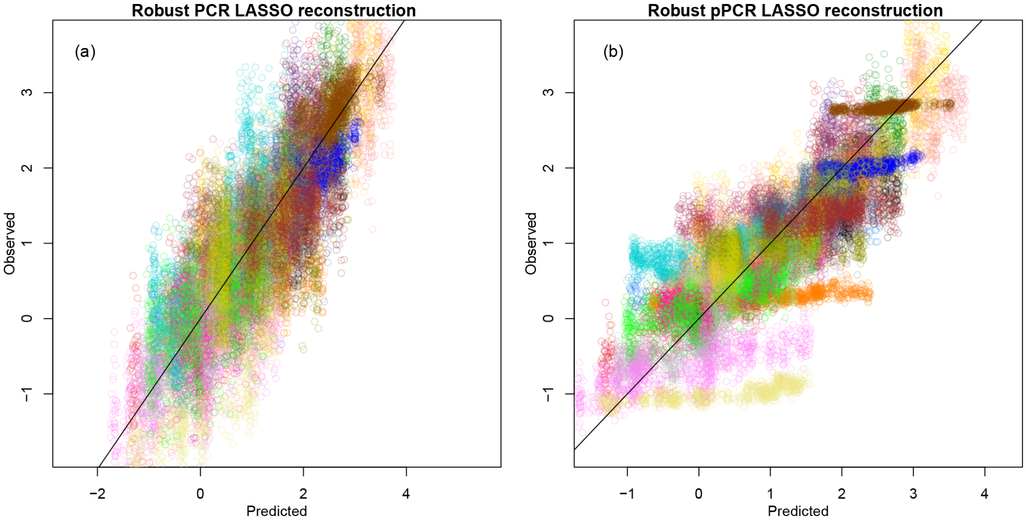

Figure 3. Simulation truth plotted against predicted temperature for the robust PCR LASSO model on the left and the robust pPCR LASSO model on the right. Predictions for each simulated year are given different colors and the clustering of colors represents annual-scale changes in the mean temperature surface. Results are shown for the LASSO model and the SSVS model performs similarly.

Table 2. Historical observer reconstruction scores. Smaller values indicate better model performance.

	Full data		Outlier removed	
Model	Gaussian	Robust	Gaussian	Robust
SSVS PCR	7499	7183	7936	7168
SSVS pPCR	7609	7378	8033	7367
LASSO PCR	7445	7065	7935	7067
LASSO pPCR	7616	7370	8053	7352

We fit the eight models to the data and present the results from LOO in Table 2. Examination of the LOO Pareto tail parameter plots in Fig. 4a and b identified an outlier occurring at data point 1452, corresponding to an unrealistic mean midday July temperature measurement of 42 F (7.8 °C). After removing the outlier, we fit the models again. Interestingly, the Gaussian model fits without the outlier showed less predictive skill (larger LOO values in Table 2) and more model misspecification (a larger tail parameter estimate in Fig. 4e and f). The decreased performance when removing the outlier is explained by the influence of the outlier on the pooled variance estimate; removing the outlier shrinks the pooled variance estimate and the Gaussian data model is less able to accommodate slightly outlying points with the smaller variance. For the robust models, quality of model fit was slightly improved when fit with the outlier removed (see Table 2), but there is little change in model misspecification as defined by large Pareto tail parameter estimates (Fig. 4c, d, g, and h).

With the outlier removed, the best predictive models are robust PCR and robust pPCR due to having the smallest LOO scores (see Table 2). Based purely on LOO, it appears the best overall predictive model is robust PCR, although the increasing sample sizes through time weight the LOO score to predictions of the most recent years. All models still show evidence of misspecification because some Pareto tail parameter estimates are greater than 0.5 (Fig. 4e, f, g, and h), but the removal of the outlier improved model fit in general. Figure 4g and h suggest that the robust pPCR model might be preferable to the robust PCR model farther back in time because, for that time period, a much greater proportion of Pareto tail parameter estimates are less than 0.5 in the robust pPCR model than the robust PCR model. Hence, there is some evidence that the choice of model is dependent on the desired inference. If inference over the entire period is desired, both robust PCR and robust pPCR predict with skill. If inference is desired on the years furthest back in time, Fig. 4 suggests that robust pPCR predictions are preferable.

To visualize our results, we plot reconstructions of 4 years of the historical temperature surfaces using the robust PCR model (Fig. 5) and the robust pPCR model (Fig. 6). Visual comparison of the reconstructions illustrates differences in the two models. The robust PCR model assumes the climate patterns in the observational data are without error; the stronger influence of these patterns is seen in the posterior predictive mean surface (Fig. 5a), particularly for year 1847. In comparison, the robust pPCR model shows less influence of these climate patterns in the posterior predictive mean (Fig. 6a), particularly for year 1847. Because the robust pPCR model includes the assumption that some of the pattern is noise, the reconstructions have less spatial structure in years with little data. Differences in posterior predictive standard deviations are also evident (Figs. 5b and 6b), where the robust PCR model shows reduction of uncertainty in the spatial locations near observations and higher uncer-

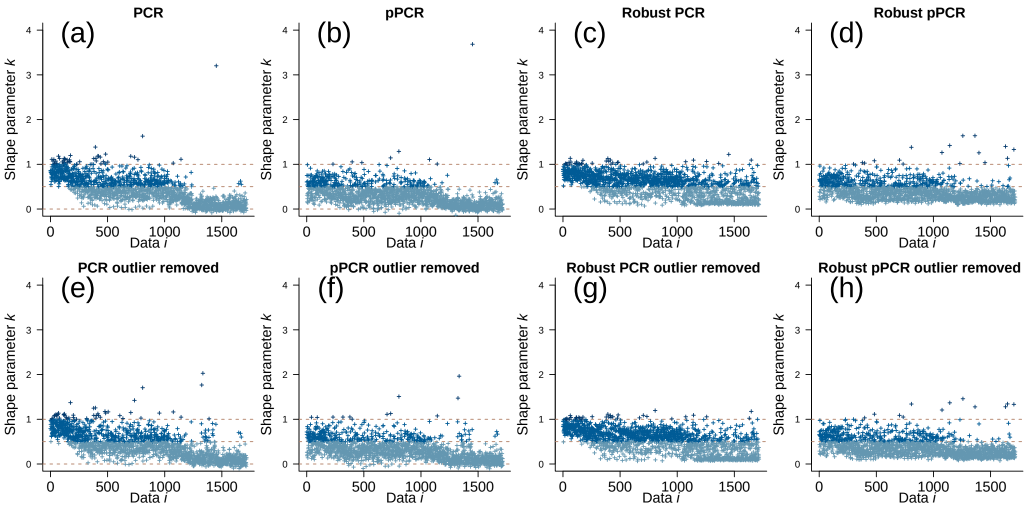

Figure 4. Historical observer station data LOO Pareto shape estimates with **(a, b, c, d)** and without **(e, f, g, h)** outlying observation. Values less than 0.5 show good model performance and values over 1.0 show poor model performance. Note the presence of the outlier in the upper right of **(a)** and **(b)** for observation 1452. Also evident is less model misspecification for the pPCR models **(b, d, f, h)** for observations furthest back in time.

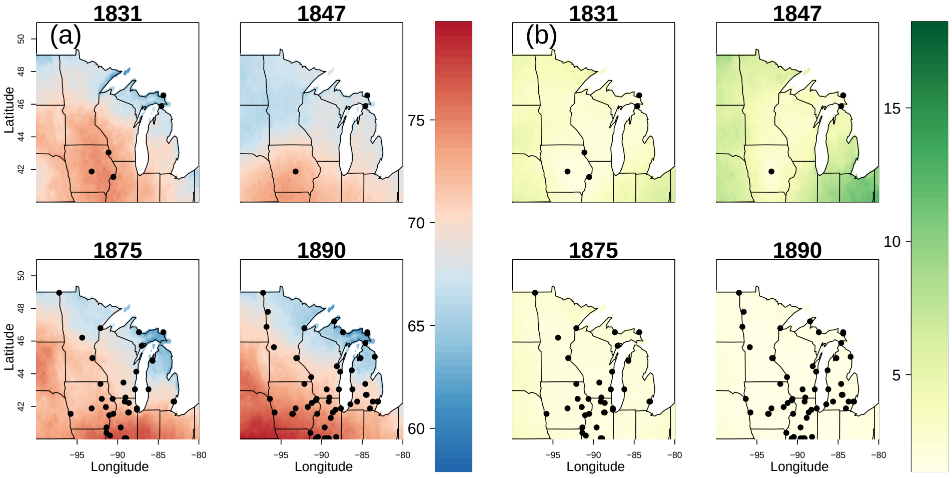

Figure 5. Reconstruction of mean mid-day July temperature using the robust PCR model for 4 years. Figures show posterior predictive mean **(a)** and standard deviation **(b)**.

tainty away from the observations, while the robust pPCR model has posterior predictive standard deviations that are more spatially diffuse and perhaps more realistic given the lack of spatial predictive ability seen in the simulation study. The reconstructed temperature surfaces for both the robust PCR and robust pPCR models for every year are included in the Supplement.

By using the spatial structure in the current-era analog data, we generated temperature predictions at unobserved locations with corresponding uncertainties. We chose four locations, Champaign, Illinois, Detroit, Michigan, Madison, Wisconsin, and Minneapolis, Minnesota, and show the time series of temperature predictions in Fig. 7a. From these time series, we see smaller standard deviations for the years with more historical observer period observations (sample sizes are shown in Fig. 7b), with greater uncertainty the further we go back in time. We can also see a general trend for the standard deviations of the robust PCR model to be smaller than the robust pPCR model (the red intervals are more often inside the blue intervals). The robust PCR model is also more likely to have structure in the mean that may not be warranted when the sample size is low (i.e., the spike at Champaign and Detroit that is not present in Madison or Minneapolis in 1847 and 1848 with sample sizes of 3). The two models show evidence of a bias/variance trade-off, with the robust PCR tending toward spatially structured predictions with smaller variance and the robust pPCR model providing less spatially structured predictions with larger variance.

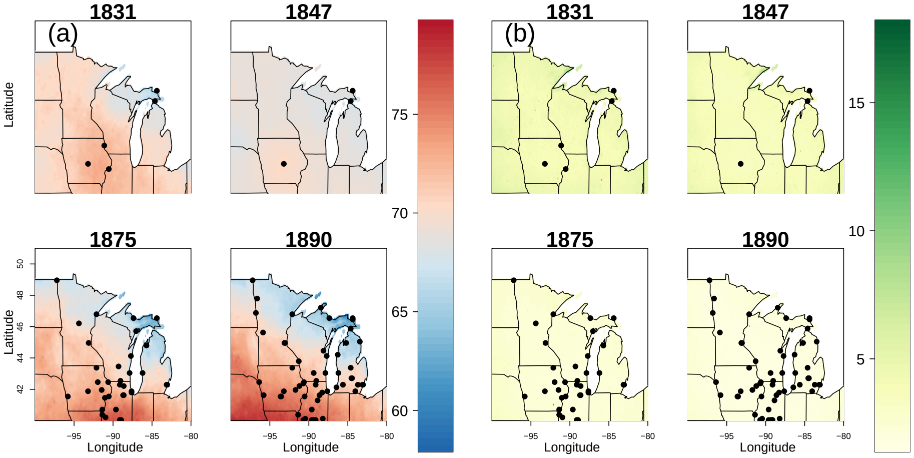

Figure 6. Reconstruction of the mean mid-day July temperature using the robust probabilistic PCR model for 4 years. Figures show posterior predictive mean **(a)** and standard deviation **(b)**.

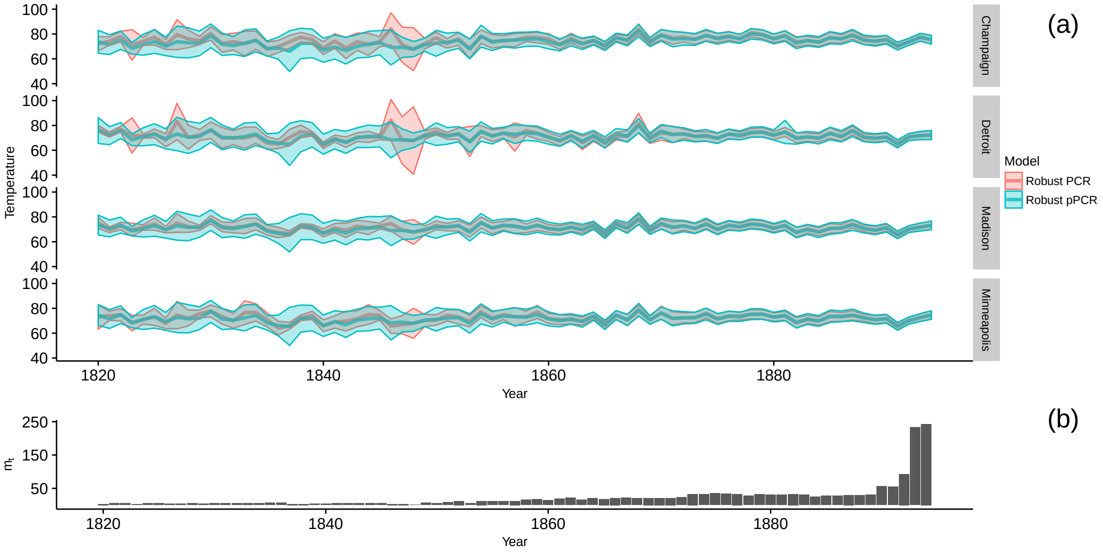

Figure 7. Posterior predictions of a time series of mid-day July temperature with associated 95 % credible intervals at Champaign, Illinois, Detroit, Michigan, Madison, Wisconsin, and Minneapolis, Minnesota **(a)**. The temporal change in the number of observations is shown in **(b)**. Note that the uncertainties in the point reconstructions are smallest in years with larger samples and largest in years with few samples. In general, the robust pPCR credible intervals are larger than the robust PCR credible intervals. The differences in reconstruction in years 1847 and 1848 at Champaign, Illinois, and Detroit, Michigan, demonstrate the bias/variance trade-offs in predictions between the two models. The robust PCR predictions tend toward lower variance and the robust pPCR predictions tend to lower bias with respect to the spatially averaged annual mean mid-day July temperature.

7 Conclusions

There are many challenges inherent in modeling paleoclimate data. Due to the lack of direct measurements of climate, paleoclimate reconstructions must rely on sparse, noisy proxies of climate. The nuances of paleoclimate data often require specialized modeling techniques and careful investigation into modeling assumptions and performance. In addition, care is needed to properly validate paleoclimate reconstruction skill. In summary, we extended principal component regression methods, applied regularization techniques to choose important principal components, developed robust models to account for the presence of outliers, and explored the use of a probabilistic principal component model to account for measurement uncertainty in the spatially rich current-era analog data. By rigorously evaluating the predictive skill of our models, we were able to explore our extensions of PCR for climate reconstruction, laying the groundwork for future developments with more complex climate data than PRISM temperature surfaces. The models presented in this paper would be good candidates for modeling climate variables that are strongly non-stationary and non-Gaussian (e.g., wind speed or precipitation), but these extensions are the subject of ongoing research.

Within our modeling framework, we presented a simulation study for evaluating paleoclimate reconstructions using proper scoring rules. By using proper scoring rules and exploring model performance in a simulation framework, we have stronger support for the quality of the reconstruction. We presented three statistical scoring rules and explored their strengths and weaknesses. MSPE is a commonly used and easy to understand scoring rule, but is not proper in general and only uses a point prediction, ignoring the probabilistic inference that is gained by using Bayesian techniques. The CRPS is proper and allows for direct comparison of point predictions and probabilistic predictions, but requires out-of-sample validation data or computationally expensive cross-validation. The use of MSPE and CRPS scoring rules allowed for exploration of the empirical properties of the computationally efficient LOO approximation to leave-one-out cross-validation. Our use of LOO to score the historical observer period model predictions not only enabled us to perform model selection, but also aided in diagnosing an outlying observation and refining model fit.

The methods presented in this paper could be applied to other historical datasets at different locations around the world, further extending the spatially explicit empirical record of climate further back in time while rigorously accounting for uncertainties. The methods we presented could also be extended to model temperature and precipitation for each month of the year by including a seasonal component in the calibration model and by modeling dynamics at appropriate timescales. There are many datasets that could be used within this framework as the current-era analog, including modern satellite data. The different options of current-era analog datasets present a trade-off between the number of records available as analogs and the quality of the data. If there are occasionally rare climate processes that occur, it seems that a longer record of climate analogs would be preferred. If the climate processes are relatively stable in time but vary highly in space, a shorter and more precise modern dataset that is not model interpolated might be preferred. In addition, use of highly precise current-era analog data could reduce or eliminate the need to account for measurement error in the current-era analog data.

Ultimately, our temperature reconstructions extend the climatological record in the Upper Midwestern US further into the past. These temperature reconstructions, with their associated uncertainties, can be used to gain better understanding of the influences of climate on the biological and ecological processes observed in the region. By backcasting mean midday July temperature with our models, we gain the potential to better understand how climate has changed, and this knowledge could be used to improve future climate reconstructions. Many of the techniques and methods we used – modeling principal components with a probabilistic model, hierarchical pooling to borrow strength among years with sparse and dense data, model selection and regularization, and proper model evaluation – can be adapted and used in future climate reconstruction problems.

8 Data availability

The historical observer data are available from http://www.isws.illinois.edu/atmos/clirecord.asp and the current-era analog data are available from http://www.prism.oregonstate.edu/. Compiled data and code can be accessed on gitHub at https://github.com/jtipton25/observer (doi:10.5281/zenodo.242996).

The Supplement related to this article is available online at doi:10.5194/ascmo-3-1-2017-supplement.

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

Acknowledgements. This paper was greatly improved based on the comments from the two anonymous reviewers and the associate editor. The authors also thank Ben Bird, Kristin Broms, Brian Brost, Franny Buderman, Trevor Hefley, and Henry Scharf for their conversations and input on this work. This research is based upon work carried out by the PalEON Project (paleonproject.org) with support from the National Science Foundation MacroSystems Biology program under grant no. DEB-1241856. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. Code and data found in this paper can be accessed at https://github.com/jtipton25/observer (doi:10.5281/zenodo.242996).

Edited by: A. M. Schmidt
Reviewed by: two anonymous referees

References

Andsager, K., Ross, T., Kruk, M.C., and Spinar, M. L.: Climate database modernization program: pre-20th century task – key climate observations recorded since the founding of America, 1700s–1800s, in: Combined preprints: 84th AMS annual meeting : 20th Conference on Weather Analysis and Forecasting/16th Conference on Numerical Weather Prediction, Seattle Washington, Boston, MA, American Meteorological Society, 2004.

Barboza, L., Li, B., Tingley, M., and Viens, F.: Reconstructing past temperatures from natural proxies and estimated climate forcings using short-and long-memory models, Ann. Appl. Stat., 8, 1966–2001, 2014.

Bell, W. and Ogilvie, A.: Weather compilations as a source of data for the reconstruction of European climate during the medieval period, Climatic Change, 1, 331–348, 1978.

Bernardo, J. M. and Smith, A.: Bayesian Theory, vol. 405, John Wiley & Sons, 2009.

Brázdil, R., Kundzewicz, Z., and Benito, G.: Historical hydrology for studying flood risk in Europe, Hydrolog. Sci. J., 51, 739–764, 2006.

Buonaccorsi, J. P.: Measurement Error: Models, Methods, and Applications, CRC Press, 2010.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M.: Measurement Error in Nonlinear Models: A Modern Perspective, CRC press, 2006.

CDMP: 19th Century Forts and Voluntary Observers Database Build Project, available at: http://www.isws.illinois.edu/atmos/clirecord.asp, last access: 21 October 2016.

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

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R.: Least angle regression, Ann. Stat., 32, 407–499, 2004.

Fuller, W. A.: Measurement Error Models, vol. 305, John Wiley & Sons, 2009.

Geisser, S. and Eddy, W.: A predictive approach to model selection, J. Am. Stat. Assoc., 74, 153–160, 1979.

Gelman, A. and Hill, J.: Data Analysis Using Regression and Multilevel/Hierarchical Models, Cambridge University Press, 2006.

Gelman, A. and Rubin, D. B.: Inference from iterative simulation using multiple sequences, Stat. Sci., 7, 457–472, 1992.

George, E. I. and McCulloch, R. E.: Variable selection via Gibbs sampling, J. Am. Stat. Assoc., 88, 881–889, 1993.

Gneiting, T.: Making and evaluating point forecasts, J. Am. Stat. Assoc., 106, 746–762, 2011.

Gneiting, T. and Raftery, A.: Strictly proper scoring rules, prediction, and estimation, J. Am. Stat. Assoc., 102, 359–378, 2007.

Gneiting, T., Balabdaoui, F., and Raftery, A.: Probabilistic forecasts, calibration and sharpness, J. Roy. Stat. Soc. B, 69, 243–268, 2007.

Gotway, C. A. and Young, L.: Combining incompatible spatial data, J. Am. Stat. Assoc., 97, 632–648, 2002.

Hadi, A. S. and Ling, R.: Some cautionary notes on the use of principal components regression, Am. Stat., 52, 15–19, 1998.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J.: The elements of statistical learning: data mining, inference and prediction, Math. Intell., 27, 83–85, 2005.

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

Hooten, M. B. and Hobbs, N.: A guide to Bayesian model selection for ecologists, Ecol. Monogr., 85, 3–28, 2015.

Jolliffe, I. T.: A note on the use of principal components in regression, Appl. Statist., 31, 300–303, 1982.

Juárez, M. A. and Steel, M. F.: Model-based clustering of non-Gaussian panel data based on skew-t distributions, J. Bus. Econ. Stat., 28, 52–66, 2010.

Kastellet, E., Nesje, A., and Pedersen, E.: Reconstructing the palaeoclimate of Jæren, Southwestern Norway, for the period 1821–1850, from historical documentary records, Geogr. Ann. A, 80, 51–65, 1998.

Koopman, S. J., Shephard, N., and Creal, D.: Testing the assumptions behind importance sampling, Journal of Econometrics, 149, 2–11, 2009.

Lorenz, E. N.: Empirical orthogonal functions and statistical weather prediction, Scientific report no. 1: Statistical forecasting project, Massachusetts Institute of Technology, Department of Meteorology, 1956.

Ogilvie, A. E.: The past climate and sea-ice record from Iceland, Part 1: Data to AD 1780, Climatic Change, 6, 131–152, 1984.

Park, T. and Casella, G.: The Bayesian lasso, J. Am. Stat. Assoc., 103, 681–686, 2008.

Plummer, M.: JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling, in: Proceedings of the 3rd international workshop on distributed statistical computing, vol. 124, 125 pp., Technische Universität Wien, Wien, Austria, 2003.

Preisendorfer, R.: Principal Component Analysis in Meteorology and Oceanography, Developments in Atmospheric Science, 17, Elsevier, 1988.

PRISM Climate Group, Oregon State University: available at: http://prism.oregonstate.edu, last access: 21 October 2016.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2016.

Rutherford, S., Mann, M., Osborn, T., Briffa, K., Jones, P., Bradley, R., and Hughes, M.: Proxy-based Northern Hemisphere surface temperature reconstructions: Sensitivity to method, predictor network, target season, and target domain, J. Climate, 18, 2308–2329, 2005.

Tibshirani, R.: Regression shrinkage and selection via the lasso, J. Roy. Stat. Soc. B, 58, 267–288, 1996.

Tingley, M. P. and Huybers, P.: A Bayesian algorithm for reconstructing climate anomalies in space and time. Part I: Development and applications to paleoclimate reconstruction problems, J. Climate, 23, 2759–2781, 2010a.

Tingley, M. P. and Huybers, P.: A Bayesian algorithm for reconstructing climate anomalies in space and time. Part II: Comparison with the regularized expectation-maximization algorithm, J. Climate, 23, 2782–2800, 2010b.

Tipping, M. E. and Bishop, C.: Probabilistic principal component analysis, J. Roy. Stat. Soc. B, 61, 611–622, 1999.

Tipton, J., Hooten, M., Pederson, N., Tingley, M., and Bishop, D.: Reconstruction of late Holocene climate based on tree growth and mechanistic hierarchical models, Environmetrics, 27, 42–54, 2016.

Vehtari, A. and Gelman, A.: Pareto Smoothed Importance Sampling, arXiv preprint arXiv:1507.02646v2, 2015.

Vehtari, A., Gelman, A., and Gabry, J.: loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models, R package version 0.1.6, available at: https://github.com/jgabry/loo (last access: 21 October 2016), 2016a.

Vehtari, A., Gelman, A., and Gabry, J.: Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC, arXiv preprint arXiv:1507.04544, 2016b.

Wang, L.: Bayesian principal component regression with data-driven component selection, J. Appl. Stat., 39, 1177–1189, 2012.

Werner, J. P. and Tingley, M. P.: Technical Note: Probabilistically constraining proxy age–depth models within a Bayesian hierarchical reconstruction model, Clim. Past, 11, 533–545, doi:10.5194/cp-11-533-2015, 2015.

Wood, S.: Generalized Additive Models: An Introduction with R, CRC press, 2006.

Wood, S. N.: Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models, J. Roy. Stat. Soc. B, 73, 3–36, 2011.