Dissertation

Improved estimation and prediction for computationally expensive ecological and paleoclimate models

Submitted by

John Tipton

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2016

Doctoral Committee:

Advisor: Mevin Hooten
Co-Advisor: Jean Opsomer

Jennifer Hoeting
Cameron Aldridge

Abstract

Improved estimation and prediction for computationally expensive ecological and paleoclimate models

In this dissertation, we present statistical methods to evaluate estimation and prediction performance for applied ecological problems. We explore a variety of applied problems and, within this context, we investigate how each method performs. We evaluate empirical performance of a model-based estimator of mean percent canopy cover using a representative United States Forest Service Forest Inventory and Analysis dataset. For two paleoclimate reconstructions, we develop novel modeling methodologies and evaluate model performance using both resampling and simulation methods. In each application, we use proper scoring rules while leveraging parallel computing and computational techniques, that allow fitting of complex models in a finite amount of time.

Acknowledgements

Throughout this journey I have been blessed by the mentorship and guidance I have received. My graduate research career started when I was introduced to Gretchen Moisen and Paul Patterson of the United States Forest Service Rocky Mountain Research Station by my co-advisor Jean Opsomer. In these relationships I learned what it means to collaborate and solve applied problems.

I am eternally grateful to my advisor and mentor Mevin Hooten. Thanks to your support I have learned an incredible amount about not only statistics but how to think creatively and succeed in academic research. You introduced me into the PalEON research team and opened many opportunities for me going forward. Without your support and friendship, this work would not be possible.

In life I have been truly fortunate for all of the support and opportunities that have allowed me to complete this work. First, I am grateful to my parents for their love, patience, and work ethic. Without them, I would not be anywhere near the person I am today. I also want to thank my intelligent, loving, talented, and beautiful wife Crystal. You are my rock and beacon in this world and I am eternally grateful for you love. I'm lucky to have made so many friends during my time working on this dissertation, many of whom I forced into discussions of this work and for that I say thanks for maintaining our friendship. And finally, I'd like to say thanks to my first daughter who entered this world two days after my defense. Feeling the love you bring into this world is all the motivation a man could ever need moving forward.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

In statistics, the goal is to estimate or predict an unobserved value: a population mean, variance, effect size, unknown variable, etc. For any quantity of interest, there are a variety of potential methods available, thus it is of interest to evaluate which methods perform well. The true values of interest are generally unknown, presenting a fundamental challenge. When the truth is unknown, model comparison is difficult. Therefore, the evaluation and selection among competing methods is important. Model comparison becomes even more difficult when the models are computationally costly to fit, such as Bayesian hierarchical models, or rely on a number of assumptions, such as independence and normality of the data.

1.1. Model Evaluation - The Two Cultures

Statisticians commonly use two methods to evaluate model performance, broadly classified as resampling methods and simulation methods. Examples of resampling methods include bootstrapping, partitioning the data into test and validation datasets, and cross-validation. Resampling methods are meant to approximate model performance on future, unobserved data that are similar to the currently observed data and are, therefore, commonly used to prevent model overfitting (Hastie et al., 2009, Chapter 7.10). In simulation methods, one simulates data from processes believed to be close to the true data-generating process and then evaluates predictive performance of the model fit using these simulated data. Simulation studies help the researcher gain a better understanding of the scientific processes that could generate the observed data and can, therefore, improve model understanding. Simulation studies also allow for exploration of model performance across a wide range of

scenarios that may not be included in the observed data but could be seen in other similar data. In addition, simulation studies are often computationally cheap when compared to resampling methods.

Both resampling and simulation methods have their challenges and trade-offs. Resampling methods are computationally costly, have the potential to be influenced by outlying observations, can be unstable in sparse data scenarios, and, for cases like paleoclimate reconstructions, the predictions that one wants to evaluate (e.g., historical temperature and precipitation) have no data available for resampling. In the case of simulation studies, the performance of the model on simulated data is only valuable if the simulated data are close to the observed data, and for many problems it is not clear how to rigorously define or evaluate closeness. In fact, assumptions made during data simulation can greatly influence how different models perform. For example, if the true data are temporally correlated but the simulated data are temporally independent, the simulation experiment will be less valuable. When viewed from the perspective of choosing between model evaluation methods, the statistician must weigh the costs and benefits associated with each method.

1.2. Keeping Score

1.2.1. The Decision Theoretic Approach. To see the connection between the evaluation of survey sampling estimators and the scoring of model predictions, we take a decision theoretic approach. In what follows, we provide a review of the mathematical framework that makes explicit the link between evaluation of survey estimates and model predictions. We begin by defining the set of all possible events or outcomes of an experiment Ω, commonly called the sample space. A set of outcomes $E \in \Omega$ is called an event. The next component needed to formally define a probability space is a σ-algebra $\mathcal{E}$ that contains subsets of Ω

that are of interest (practically, these are the outcomes that one could actually observe). Finally, we need a probability measure $\mathbb{P} : \mathcal{E} \to [0, 1]$ that assigns probabilities to the different outcomes in $\mathcal{E}$. Within this framework, we represent the probability of observing the event $E \in \mathcal{E}$ as $P(E)$. Taken together, the triple $(\Omega, \mathcal{E}, \mathbb{P})$ formally define a probability space (Grimmett and Stirzaker, 2001). To simplify notation, we use $\mathcal{F}$ to represent the class of probability distributions associated with the probability measure $\mathbb{P}$ and associated σ-field $\mathcal{E}$.

A real-valued continuous random variable $Y = Y(\omega) : \Omega \to \mathbb{R}$ is a measurable $\mathcal{E}$ function that maps from the sample space Ω to the real numbers $\mathbb{R}$. Hence, we write probability statements $P(Y = 1)$ and $P(-1 \leq Y \leq 1)$ which formally mean $P(\{\omega \in \Omega : Y(\omega) = 1\})$ and $P(\{\omega \in \Omega : -1 \leq Y(\omega) \leq 1\})$, respectively. Then, assuming a probability density function $f_Y : \mathbb{R} \to \mathbb{R}$ exists (we assume the usual measurability conditions are met) and defining $\mathbf{y}$ as the observed value of the random variable Y, we can write our probability statement as $P(\{\omega \in \Omega : Y(\omega) \in E_F\}) = \int_{E_F} f_Y(\mathbf{y})\, d\mathbf{y}$ (Billingsley, 2008).

In the decision theoretic framework, one wants to take actions in the presence of uncertainty. In practice, one assumes there is a state in nature θ that can take one of a set of possible values denoted by Θ (commonly called the parameter space). Often, θ is an unknown quantity of interest (a parameter or prediction) that one wishes to make inference about. The inference is formally called an action and the set of possible actions A is called the action space, where we denote a particular action as a. If we assume that the random variable Y depends on the parameter θ (and hence, the underlying probability measure $\mathcal{P}$ depends on the unknown state of nature θ), we can write the probability of event E as $P_\theta(E) = \int_E f_Y(\mathbf{y}|\,\theta)\, dy$, for probability density function $f_Y(\cdot|\theta) \in \mathcal{F}$ that depends on the unknown θ, where the dependence is suppressed in the notation for compactness. For most

estimation and prediction problems, it is assumed that the action space A is equal to the parameter space Θ, and we make this assumption in what follows.

To evaluate an action under a state of nature, there must be a function that maps actions and states of nature to the real numbers. In other disciplines, these functions are called utility or objective functions, but statisticians call these functions loss functions. The loss function is a non-negative function that generally increases as the distance between a and θ increases to represent the fact that if a is close to θ, there is little loss incurred and if a is far from θ a large loss is incurred. Common forms of loss functions include squared-error (L_2) loss where $L(a,\theta) = (a-\theta)^2$ and absolute error (L_1) loss where $L(a,\theta) = |a-\theta|$.

The final element of the decision theoretic framework is the functional called the decision rule $m(\mathbf{y})$ that maps the sample space of $\mathbf{y}$ (which is a random function with probability distribution $F_Y \in \mathcal{F}$) to the action a. Thus, after a loss function is chosen, the decision theoretic problem is to choose $m(\mathbf{y})$ such that the risk (or expected loss) is minimized. In other words, one chooses the functional

$$\widehat{m}(\mathbf{y}) = \underset{m(\mathbf{y})}{\operatorname{argmin}} \mathbb{E}_{F_Y} L(m(\mathbf{y}), \theta), \tag{1}$$

that minimizes the risk. Because the decision is made under uncertainty about the true state of nature θ with the goal of minimizing risk, the correct action is the Bayes rule $\widehat{m}(\mathbf{y})$ in (1). Hence, given a loss function $L(m(\mathbf{y}), \theta)$, we find the estimator $\widehat{m}(\mathbf{y})$ satisfying the Bayes rule (given the Bayes rule exists and is unique, which is frequently true for most practical problems).

1.2.2. Estimation. Evaluation of survey sampling estimators often occurs within a decision-theoretic framework (Casella and Berger, 2002). For example, Lehmann and Casella

(1998, p.157) define the concept of risk-unbiasedness as a criteria for evaluating estimators. An estimator $m(\mathbf{y})$ is said to be risk-unbiased if

$$\mathbb{E}_{F_Y} L(m(\mathbf{y}), \theta) \leq \mathbb{E}_{F_Y} L(m(\mathbf{y}), \theta') \tag{2}$$

for all $\theta \neq \theta'$. In other words, an estimator is risk-unbiased if the expected loss of the estimator is minimized at the true, unknown value of θ.

In survey sampling estimation problems, squared-error loss is commonly used to evaluate estimators. There are many reasons for the popularity of squared-error loss including: evaluating unbiased estimators (where comparisons on squared-error loss simplify to choosing the unbiased estimator with the smallest variance), the similarity of squared-error loss to classical least squares theory, and the ease of calculations in a decision theoretic framework. One can argue that the reasons for squared-error loss above have little merit, in that squared-error loss might not reflect the loss function that is truly of interest (Berger, 2013, Chhapter 2).

Instead of starting by choosing a loss function to evaluate an estimator, the statistician often starts by considering a class of potential estimators and then evaluating the performance of estimators within this class. One such class of estimators is the class of unbiased estimators of the mean. Among this class of estimators, a common criterion is to choose the estimator that has minimum variance. For the class of unbiased estimators of the mean, the risk-unbiased estimator under squared-error loss is the estimator with minimum variance and therefore squared-error loss is implicitly chosen to evaluate estimator performance (Särndal et al., 2003, Chapter 2.7).

We can also view the problem of evaluating estimators from a different perspective; we can ask what loss function is consistent with the functional we aim to evaluate. First,

we define what is meant by a statistical functional (or more simply a functional). For our purposes, we define a functional as a set of mappings from a class of probability distributions $\mathcal{F}$ to the real numbers $\mathbb{R}$. For example, if θ is the population parameter of interest for the probability distribution function F_Y, then θ can typically be written in the form $\theta = \mathbf{M}(F_Y)$ for some unknown, continuous functional $\mathbf{M}$ (Horowitz and Manski, 2006). Extending this idea to a sample $\mathbf{y}$ with empirical probability distribution function F_n and given the limit of F_n exists, the functional

$$\mathbf{M}(F_Y) = \lim_{n\to\infty} \mathbf{M}(F_n), \tag{3}$$

where F_Y is the true underlying distribution of the observation $\mathbf{y}$. A functional meeting the condition in (3) is called Fisher consistent at F_Y (Huber and Ronchetti, 2011). Murphy and Daan (1985, p.391) extend this idea further, where a loss function $L(\cdot,\cdot)$ is called consistent for the functional $\mathbf{M}$ with respect to the class of probability density functions $\mathcal{F}$ if

$$\mathbb{E}_{F_Y} L(m(\mathbf{y}), \theta) \leq \mathbb{E}_{F_Y} L(\mathbf{t}, \theta) \tag{4}$$

for all probability distributions $F_Y \in \mathcal{F}_\theta$, all functions $m(\mathbf{y}) \in \mathbf{M}(F_Y)$, and $\mathbf{t} \in \mathbb{R}$, the support of θ. If the loss function in (4) holds with equality only when $\mathbf{t} \in \mathbf{M}(F_Y)$, then the loss function $L(\cdot,\cdot)$ is called strictly consistent (Gneiting, 2011). Noorbaloochi and Meeden (1983) demonstrate a duality between risk unbiasedness and consistency, arguing that the problem of evaluating estimators (risk-unbiasedness) and finding optimal estimators (Bayes rule) are connected in a dual nature.

For example, if we consider an estimator of the mean that is in the class of estimators of the mean that are asymptotically normal, then any estimator in this class is asymptotically

risk-unbiased under squared error loss. Hence, squared-error loss functions are asymptotically consistent for each member $m(\mathbf{y}) \in \mathcal{M}$ of the class of asymptotically normal estimators of θ. Many survey estimators have some form of central limit theorem, where, for large sample sizes, the distribution of the estimator is asymptotically Gaussian (Madow et al., 1948; Erdös and Rényi, 1959; Hájek, 1960, 1964; Rosen, 1972; Holst, 1973).

For survey sampling estimators that assume asymptotic Gaussian distributions, use of squared-error loss is asymptotically consistent for evaluation of estimators of the mean, and use of absolute-error loss is asymptotically consistent for evaluation of estimators of the median, where consistency is defined in (4).

Thus, instead of finding the theoretically best Bayes rule (1), we can instead choose a consistent loss (4) for the class of estimators of interest and appeal to the duality between risk-unbiasedness and finding optimal estimators (1). For estimators of the mean, it can be shown that the class of consistent loss functions is the Bregman loss, of which squared-error loss is a special case (Williams and Hooten, 2016). Thus, the problem of choosing an optimal estimator can be replaced by choosing a consistent loss for the functional of interest and demonstrating the estimator is risk-unbiased, providing some rigor in the choice of loss function and avoiding the concern over the arbitrariness of squared-error loss expressed by Berger (2013).

1.2.3. Prediction. The evaluation of predictive performance can also be framed within the decision theoretic framework presented in Section 1.2.1. Given data $\mathbf{y}$ and unobserved random variable $\mathbf{Z}$ drawn from the probability distribution F_Z on which we desire prediction, we define a scoring function $S(m(\mathbf{y}), \mathbf{z}) : \mathbb{R} \to [0, \infty)$ which maps the prediction-observation domain (the real numbers $\mathbb{R}$ for our real-valued random variable $\mathbf{Z}$) to the positive real numbers. Thus, $S(m(\mathbf{y}), \mathbf{z})$ represents the penalty incurred when the statistician predicts

$m(\mathbf{y})$ and the observation $\mathbf{z}$ is realized. Note that we assume the statistician is using observed data $\mathbf{y}$ in formulating the prediction $m(\mathbf{y})$. Similar to the Bayes rule in (1), the optimal prediction under the probability distribution $F_Z \in \mathcal{F}$ for the unobserved value $\mathbf{Z}$ is

$$\widehat{m}(\mathbf{y}) = \underset{m(\mathbf{y})}{\operatorname{argmin}} \mathbb{E}_{F_Z} S(m(\mathbf{y}), \mathbf{Z}), \tag{5}$$

which depends on the scoring function $S(m(\mathbf{y}), \mathbf{Z})$ and the unknown probability distribution of the unobserved random variable $\mathbf{Z}$. In general, one is free to choose any scoring function (just as one is free to choose any loss function), but (5) has two fundamental flaws. First, it must be the honest belief of the statistician who the optimal prediction is probabilistic whereas $\widehat{m}(\mathbf{y})$ is a point prediction (Dawid, 1984; Gneiting, 2008). Secondly, the statistician is incentivized to express different predictions when presented with different scoring functions, regardless of the statisticians true belief about nature. Thus, the Bayes prediction in (5) is unsatisfactory because use of the Bayes prediction causes a statistician to deviate from her true belief about nature (the predictive probability distribution function) to minimize the loss function under consideration.

Fortunately, there are properties of scoring functions which are direct analogs of the properties of risk-unbaisedness in (2) and consistency in (4) that are considered desirable. In particular, a statistician who wants to optimize predictive performance would like a class of scoring functions which guarantees, under expectation, that the prediction with the best score is optimal. Additionally, it is desirable to have a scoring rule that evaluates a probabilistic prediction rather than point prediction. Many commonly used scoring functions, including mean square prediction error and mean absolute prediction error, do not meet the two previous criteria.

Following the definition in (4) and Gneiting (2011), we call the scoring function S consistent for the functional $\mathbf{M}$ with respect to the class of probability densities $\mathcal{F}$ if

$$\mathbb{E}_{F_Z} S\left(m\left(\mathbf{y}\right), \mathbf{Z}\right) \leq \mathbb{E}_{F_Z} S\left(\mathbf{z}, \mathbf{Z}\right) \tag{6}$$

for all probability distributions $F \in \mathcal{F}$, all $m\left(\mathbf{y}\right) \in \mathbf{M}(F_Z)$, and all realizations $\mathbf{z}$ of the unobserved random variable $\mathbf{Z}$. The scoring function S is called strictly consistent if it is consistent and equality in (6) implies $\mathbf{z} \in \mathbf{M}(F_Z)$.

As in Section 1.2.2, there is a duality between a scoring function being consistent and a prediction being an optimal point forecast (Bayes). From Theorem 1 in Gneiting (2011), we obtain the result that the class of scoring functions that are consistent for a given functional is the same as the class of loss functions under which that functional is an optimal point forecast.

With the definition of a consistent scoring function, we can find an optimal point prediction, but our stated goal is to predict probabilistically and honestly. Thus, we need to define a proper scoring rule. Within our decision theoretic framework, a proper scoring rule is a function $\mathbf{R} : \mathcal{F} \times \mathbb{R} \to \mathbb{R}$ such that

$$\mathbb{E}_{F_Z} R\left(F_Z, \mathbf{Z}\right) \leq \mathbb{E}_{F_Z} R\left(G_Z, \mathbf{Z}\right) \tag{7}$$

for all probability distributions $F_Z, G_Z \in \mathcal{F}$, assuming that the expectations are well defined. Thus, a proper scoring rule assigns a score $R(F_Z, \mathbf{z})$ when the prediction is the probability distribution F_Z and the observation $\mathbf{z}$ is realized. Thus, if a statistician believes that the true data generating distribution for the random variable $\mathbf{Z}$ is F_Z, the statistician minimizes her expected loss by following her true beliefs. Hence, use of proper scoring rules satisfies the

two goals of predictive inference by allowing for evaluation of probabilistic predictions and reinforcing honest beliefs. Theorem 3 of Gneiting (2011) states that any scoring function S induces a proper scoring rule R in a straightforward manner. Thus, one can evaluate any probabilistic prediction using a proper scoring rule that results from a proper scoring function.

Propriety is essential for any practical application of a scoring rule; without propriety a researcher can (under expectation) reach incorrect conclusions about predictive skill. When evaluating predictions using improper scoring rules, one is not optimizing expected predictive performance, nor is it entirely obvious what one is optimizing; this can result in overly optimistic beliefs of predictive performance by researchers using improper scoring rules. Therefore, in practical applications, much care is needed to evaluate model performance and this has led to development of scoring rules that are proper in general (see Gneiting and Katzfuss (2014) for a review). For continuous data, the log score and the continuous ranked probability score (CRPS) are proper scoring rules, and for binary data, scores like the Brier and Quadratic score are proper (Gneiting and Raftery, 2007). Most often, scoring rules are negatively oriented; thus, the goal is to minimize the score, a convention I use throughout the dissertation.

1.3. Computation

Model comparison using proper scoring rules is useful only if one can compute the model fit and scoring rules in a reasonable amount of time. Modern central processing units are not experiencing the exponential increase in clock speeds seen in the past few decades, but are instead being developed with multiple cores that can run many processes in parallel. To take advantage of the changing computing environment, utilization of parallel computing

resources is quickly becoming essential to any large statistical analysis. Many statistical problems are trivial to parallelize. For instance, resampling methods that do not depend on previous samples, like bootstrapping and cross-validation, can be done in parallel, resulting in computational speedups that are nearly linear in the number of processors used. Thus, under modern computing architectures, many computational problems associated with resampling methods are ameliorated with parallel computing. Many of the applications presented in this dissertation use Bayesian models that are fit using Markov Chain Monte Carlo (MCMC) methods. There is no generic way to parallelize an MCMC algorithm, although it is possible to parallelize each independent MCMC chain, combining the results after model fitting for use in diagnostics, predictions, and evaluation. Rather than trying to improve MCMC performance through parallelization, one can increase the performance of each independent MCMC chain by using compiled code, like C++, to perform estimation. By switching to compiled code, computational speed can be increased by up to 50-100 fold relative to code in scripting languages like R. For many of the computations in this dissertation, I use both parallel computation and compiled C++ code to reduce computation time, allowing exploration of models that would otherwise be too computationally expensive.

1.4. Overview

In this dissertation, I explore the statistical properties of different statistical models applied to a variety of environmental and ecological problems. In Chapter 2, I explore survey sampling methods that improve estimation of landscape-level forest quantities using United States Forest Service Forest Inventory Analysis data. I start by fitting a number of models to predict percent canopy cover using Geographic Information System and environmental

covariates. Using the predictions from the models to construct stratified estimators of percent canopy cover over the study region allows investigation into the empirical consequences of model choice and type of stratification used to construct the estimator. Evaluation of estimator performance is achieved through a resampling experiment, treating the entire dataset as a population and constructing estimates using samples from the population. Designing the experiment in this way allows for comparison of model estimates to the true values in the population while retaining the correlations and interrelationships in the data. Each of the stratification methods result in estimators of the mean that are are unbiased, hence the focus is on the statistical properties of the variance estimator. Performance of the variance estimator is evaluated by comparing the mean variance estimate from the resampling experiment to the simulation variance of the mean estimates. By using the resampling experiment design, I explore the empirical properties of many different estimators based on complicated, real-world data, gaining insight into which methods perform well and which methods severely underestimate the variance, resulting in overly optimistic and unrealistic inference for the end user.

In Chapter 3, I use a biologically motivated model of tree ring growth to link climate variables (temperature and precipitation) to observed tree ring growth. The fitted model is then used to predict unobserved temperature and precipitation for the years 1450-1895 using tree ring width data over the period 1450-2010 and observed climate data from 1895-2010. Because the target for prediction is past, unobserved climate, there is no population dataset from which to resample climate before 1895 to validate the predictions. As an alternative, one could hold out the last few years of observed climate and test the reconstruction method's ability to predict these years; however, this is only a proxy for the true target of a long temporal backcast and one can only leave out a relatively small number of observation years

before parameter estimation is affected. Often, paleoclimate reconstructions hold-out the last 10 or 20 years of the training data and evaluate predictive performance using scores like the coefficient of efficiency (CE) and relative efficiency (RE) (Cook et al., 1994; Rutherford et al., 2005; Tingley and Huybers, 2010a,b), but the use of such short-term validation methods is fraught with problems. First, one could easily imagine a scenario where a model that can accurately reconstruct 10-20 years of climate could diverge from the unknown true climate values (i.e., a high order polynomial regression model). Also, many of the scoring rules commonly used in the paleoclimate literature, like CE and RE, are improper; thus, these scoring rules are not statistically optimal (Gneiting and Raftery, 2007). Therefore, I validate the reconstruction method using a simulation study. Simulating data as close as possible to the process that grows tree rings given climate allows for evaluation of the reconstruction method's ability to predict far backward into the past. The simulation experiment allows me to investigate how well the model estimates simulated parameters and learn about the mechanisms linking tree growth to climate.

In Chapter 4, I perform a spatio-temporal reconstruction of mean July temperature for the Upper Midwestern United States using compiled historical data. The data consist of measurements made at United States military forts over the years 1820-1893. Over these 73 years, there are a small number of fort locations at which non-standardized records of temperature were recorded (two to 36 observations per year), whereas there are approximately 20,000 spatial locations at which temperature predictions are desired. To perform the reconstruction, I use a set of 110 mean July temperature model interpolated surfaces over the study region to create a basis of possible temperature surfaces. Using the basis in a Bayesian hierarchical regression, the model generates spatial predictions over the spatial domain. I use model selection and hierarchical pooling priors to improve predictive skill. In

addition, I extend traditional principal component analysis to a framework that accounts for noisy observation of the principal components and is robust to the presence of outliers, using both simulation and resampling methods to evaluate model performance. In the simulation study, direct comparisons are made between different models using proper scoring rules, providing evidence as to which models predict with skill on the synthetic data. To validate the reconstruction using the historical data, I apply a computationally efficient approximation to leave-one-out cross-validation to investigate model performance and diagnose outlying observations.

By applying rigorous statistical model validation techniques using both resampling and simulation methods, I present a general framework for evaluating model performance in applied statistical problems, developing and evaluating novel statistical methodologies that allow for principled inference for practical environmental and ecological applications.

CHAPTER 2

Properties of the Endogenous Post-Stratified Estimator Using a Random Forests Model

2.1. Introduction

Post-stratification is used in survey estimation as a method to improve the precision of estimates by calibrating to known population quantities (Särndal et al., 2003, Ch. 7.6). The calibration is done by classifying survey observations into two or more categories called post-strata, where the population counts within each of these categories are known at the population level from some source, census or other register outside the survey. The sample stratum weights are then matched to the known population stratum weights and an estimator is constructed using the updated weights. Traditional post-stratification requires the variable on which the data are being stratified be known without error at the population level.

Post-stratum categories can be any category of interest, but are often demographic variables for surveys of human populations or land cover classes in natural resource surveys. In the United States Forest Service Forest Inventory and Analysis Program (FIA), the stratum categories are often landcover or other classes where the population counts are obtained using remote sensing data (see Bechtold and Patterson (2005) for details). In the FIA survey, field data are collected annually and used to produce estimates for a wide variety of forest attributes. The estimators are calculated as post-stratified estimators based on strata defined by landcover and other forest related categories determined from maps maintained in a geographic information system (GIS).

The maps of forest variables are often estimated from models that use reflectance values from sensors such as the Landsat Enhanced Thematic Mapper Plus (ETM+) or the Moderate

Resolution Image Spectroradiometer (MODIS), as well as bioclimatic and other ancillary GIS layers to map vegetation over a large geographic area. Each pixel of the map corresponds to a useable class, such as forest type, stand height, or crown cover. The classification is obtained by training a classification algorithm on satellite and ancillary data and then predicting the category membership for each pixel in the study area. The relationships between variables are often highly complex and non-linear, therefore motivating the common use of classification algorithms such as neural networks and Random Forests (Moisen and Frescino, 2002; Gislason et al., 2006). The end product is a digital map representing category membership for each pixel in the study area. These maps are often used by scientists and land managers for a variety of purposes, including management decisions and estimation.

In many cases, it is not realistically possible to know post-stratum counts at the population level, but it is possible to use a model to predict post-stratum membership using covariates and then stratify on these predicted values over fixed stratum boundaries. The approximating of post-stratification using model predictions is called Endogenous Post-Stratification Estimation (EPSE) and the theoretical properties of the EPSE approximation to traditional post-stratification have been discussed in Breidt and Opsomer (2008) for a parametric estimator and in Dahlke et al. (2013) for a nonparametric estimator. In traditional post-stratification, it is assumed the category membership counts over which post-stratification is performed are known without error, and the stratum membership counts are ancillary to the survey sample used for estimation. In the EPSE framework, both of these assumptions are relaxed. First, the category membership counts over which post-stratification is performed are estimated using a classification algorithm. The algorithm classifies the data, but the classification is performed with errors that result in an unknown amount of misclassification. The post-stratification categories are classified with error violates an assumption

of post-stratification, and the consequences for violating this assumption are in need of investigation. In constructing the classification algorithm, data from the survey sample are used to train the classifier, which is then used to post-stratify the sample, violating the second assumption of post-stratification that the categories over which post-stratification is performed are ancillary to the survey sample.

The use of FIA data to construct maps, and then the subsequent use of these maps as a basis for constructing post-stratified estimates, has raised questions about the appropriateness of their use based on the relaxation of the assumptions of post-stratification (Scott et al., 2005). The FIA maps are created from sample observations that are used for estimation, and the categories from the map are not without errors, violating the two assumptions of traditional post-stratification described previously. Despite the fact that EPSE violates the assumptions of post-stratification, it is useful in practice because the EPSE method makes it possible to take advantage of the considerable work that has been applied in creating good classifiers for satellite imagery and land-use mapping. When applied properly, EPSE results in significant improvements in precision for the survey estimators, thus it is of interest to continue to investigate the validity of the approach in realistic settings.

The EPSE estimator is an approximation to traditional post-stratification. In the EPSE approximation to post-stratification, the stratification variable is used to build a model based on covariates, the model is used to predict the values of the stratification variable for the entire population, and then the sample is post-stratified based on these predictions. Throughout this chapter, I focus on the EPSE approximation to post-stratification using model predictions of the stratification variable to approximate the true values of the stratification variable. A note of caution is that using the true values of the stratification variable for post-stratification of the sample is incorrect and violates the validity of this method;

when calculating sample and population stratum weights, only the modeled counts should be used. From the post-stratification approximation, I apply the post-stratification to other variables of interest. If there is a correlation between the stratification variable and the new variable of interest, the post-stratification (using the weights from the stratification variable) will increase the precision of the estimate for the new variable of interest.

Currently, the properties of the EPSE variance estimator have been investigated using fixed stratum boundaries, for example, defining strata by taking a continuous tree canopy cover layer and dividing the canopy cover values into 10 percent increments. However, it is conceptually attractive to be able to use adaptive stratum boundaries (e.g., sample quantiles in EPSE), thus guaranteeing a better spread of the sample across the categories and in particular, avoiding empty strata. In this chapter, I investigate the use of estimated stratum boundaries set as the quantiles of the model predictions in addition to the fixed boundary case. Other stratification methods are of interest but were not considered in this chapter.

Since the publication of Breidt and Opsomer (2008), there have been a number of studies that have applied an EPSE estimator explicitly stating that the assumptions of traditional post-stratification are violated. For instance, McRoberts et al. (2005) used a logistic model based on sample and ancillary information to create strata, and then used those strata for post-stratification. A similar process was performed in McRoberts (2010) and McRoberts et al. (2013). A logistic model meets the assumptions in Breidt and Opsomer (2008) and Dahlke et al. (2013) of a monotone model, but many commonly used models do not, including spline models, Random Forests, and k-nearest neighbors, a technique used in McRoberts et al. (2012). Because the EPSE estimator has been shown to be effective in reducing variance estimates, a simulation study is needed to show the conditions under which the

EPSE variance estimator is accurately estimating the true variance. For example, McRoberts et al. (2012) use a complex optimization scheme and stratum boundaries not set *a priori*. The empirical and theoretical properties of the EPSE variance estimator under these conditions have not been previously explored. As shown in the simulation, the EPSE variance estimator with an optimization step can result in underestimation of the variance. Due to the potential for underestimation of the variance, verification that the variance estimator is reliable is important in applying an EPSE estimator.

This chapter has three objectives: first, to investigate the EPSE variance estimator properties using a linear model, a spline model, and Random Forests; second, to investigate the effects of using estimated stratum boundaries instead of fixed stratum boundaries on the EPSE variance estimator; third, to investigate the effects of optimizing the stratum boundaries to minimize the variance estimate on the EPSE variance estimator. The linear model and the spline models are chosen because the theoretical properties for the EPSE estimator using these models have been investigated under simple conditions and their performance in a complex data set is of interest. Random Forests is chosen for two reasons. First, the Forest Service FIA is currently using Random Forests to create a nationwide map of percent tree canopy cover for the whole United States at a resolution of 30m×30m pixels (Coulston et al., 2012). Because the canopy cover map will likely be used as a basis for constructing post-stratified estimates, the properties of estimates constructed using this map are directly relevant to forest scientists. Second, Random Forests is a powerful and flexible tool and can easily be applied to different problems and data sets with little input from the user (Liaw and Wiener, 2002).

2.2. Endogenous Post-Stratified Estimator

Following the EPSE framework described by Breidt and Opsomer (2008), a sample s of size n is taken from a population $U \equiv \{1, ..., i, ..., N\}$ of size N according to a probability design $p(\cdot)$, where $p(s)$ is the probability of drawing the sample s. For each $i \in U$ a vector of covariates $\boldsymbol{x_i}$ and a response y_i are observed. There is assumed to be a true relationship between $\boldsymbol{x_i}$ and y_i denoted by $m(\cdot)$ where

$$\mathrm{E}\left(y_i|\boldsymbol{x_i}\right) = m(\boldsymbol{x}_i), \tag{8}$$

and $m(\boldsymbol{x}_i)$ is estimated by $\hat{m}(\boldsymbol{x}_i)$ when fit on the sample data and $\tilde{m}(\boldsymbol{x}_i)$ when fit on the population data. The EPSE approximation to post-stratification is as follows. The model predictions for the stratification variable $\tilde{m}(\boldsymbol{x}_i)$, for $i \in U$, are an approximation to the true values of the stratification variable $m(\boldsymbol{x}_i)$, for $i \in U$. The $\tilde{m}(\boldsymbol{x}_i)$, for $i \in U$, are further approximated by the values $\hat{m}(\boldsymbol{x}_i)$, for $i \in U$, where the model $\hat{m}$ is fit on the sample, and hence is endogenous to the sample.

The relationships $\hat{m}$ and $\tilde{m}$ used in this chapter will assume four different forms; a linear model, two different penalized spline models that I will differentiate as the spline model and the thin-plate spline model, and Random Forests. The linear model is of the form

$$y_i = \boldsymbol{x_i}'\boldsymbol{\beta} + \varepsilon_i \tag{9}$$

where ε_i is an independent and identically (iid) mean 0 random variable and $\boldsymbol{\beta}$ is to be estimated by ordinary least squares. Model selection for the linear model is performed using backward stepwise model selection based on Akiake's Information Criterion (AIC) to choose the variables (Akaike, 1974).

The spline model is of the form

$$y_i = \sum_{p=1}^{P} \sum_{j=0}^{J_p} \beta_{j,p} \phi_{j,p}(x_i^p) + \varepsilon_i \tag{10}$$

where the $\phi_{j,p}$s are orthonormal basis functions at the j^{th} knot for the p^{th} variable, J_p is the number of knots for the p^{th} variable, x_i^p is the value of the p^{th} variable for the i^{th} element in the sample, and the β_{jp}s are the associated coefficients to be estimated using penalized iteratively re-weighted least squares subject to a smoothness penalty (Ruppert et al., 2003). For this chapter the two spline models are fit using R and the package $mgcv$ (R Core Team, 2015; Wood, 2006). The spline model is fit with a penalty on the integral of the square of the second derivative of the fit, thereby limiting the "wiggliness" of the fit. The thin plate spline model uses a thin plate spline basis to represent the P predictors with the effective degrees of freedom for the smoothness penalty estimated through generalized cross-validation (Wood, 2006). The thin plate spline model uses a modified smoothing penalty that causes small coefficients to be penalized to zero, thereby acting as a model selection step and reducing the effects of overfitting the data (Wood, 2003). The different spline models imply different levels of smoothness and are considered to the determine the effects of model smoothness on the EPSE variance estimator.

Random Forests is implemented in R using the $randomForests$ package and fit using 4 predictors for the number of parameters to choose at each node $mtry$ and the default number of trees $ntree$ of 500 (Liaw and Wiener, 2002), corresponding to defining a subset of four predictors for each node in each tree. For details about Random Forests see Breiman (2001) and Liaw and Wiener (2002).

In EPSE with fixed stratum boundaries, the predictions $\hat{m}(\boldsymbol{x_i})$, $i = 1, \ldots, N$ are sorted into H fixed strata based on the stratum boundaries $\tau_0, \tau_1, ..., \tau_H$ where $\hat{m}(\boldsymbol{x_i})$ is in the h^{th}

stratum if $I(\tau_{h-1} < \hat{m}(\boldsymbol{x_i}) \leq \tau_h) = 1$. The estimated sample counts in stratum h are given by $\hat{n}_h$ where

$$\hat{n}_h = \sum_{i \in s} I_{\{\tau_{h-1} < \hat{m}(\boldsymbol{x_i}) \leq \tau_h\}}. \tag{11}$$

The estimated population counts in stratum h are given by $\hat{N}_h$ where

$$\hat{N}_h = \sum_{i \in U} I_{\{\tau_{h-1} < \hat{m}(\boldsymbol{x_i}) \leq \tau_h\}}. \tag{12}$$

The stratum mean $\hat{\mu}_h$ is calculated for each stratum h by

$$\hat{\mu}_h = \frac{1}{\hat{n}_h} \sum_{i \in s} y_i I_{\{\tau_{h-1} < \hat{m}(\boldsymbol{x}_i) \leq \tau_h\}}. \tag{13}$$

The EPSE estimator $\hat{\mu}_y$ for the population mean is calculated by

$$\hat{\mu}_y = \frac{1}{N} \sum_{h=1}^{H} \hat{N}_h \hat{\mu}_h. \tag{14}$$

The estimator $\hat{V}(\hat{\mu}_y)$ for the variance $\text{Var}\,(\hat{\mu}_y)$ is calculated using the post-stratified formula in Särndal et al. (2003, Chapter 7.6),

$$\hat{V}(\hat{\mu}_y) = (1 - \frac{n}{N}) \frac{1}{N^2} \sum_{h=1}^{H} \hat{N}_h^2 \frac{s_h^2}{\hat{n}_h}, \tag{15}$$

where $s_h^2 = \frac{1}{\hat{n}_h - 1} \sum_{i \in s} [(y_i - \hat{\mu}_h)^2 I_{\{\tau_{h-1} < \hat{m}(x_i) \leq \tau_h\}}]$, the sample variance for stratum h. In the simulation study, I first consider fixed stratum values τ_h following Breidt and Opsomer (2008). Also, estimated stratum values $\hat{\tau}_h$ based on quantiles of the model predictions for the set of population covariates $\boldsymbol{x_i}$, for $i \in U$, will be considered.

2.3. Empirical Properties of the Estimator

The data used in this study are from a pilot study for the 2011 National Land Cover Data (NLCD) set canopy cover map conducted in Utah (Coulston et al., 2012). The study areas consisted of 4151 sites in Utah. At each location, an aerial photograph is interpreted to estimate the values for percent canopy cover. The aerial photographs are considered highly accurate because there was a very high level of agreement among independent observers used to create the estimates (Jackson et al., 2012). I treat the aerial photograph interpreted percent canopy cover as the variable of interest. I am not claiming that the aerial interpreted estimates represent truth on the ground, only a reasonably accurate approximation to the truth on the ground. I treat these data as a population of percent canopy cover variables and ignore any discrepancies between the measured value in the pilot study and values measured in the field, allowing evaluation of the performance of the EPSE variance estimator in a population that reasonably approximates a true population with complex inter-relationships between variables. The covariates used for this study, which include a mixture of Landsat TM reflectance values, Tasseled Cap transformations (Crist and Cicone, 1984), and topographic values, are summarized in Table 2.1. Decisions about sampling intensity and variable selection for these data are described in Tipton et al. (2012).

I use two simulation studies to evaluate the effectiveness of the EPSE estimator. The first simulation in Section 2.3.1 addresses the first two goals of the chapter; the effectiveness of using Random Forests in the EPSE framework relative to other models in a realistic setting, and the effects of fixed versus estimated stratum boundaries. The second simulation in Section 2.3.2 addresses the final goal of this chapter, investigating the effects on the EPSE variance estimator due to minimizing the variance estimates by allowing the number of strata

TABLE 2.1. Remote sensing and topographic covariates used in the simulation study.

Landsat TM Band 1
Landsat TM Band 2
Landsat TM Band 3
Landsat TM Band 4
Landsat TM Band 5
Landsat TM Band 6
First Tasseled Cap Transformation of TM Bands 1-6
Second Tasseled Cap Transformation of TM Bands 1-6
Third Tasseled Cap Transformation of TM Bands 1-6
Normalized Difference Vegetative Index (NDVI) Transformation
Average Slope over 90m×90m pixel
Average value of Compound Topographic Index over 90m×90m pixel
Average value of Digital Elevation Map over 90m×90m pixel
Average value of Slope Aspect in degrees over 90m×90m pixel
sin Transformation of Slope Aspect
cos Transformation of Slope Aspect

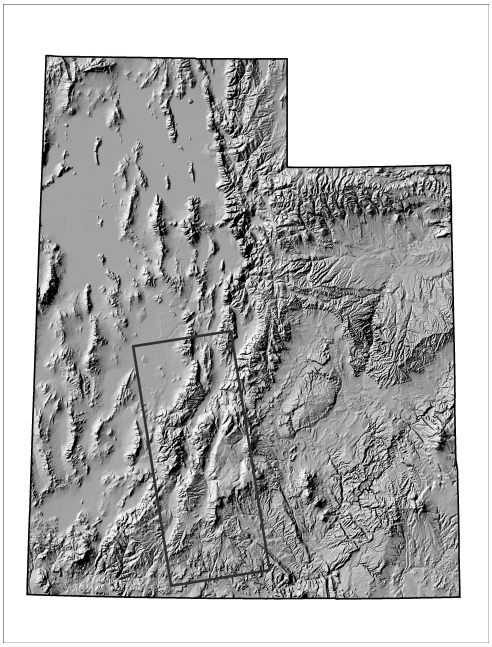


FIGURE 2.1. Image of the study region in Utah where the 4151 plots are located.

to vary. The minimization is performed by choosing the number of strata that produces the smallest variance estimate.

2.3.1. Comparison of Fixed vs. Estimated Stratum Boundaries. The first simulation study is designed to address two questions: First, how does the variance estimator performance compare with the linear model EPSE, the two spline model EPSEs, and the Random Forests EPSE? Second, how do the EPSE variance estimators perform when the stratum boundaries are fixed (this is the case where the theory is well known) versus when the stratum boundaries are estimated by sample quantiles? For notation, I define the sample-level predictions $\hat{\boldsymbol{y}}_s$ as the vector of all predictions $\hat{y}_i = \hat{m}(\boldsymbol{x_i})$ where $i \in s$, and the population level predictions $\hat{\boldsymbol{y}}_U$ as the vector of all predictions $\hat{y}_i = \hat{m}(\boldsymbol{x_i})$ for $i \in U$. For $i \in s$ the two predictions are identical (i.e., $\hat{y}_{s_i} = \hat{y}_{U_i}$ for all i); this is important because otherwise the method is not EPSE and is not valid. I also need the vector of fitted values $\tilde{\boldsymbol{y}}_U$, where the values $\tilde{y}_i = \tilde{m}(\boldsymbol{x}_i)$, for $i \in U$, are fit using the full population data. Unlike $\hat{\boldsymbol{y}}_U$, the values $\tilde{\boldsymbol{y}}_U$ are not dependent on the sample and are therefore fixed. Hence, splitting the sample into post-strata based on $\tilde{\boldsymbol{y}}_U$ can be viewed as the reference post-stratification that I am approximating when applying the EPSE estimator based on $\hat{\boldsymbol{y}}_U$. Note that, as sample size increases, I assume that the predictions $\hat{\boldsymbol{y}}_U$ are converging to the predictions $\tilde{\boldsymbol{y}}_U$.

In this investigation, I consider two stratification schemes. For both of these strata definitions the behavior of the variance of the EPSE estimator $\text{Var}\,(\hat{\mu}_y)$ and its variance estimator $\hat{V}(\hat{\mu}_y)$ are investigated for the four different estimation methods at different sample sizes n. The first stratification scheme of fixed stratum boundaries consists of the quartiles $Q_i(\tilde{\boldsymbol{y}}_U)$ of $\tilde{\boldsymbol{y}}_U$, where $Q_i(\tilde{\boldsymbol{y}}_U)$ is the $\frac{25*i}{100}$ quartile of fitted values for $\tilde{m}$, where the method of interest is fit on the full population data. The strata for the fixed simulation are set as

$$(-\infty, Q_1(\tilde{\boldsymbol{y}}_U)], (Q_1(\tilde{\boldsymbol{y}}_U), Q_2(\tilde{\boldsymbol{y}}_U)], (Q_2(\tilde{\boldsymbol{y}}_U), Q_3(\tilde{\boldsymbol{y}}_U)] \text{ and } (Q_3(\tilde{\boldsymbol{y}}_U), \infty). \tag{16}$$

These strata allow for values outside of the range [0,1] because the methods used to predict percent tree cover are not constrained to predict values between 0 and 1. Note that these quartiles are not sample dependent, hence they are fixed relative to the sample.

The second stratification scheme uses estimated stratum boundaries defined by the quartiles of the predicted tree canopy cover $\hat{m}$ fit using the sample data. The estimated strata are defined by the boundaries below which $\hat{Q}_i(\hat{\boldsymbol{y}}_U)$ is the i^{th} quartile of $\hat{\boldsymbol{y}}_U$. Note that quantiles are dependent on the sample s through the fitting of $\hat{m}$ on the sample. The estimated stratum boundaries are defined as

$$(-\infty, \hat{Q}_1(\hat{\boldsymbol{y}}_U)], (\hat{Q}_1(\hat{\boldsymbol{y}}_U), \hat{Q}_2(\hat{\boldsymbol{y}}_U)], (\hat{Q}_2(\hat{\boldsymbol{y}}_U), \hat{Q}_3(\hat{\boldsymbol{y}}_U)], \text{ and } (\hat{Q}_3(\hat{\boldsymbol{y}}_U), \infty). \tag{17}$$

Random Forests has two different methods for calculating predictions for covariates in the sample: *In Bag* (*IB*) and *Out of Bag* (*OOB*). For the simulation using Random Forests, the predictions $\hat{\boldsymbol{y}}_s$ for the sample are set to equal the *OOB* predictions. Random Forests is then used to predict the percent tree cover for the observations not in the sample $\hat{\boldsymbol{y}}_{U\setminus s}$, defined as the vector of elements $\hat{y}_i$ where $i \in U \setminus s$. These predictions are combined with the *OOB* sample predictions $\hat{\boldsymbol{y}}_s$ to create predictions at the population level $\hat{\boldsymbol{y}}_U$. The need to combine predictions is necessary because the Random Forests predictions for the population use the *IB* predictions by default and therefore $\hat{\boldsymbol{y}}_s \neq \hat{\boldsymbol{y}}_U$ for the set $\{i \in s\}$, causing the EPSE estimator to fail. To ensure that the predictions for observations in the sample agree with population level predictions, predictions are made for the sample values based on the model fit (using *OOB* predictions for the Random Forest model) and are combined with the predictions for the elements not in the sample for all of the different EPSE methods.

For each iteration of the simulation, a simple random sample is taken from the population of 4151 sites, the model of interest is fit, and $\hat{\mu}_y$ and $\hat{V}(\hat{\mu}_y)$ are calculated using equations

(14) and (15). The process is repeated for 1000 iterations, each with a different simple random sample, and the mean of the simulation variance estimates $\mathrm{E}\left(\hat{V}(\hat{\mu}_y)\right)$ are compared to the simulation variance of the post-stratified estimator $\mathrm{Var}\,(\hat{\mu}_y)$. The 1000 iterations for the simulation were chosen to balance quality of the estimates with computation time. The relative bias is computed to allow a comparison between the different sample sizes and models by the formula

$$\text{Relative Bias} = \frac{\mathrm{E}\left(\hat{V}(\hat{\mu}_y)\right) - \mathrm{Var}\,(\hat{\mu}_y)}{\mathrm{Var}\,(\hat{\mu}_y)}. \tag{18}$$

Relative bias is a measure of the relative difference between the expectation of the variance estimator, $\mathrm{E}\left(\hat{V}(\hat{\mu}_y)\right)$, and the variance of the estimator, $\mathrm{Var}\,(\hat{\mu}_y)$. The relative bias allows comparison between the average of the target value and its estimator on a scale relative to the magnitude of the target value. For instance, a relative bias value $RB = 0.05$ indicates that the estimated value is 0.05 times greater than the true value on average.

As seen in Table 2.2, for the smaller sample sizes and both stratification schemes, the EPSE estimators using the spline model and Random Forests generally have the smallest variances among the four EPSE estimators, the linear model and thin plate spline model EPSE estimators also have similar variances, but are generally slightly larger. For larger sample sizes and for both stratification schemes, the EPSE estimator using Random Forests has the smallest variance most often, the spline model EPSE estimator generally has the next smallest variances, and the linear model and thin-plate spline EPSE estimators have the largest variances.

In Figure 2.2, all four EPSE methods appear to have relative biases that are reasonably small across all sample sizes and across both stratification schemes, except for small sample sizes for the spline model, indicating that the EPSE framework is relatively robust. The

TABLE 2.2. Table of variances for the four EPSE estimators for the fixed (Fix) and estimated (Est) stratification schemes. The linear model is abbreviated LM, the spline model is abbreviated SM, the thin-plate spline model is abbreviated TPS, and Random Forests is abbreviated RF. The values of the variances in the table are scaled by a factor of 10^{-5}. The NA for Random Forests with fixed stratum boundaries comes from having empty strata when Random Forests was fit on a sample of size 100.

n	LM Fix	SM Fix	TPS Fix	RF Fix	LM Est	SM Est	TPS Est	RF Est
100	41.35	50.30	41.77	NA	41.95	47.88	41.37	41.86
200	19.87	18.74	19.66	19.86	19.80	18.73	19.82	19.44
300	13.29	11.63	13.44	11.37	13.15	11.34	13.16	11.27
400	9.470	8.302	9.605	8.419	9.394	8.326	9.297	8.251
500	7.470	6.382	7.480	6.470	7.289	6.386	7.390	6.308
1000	3.095	2.766	3.250	2.589	3.067	2.788	3.232	2.571
1500	1.820	1.576	1.777	1.317	1.823	1.595	1.797	1.321
2000	0.953	0.953	1.191	0.840	0.938	0.965	1.179	0.829
2500	0.663	0.543	0.639	0.572	0.664	0.542	0.643	0.579
3000	0.374	0.340	0.382	0.314	0.373	0.337	0.381	0.321
3500	0.191	0.169	0.192	0.155	0.191	0.168	0.192	0.156
4000	0.037	0.031	0.035	0.029	0.037	0.031	0.035	0.029

linear model EPSE shows evidence of a symmetric distribution of relative bias about 0. For the spline EPSE, it can be seen that the relative bias appears to be skewed to negative values, especially for small sample sizes, suggesting that the variance estimate is underestimating the variance of the estimator. In some of these scenarios, underestimation of the variance by as much as 40% occurs. The plot of relative bias for thin plate spline EPSE shows some evidence of underestimation of the variance, as the estimate seems to be skewed toward negative values. For the Random Forests EPSE, the relative bias appears to have a symmetric distribution of positive and negative values, suggesting that the variance estimator is unbiased for the variance of the estimator. The underestimation of the variance by the variance estimator for the two spline model EPSEs is a problem because underestimation of the variance can cause reduced coverage of confidence intervals and overly optimistic results in hypothesis testing situations.

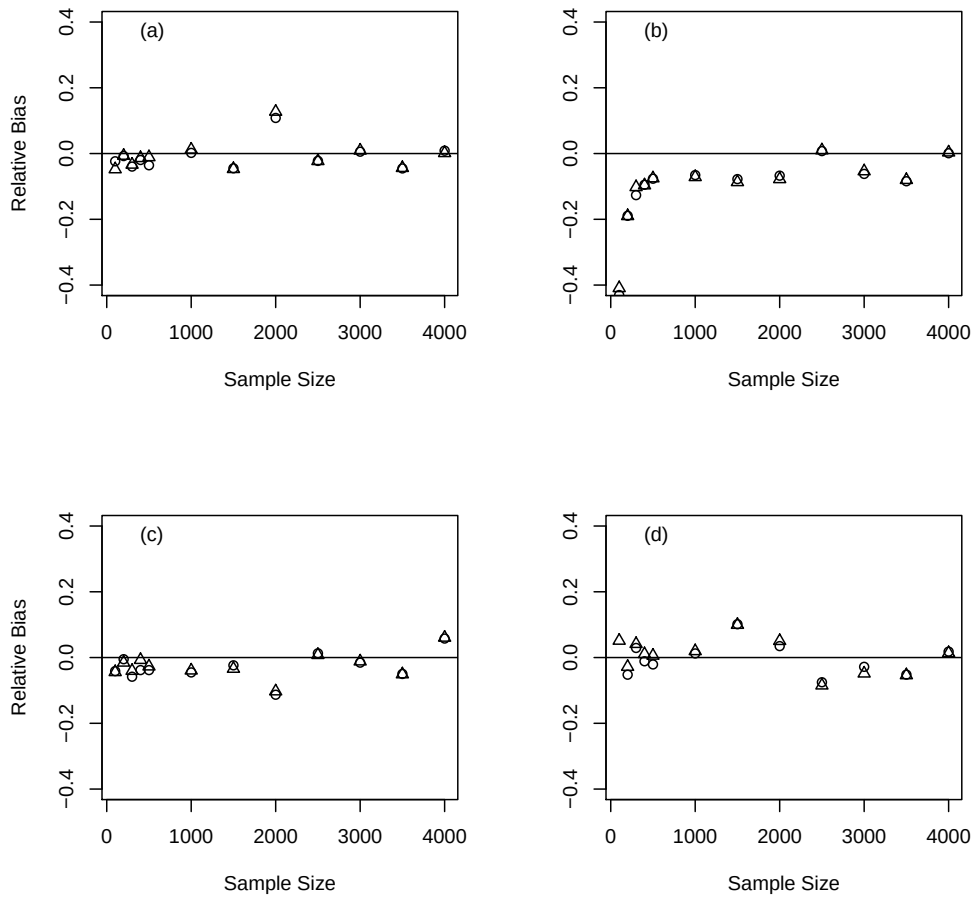


FIGURE 2.2. Plot of Relative Bias for EPSE variance estimators. Subplot (a) is for EPSE estimates constructed using a linear model, subplot (b) is for EPSE estimates constructed using a spline model with B-spline basis, subplot (c) is for EPSE estimates constructed using a spline model with a thin-plate spline basis, and subplot (d) is for EPSE estimates constructed using Random Forests. Circles represent the EPSE estimates constructed using 4 fixed strata, and triangles represent the EPSE estimates constructed using 4 strata estimated by the quartiles of the model predictions.

For a given EPSE method and sample size, the variance of the EPSE estimator for the fixed stratum boundaries stratification scheme appears quite similar to the variance of the EPSE estimator for the estimated stratum boundaries stratification scheme (Table 2.2). These similarities in EPSE variances suggest that the estimated quantiles are accurately estimating the population quantiles. That the variance of the EPSE estimator using the

estimated quantiles appears to be approximating the variance of the EPSE estimator using the population quantiles is reasonable because it has been shown that sample quantiles of a distribution converge to the true population quantiles under mild assumptions (Bahadur, 1966).

In light of the results from Table 2.2 and Figure 2.2, I summarize the findings as follows. The best EPSE estimator in the simulation is the Random Forests EPSE due to having small variances and having relative biases of the variance estimator distributed evenly around 0. The spline model EPSE estimator has small variance, but the variance estimator appears to be underestimating the variance, thus calling into question the spline model EPSE's validity. The linear model EPSE's variance estimator appears not to be biased, but the EPSE estimator has larger variance than either the spline model or Random Forest EPSE. The thin plate spline EPSE variance estimator appears to be only slightly underestimating the variance, but with a variance comparable to the linear model EPSE.

The EPSE estimator appears to be relatively robust for the different methods and under estimated stratum boundaries as long as care is taken to ensure that $\hat{\boldsymbol{y}}_s$ and $\hat{\boldsymbol{y}}_U$ agree for all $i \in s$. The simulation supports that the class of methods used to construct the EPSE estimator can be extended to include Random Forests. The simulation also supports the use of the EPSE estimator when the stratum boundaries are estimated quantiles from the model predictions instead of fixed stratum boundaries. One benefit of using quantiles from the predicted values is that the construction guarantees non-empty strata.

2.3.2. Properties of Minimization of EPSE Variance Estimates. For the second simulation, I address the third goal of this chapter. The question of interest is whether one can use the EPSE variance estimates to optimize the number of EPSE strata to minimize the variance estimate. This is done by selecting the number of strata to construct the

smallest variance estimate $\hat{V}(\hat{\mu}_y)$. The aim is to determine whether it is still acceptable to use the minimized variance estimate as a valid estimate of the variance. For computational efficiency, the predictions for the EPSE method of interest are split into strata of equal number of elements based on the quantiles of the model predictions using the covariates for the sample. Minimization of the variance estimate is then performed by selecting the number of strata that results in the smallest estimate $\hat{V}(\hat{\mu_y})$. Other ways of dividing the predictions into strata are possible but not considered in this chapter. For instance, McRoberts et al. (2012) and McRoberts et al. (2013) divided model predictions into 100 equally spaced bins and then collapsed these bins into between one to six strata requiring at least 10 plots per stratum. The collapsing of the bins into strata was done to minimize the variance estimate. A future extension of the work in this chapter would be to consider the EPSE variance estimator under the previous optimization scheme.

I begin the second simulation experiment as before, taking a random sample of size n from the population of 4151 sites in the Utah data set. For each sample, I fit a linear model, the two spline models, and Random Forests using the covariates described previously. Using predictions generated from the model fits, I combine the predictions with the stratification scheme to construct the classifiers used in constructing the EPSE estimators. To ensure that the predictions for observations in the sample agree with population level predictions, predictions are made for the sample values based on the model fit (using OOB predictions for the Random Forest model) and are combined with the predictions for the elements not in the sample, as described in Section 2.3.1.

After computing the predictions from the methods of interest for the population, the strata over which optimization is to be performed can be created. Let k denote the number of strata being considered. For $k = 1$, there is one stratum and the estimator is equivalent

to simple random sampling. For $k = 2$, the stratification scheme is equivalent to stratifying the data based on the median of the predicted values. For arbitrary $k, k = 1, \ldots, 10$, define the quantiles over which stratification will be performed as

$$\hat{Q}_{h,k}(\hat{\boldsymbol{y}}_U) = \left[\frac{h}{k}\right]^{th} \text{ percentile of } \hat{\boldsymbol{y}}_U \text{ for } k = 1, 2, ..., 10, h = 1, ..., k. \tag{19}$$

I count the number of population-level prediction counts in each stratum and define

$$\hat{N}_{h,k} = \sum_{i \in U} I(i \in \text{stratum } h \text{ out of } k \text{ strata}). \tag{20}$$

By stratifying the sample-level predictions $\hat{\boldsymbol{y}}_s$ based on (19) I count the number of sample-level predictions in each stratum and define

$$\hat{n}_{h,k} = \sum_{i \in s} I(i \in \text{stratum } h \text{ out of } k \text{ strata}). \tag{21}$$

The stratum means for the h^{th} stratum out of a total of k strata are defined as

$$\hat{\mu}_{h,k} = \frac{1}{\hat{n}_{h,k}} \sum_{i \in s} y_i I(i \in \text{ stratum } h \text{ out of the total } k \text{ strata}). \tag{22}$$

The estimator $\hat{\mu}_{y,k}$ for the population mean based on k strata is defined as

$$\hat{\mu}_{y,k} = \frac{1}{N} \sum_{h=1}^{k} \hat{N}_{h,k} \hat{\mu}_{h,k}. \tag{23}$$

The variance estimator for each stratum is similar to (15), but has been adapted to reflect the notation that allows for different numbers of strata. The variance estimator for k strata

is defined as

$$\hat{V}(\mu_{y,k}) = (1 - \frac{n}{N})\frac{1}{N^2}\sum_{h=1}^{k}\hat{N}_{h,k}^2\frac{s_{h,k}^2}{\hat{n}_{h,k}}, \tag{24}$$

where

$$s_{h,k}^2 = \frac{1}{\hat{n}_{h,k}-1}\sum_{i\in s}(y_i - \hat{\mu}_{h,k})^2 I(i \in \text{stratum } h \text{ out of } k) \tag{25}$$

is the sample variance in stratum h out of a total of k strata.

For each iteration in the simulation, choose the minimum variance estimate across the 10 possible strata choices, $\hat{V}(\hat{\mu}_{y,k^*})$, where $\hat{V}(\hat{\mu}_{y,k^*}) = \min_k \hat{V}(\hat{\mu}_{y,k})$, $k = 1, \ldots, 10$. In addition, select the EPSE estimator of the mean $\hat{\mu}_{y,k^*}$ for the optimal number of strata k^*. After 1000 iterations, I compute the mean of the simulation variance estimates $\mathrm{E}\left(\hat{V}(\hat{\mu}_{y,k^*})\right)$ and the simulation variance of the EPSE estimates $\mathrm{Var}\,(\hat{\mu}_{y,k^*})$. The relative bias of the variance estimator is defined as

$$\text{Relative Bias } = \frac{\mathrm{E}\left(\hat{V}(\hat{\mu_{y_{k^*}}})\right) - \mathrm{Var}\,(\hat{\mu}_{y,k^*})}{\mathrm{Var}\,(\hat{\mu}_{y,k^*})}. \tag{26}$$

The procedure is repeated for each of the four EPSE methods considered and for various sample sizes. The results for the variance $\mathrm{Var}\,(\hat{\mu}_{y,k^*})$ are summarized in Table 2.3 and the results for the relative bias are summarized in Figure 2.3.

From Table 2.3 I can rank the EPSE estimator methods based on their variances. For sample sizes less than 500, the spline model EPSE has the smallest variance, the Random Forests EPSE has the next smallest variance, the linear model EPSE has the next smallest variance, and the thin-plate spline EPSE has the largest variance. For sample sizes over 500, the Random Forests EPSE has the smallest variance, followed by the spline model EPSE

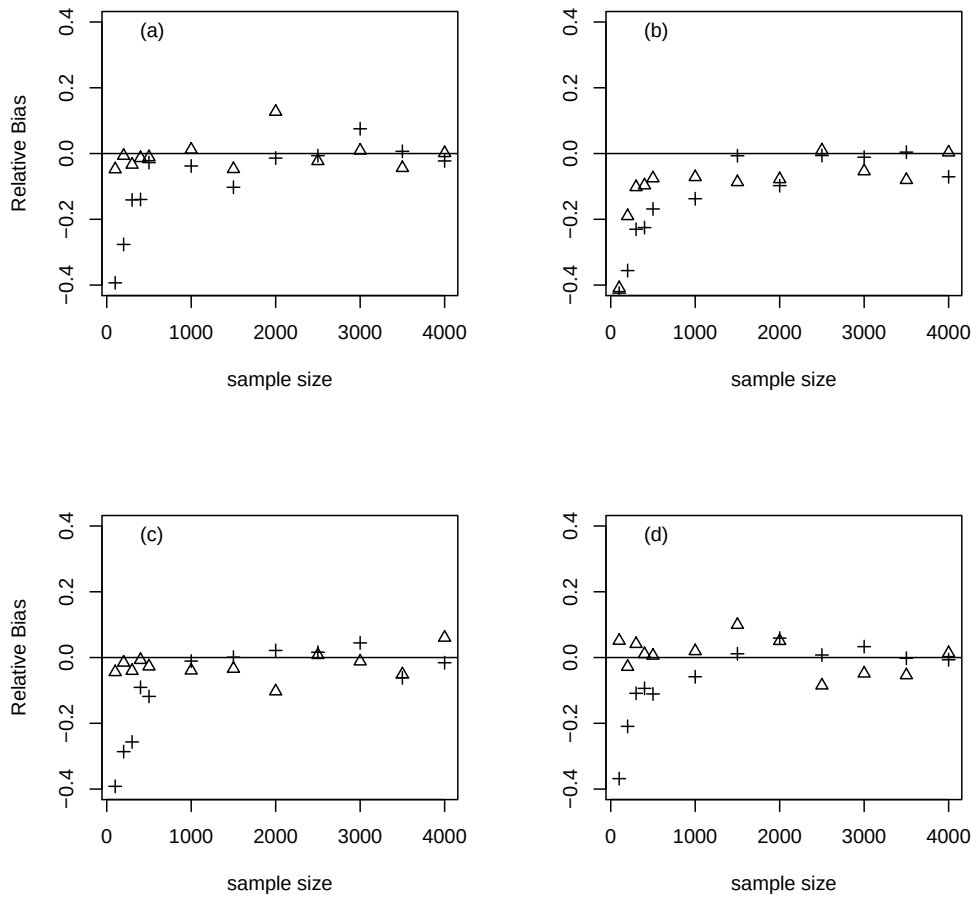


FIGURE 2.3. Plot of Relative Bias for optimized EPSE estimates. Subplot (a) is for EPSE estimates constructed using a linear model, subplot (b) is for EPSE estimates constructed using a spline model with B-spline basis, subplot (c) is for EPSE estimates constructed using a spline model with a thin-plate spline basis, and subplot (d) is for EPSE estimates constructed using Random Forests. Crosses represent the EPSE estimates constructed optimizing up to 10 strata, and triangles represent the EPSE estimates constructed using exactly 4 strata estimated by the quartiles of the model predictions.

with the next smallest variance, and then both the thin-plate spline model EPSE and the linear model EPSE appear to have similar variances.

For comparison, the variances for the four EPSE methods using the estimated boundary stratification scheme from the first simulation are also listed in Table 2.3. For sample sizes less than 500, the optimized EPSE methods have variances larger or about the same as

TABLE 2.3. Table of variances for the four EPSE estimators for the optimized (Opt) stratification scheme and the estimated (Est) stratification scheme from the first simulation. The linear model is abbreviated LM, the spline model is abbreviated SM, the thin-plate spline model is abbreviated TPS, and Random Forests is abbreviated RF. The values of the variances in the table are scaled by a factor of 10^{-5}.

n	LM Opt	SM Opt	TPS Opt	RF Opt	LM Est	SM Est	TPS Est	RF Est
100	54.26	48.29	52.58	56.25	41.95	47.88	41.37	41.86
200	23.41	19.32	23.61	20.96	19.80	18.73	19.82	19.44
300	12.97	11.06	14.94	11.52	13.15	11.34	13.16	11.27
400	9.596	8.137	8.990	8.110	9.394	8.326	9.297	8.251
500	6.588	6.013	7.226	6.263	7.289	6.386	7.390	6.308
1000	2.893	2.553	2.811	2.432	3.067	2.788	3.232	2.571
1500	1.739	1.255	1.558	1.249	1.823	1.595	1.797	1.321
2000	0.965	0.842	0.929	0.714	0.938	0.965	1.179	0.829
2500	0.588	0.470	0.575	0.456	0.664	0.542	0.643	0.579
3000	0.316	0.275	0.325	0.256	0.373	0.337	0.381	0.321
3500	0.164	0.131	0.175	0.128	0.191	0.168	0.192	0.156
4000	0.034	0.029	0.034	0.026	0.037	0.031	0.035	0.029

the EPSE variances using the estimated strata boundaries, suggesting that optimizing the EPSE stratum boundaries for small sample sizes can actually increase the variance of the estimates in ways not accounted for in the variance estimator. The presence of unaccounted for variance explains why the optimized EPSE variance estimators are underestimating the true variance for small sample sizes. For sample sizes greater than 500, the optimized EPSE variances are smaller than the EPSE estimators using four strata, but it is questionable if the gain in precision is large enough to offset the concerns over the potential for negative bias in the variance estimator.

From Figure 2.3 I see that the optimization step introduces a negative bias into the four EPSE variance estimators for small sample sizes that is not seen when using a fixed number of strata, suggesting that optimizing the EPSE estimator has the consequence of underestimation of the variance for small sample sizes. For sample sizes greater than 500, Figure 2.3 suggests that the optimized linear model, the optimized thin-plate spline model,

and the optimized Random Forests EPSE variance estimators appear to be symmetrically distributed around a relative bias of 0. The spline model EPSE variance estimator appears to be negatively biased across all sample sizes.

The second simulation study examines the use of Random Forests, two spline models, and a linear model in constructing an optimized EPSE variance estimator, suggesting that the EPSE variance estimator has potential to underestimate the variance when optimization is performed. From the results discussed above, I rank the methods used for constructing the EPSE estimates. If the goal is to have the best trade-off between bias of the optimized EPSE variance estimator and a minimum variance for the optimized EPSE estimator, the Random Forests EPSE variance estimator performs the best overall, followed by the linear model EPSE variance estimator, the thin-plate spline EPSE variance estimator, and then the spline model EPSE variance estimator.

2.4. Conclusion

I have shown that the use of complex multivariate parametric, semiparametric, and nonparametric classifiers in the EPSE framework appears to be relatively robust in constructing variance estimates, but some cautions are raised to avoid underestimating the variance. This study lends strength to the idea that EPSE can be applied to stratum boundaries that are estimated quantiles of the predictions rather than fixed stratum boundaries, although only the latter case of fixed stratum boundaries has been supported by theory. The theory exploring the effects on the EPSE estimator using estimated stratum boundaries is an area for further research, as in practice, it is often attractive to use estimated stratum boundaries not fixed *a priori*. Investigation into using estimated stratum boundaries in a theoretical

context would provide more justification of the use of estimated quantiles as post-stratum boundaries.

The best performing, most robust method throughout the study has been the Random Forests EPSE variance estimator with the strata boundaries set as the quantiles of the model predictions. The next best performing EPSE method has been the linear model EPSE variance estimator with the strata boundaries set as the quantiles of the model predictions. From the results in simulation two, I see that optimization of the EPSE variance estimates can have the unintended consequence of introducing negative bias into the EPSE variance estimator.

The EPSE variance estimator using Random Forests performed well in each simulation and across all sample sizes, except for the optimized EPSE small sample sizes. This is a practically important result in that there is almost no tuning needed by the user to fit Random Forests, thus supporting the FIA's use of maps of landcover and percent tree cover created by Random Forests as a basis for using Endogenous Post-Stratification as a way to increase precision of FIA estimates.

Recognition of Support

The authors would like to thank FIA for support of this research. In addition, thanks go out to FIA and the Remote Sensing Applications Center for the dataset used in this study. The authors would also like to thank the three reviewers for their contributions to the writing of this chapter.

CHAPTER 3

Reconstruction of late Holocene climate based on tree growth and mechanistic hierarchical models

3.1. Introduction

Statistical estimation of past climate is important for understanding climate change in a historical context and for predicting how climate will respond in the future (Stocker et al., 2013). Ideally, one would model climate with a long time-series of spatially explicit, highly precise instrumental measurements. However, the instrumental record only spans the last one to two hundred years; perhaps less in many areas. Paleoclimate reconstructions allow for investigation of climate dynamics at longer time scales than instrumental records and serve as a test bed to evaluate performance of complex modern climate models. In the absence of a dense network of instrumental observations, we must rely on climate proxy data to gain a better understanding of climate history. Evans et al. (2013) describe a conceptual model for how proxy processes integrate physical, chemical, biological, and geological climate information to yield the observed data. Their work calls for the development of mechanistic proxy system models to describe how climate influences the proxy observations. Among available proxy data sources, many late Holocene paleoclimate reconstructions focus on tree ring widths because tree ring width data are widely available on a regional or hemispheric scale, can contain hundreds or thousands of years of observations, are sensitive to temperature, precipitation, and drought, and have a very clear annual to seasonal resolution (Jones et al., 1998; D'Arrigo et al., 2004; Moberg et al., 2005; Mann et al., 2008; Christiansen and Ljungqvist, 2011; Griffin et al., 2013). Our contribution is to improve current statistical

models for reconstructing late Holocene climate from annually resolved tree ring proxy data and develop a framework for future model development.

The statistical reconstruction of paleoclimate histories from tree ring width data poses many challenges (Jones et al., 2009). First, tree rings are formed through a broad range of climatic, ecological, and growth allocation factors that make them noisier than instrumental records. Dendrochronologists typically process the tree ring width data in an attempt to remove the non-climate factors influencing growth (Cook and Kairiukstis, 1990). After removing most of these non-climatic effects, the dendrochronologist aggregates the many tree ring widths at a site into a tree ring chronology. The final filtered tree ring chronology consists of one time series derived from a number of tree cores of a particular species exhibiting a coherent signal (Cook and Kairiukstis, 1990). The tree ring standardization procedures have been thoroughly discussed in the literature and, as such, we treat the chronologies as observed data (Melvin and Briffa, 2008).

One challenge in the reconstruction of paleoclimate from tree ring width data is that the climate signal influencing tree growth occurs in continuous time whereas we typically summarize climate and tree ring widths in discrete time. A tree growth increment represents the integrated response of the tree to climate conditions over a growing season(s), collapsing sub-annual climate information into a univariate value, annual tree ring width (Fritts, 1976; Bradley, 2011; Carbone et al., 2013). The model of tree rings we develop considers climate on a monthly time step. However, there is only one tree ring observation per year, leading to a temporal change of support problem (Gotway and Young, 2002).

Another factor to consider is that climatic influences on tree ring growth are multivariate, typically involving temperature and precipitation. The joint estimation of temperature and precipitation from a univariate tree ring width observation requires inverting a multivalued

functional which has an infinite number of equally likely solutions (Tolwinski-Ward et al., 2014). Without additional information or constraints, it is impossible to overcome this loss of climate information. Further complicating a multivariate climate reconstruction is the development of site-selection techniques in dendrochronology that select for univariate climate signals in tree ring chronologies, resulting in a nonrandom sample of chronologies (Cook and Kairiukstis, 1990).

The problem of reconstructing climate using univariate tree ring chronologies has been addressed in the literature using a variety of methods. Many authors have attempted to solve the climate reconstruction problem with linear statistical methods that allow for estimation in the presence of a rank deficient design matrix, including regularized expectation-maximization algorithms, truncated total least squares, and multivariate calibration methods including partial least squares (Rutherford et al., 2003; Zhang et al., 2004; Rutherford et al., 2005; Mann et al., 2008; Steig et al., 2009). These methods all assume a linear relationship between the observed climate and tree ring chronologies. More recently, investigators have developed new methods using correlated spatial random effects (Guillot et al., 2015) or non-linear processes (Tolwinski-Ward et al., 2014) to link climate and proxy data. Other statistically rigorous work has been done to investigate the environmental mechanisms of tree growth using dendrochronological data (Hooten and Wikle, 2007). To develop the best methodology for paleoclimate prediction, Tingley et al. (2012) recommend collaboration between climate scientists and statisticians to develop a Bayesian hierarchical framework that combines scientifically motivated processes with flexible spatio-temporal methods. Our contribution is to bridge the gap between the linear multivariate calibration and regularized expectation-maximization methods and the more mechanistic, ecologically motivated

methods. Thus, we approach the problem of multivariate climate reconstruction in a novel way.

Our work is based on Tolwinski-Ward et al. (2014), where we make many computational and methodological improvements. First, we propose different forms of the deterministic growth function linking climate observations to tree ring widths within a framework that allows for a statistically principled evaluation of the effects of different growth model forms. Next, we constrain the multivariate climate predictions by modeling a differential growth response for each tree species, as is common in multispecies ecological modeling. The multispecies approach ameliorates the multivalued inverse problem while gaining the additional benefit of inferring niche climate response of each tree species. We propose a computationally efficient calibration model to link observed tree ring widths to the deterministic growth model output speeding algorithmic convergence and improving mixing during parameter estimation. We use an upscaling data model that links monthly scale climate to annual scale observed tree ring chronologies and propose a dynamic model for downscaling annual tree ring observations to monthly climate anomalies. To facilitate climate backcasting, we propose a novel model for temperature and precipitation using a dynamic, flexible, multi-scale process. Finally, we conduct a simulation experiment to validate the model's predictive ability using a proper scoring rule that selects the optimal predictive model. We do not include a direct comparison of our model to that of Tolwinski-Ward et al. (2014) for a number of reasons. First, Tolwinski-Ward et al. (2014) models soil moisture, a nonlinear function of temperature and precipitation, instead of precipitation, making a direct comparison difficult. In addition, the model of Tolwinski-Ward et al. (2014) was fit to a reconstruction period of 50 years and took approximately 350 CPU hours to fit in a high performance computing

environment. We aim to reconstruct 456 years thus our model needs to be computationally feasible for longer time scales.

We introduce the climate data in Section 3.2.1, detailing the transformation of the climate data to anomaly space. In Section 3.2.2, we propose a calibration model to align the observed tree ring data with output from a deterministic tree ring growth model, thereby assimilating the climate and tree ring data. We describe the deterministic link function that takes climate inputs and grows synthetic tree ring widths in Section 3.2.3 and present a dynamic, multi-scale model that facilitates backcasting at a monthly scale in Section 3.2.4. Finally, we formulate the posterior distribution on which inference is desired and outline the sampling algorithm used for estimation in Section 4.4.

Ultimately, our modeling effort is successful if accurate predictions of historical climate are obtained along with associated uncertainty. In Section 3.3, we discuss a scoring rule commonly used in dendrochronology and describe an alternative that is proper. In Section 4.6, we present the pseudoproxy simulation study and interpret the results of this experiment. We conclude by presenting our reconstruction of temperature and precipitation for the Hudson Valley data in Section 3.5 and discuss these results in Section 3.6.

3.2. The Model

3.2.1. Climate Data Model. Our Instrumental period climate data are monthly PRISM (Parameter-elevation Relationships on Independent Slopes Model) gridded data products from 1895 to 2005 (PRISM Climate Group, Oregon State University). From PRISM, we obtain temperature and precipitation at $I = 16$ sites within the Hudson Valley of New York, USA. Monthly temperature T_{ts} and log total precipitation P_{ts} represent regional averages

across the I sites for year t and month s. For our data, there are no months with zero precipitation and the log transform is well defined, although this is not true in general. Within a given month, the temperature and log precipitation measurements are approximately normally distributed, hence, we model them using Gaussian distributions. It is common to model the climate dynamics in anomaly space, resulting in standard normal temperature and log precipitation anomalies

$$w_{Tts} = \frac{T_{ts} - \bar{T}_s}{s_{T_s}} \tag{27}$$

$$w_{Pts} = \frac{P_{ts} - \bar{P}_s}{s_{P_s}}, \tag{28}$$

where $\bar{T}_s$, $\bar{P}_s$, s_{T_s}, and s_{P_s} are the sample means and standard deviations of temperature and log precipitation for month s, respectively. We define the monthly scale anomaly temperature and log precipitation vectors for year t, $\boldsymbol{w}_{Tt} \equiv (w_{Tt1}, \ldots, w_{Tt12})'$ and $\boldsymbol{w}_{Pt} \equiv (w_{Pt1}, \ldots, w_{Pt12})'$, the bivariate climate anomaly vector for year t, $\boldsymbol{w}_t \equiv (\boldsymbol{w}'_{Tt}, \boldsymbol{w}'_{Pt})'$, and the vector of all climate anomalies, $\boldsymbol{w} \equiv (\boldsymbol{w}'_1, \ldots \boldsymbol{w}'_\tau)'$, for years $t = 1, \ldots, \tau$.

3.2.2. Tree Ring Data Model. The 34 tree ring chronologies contain measurements from $J = 12$ different tree species in the Hudson Valley region of New York, shown in Table 3.1. Each of the tree ring chronologies are at least 160 years long and three chronologies date back to 1453. Further details about how the tree ring chronology data were collected and processed can be found in Pederson et al. (2013).

The tree ring observation y_{itj} represents the annual observed tree ring width from the i^{th} location for species j at time t. We model the tree ring width data as arising from a mixture of two distributions that depend on different forms of a deterministic growth model response

TABLE 3.1. Species used in the reconstruction.

Species	Number of Chronologies
T. canadensis	3
L. tulipifera	3
J. virginiana	1
C. glabra	3
Q. stellata	1
B. lenta	2
P. rigida	1
Q. montana	5
Q. rubra	4
Q. alba	5
Q. velutina	3
C. ovata	1
C. thyoides	2

to climate:

$$y_{itj}|\boldsymbol{w}_t, \boldsymbol{\theta}_j^{VS}, \boldsymbol{\theta}_j^{Pro}, z_j, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\sigma}^2 \sim \begin{cases} \mathrm{N}\left(\beta_{0_j} + \beta_{1_j} f\left(\boldsymbol{w}_t, \boldsymbol{\theta}_j^{VS}\right), \sigma_j^2\right) & \text{if } z_j = 0 \\ \mathrm{N}\left(\tilde{\beta}_{0_j} + \tilde{\beta}_{1_j} \tilde{f}\left(\boldsymbol{w}_t, \boldsymbol{\theta}_j^{Pro}\right), \tilde{\sigma}_j^2\right) & \text{if } z_j = 1, \end{cases} \tag{29}$$

where $f(\boldsymbol{w}_t, \boldsymbol{\theta}_j^{VS})$ and $\tilde{f}(\boldsymbol{w}_t, \boldsymbol{\theta}_j^{Pro})$ are deterministic link functions that "grow" tree rings given the climate anomaly $\boldsymbol{w}_t$ and species specific model parameters $\boldsymbol{\theta}_j^{VS}$ and $\boldsymbol{\theta}_j^{Pro}$ for the two different growth models denoted as VS and Pro to be discussed in more detail later. In Sections 3.2.3, 3.2.3.1, and 3.2.3.2, we describe the structure of the growth link function and the growth model parameters $\boldsymbol{\theta}_j$. The stochastic indicator variable z_j selects the growth model form appropriate for species j. Examination of the posterior distribution of $\boldsymbol{z}$ provides a statistically principled method for comparing proposed growth model forms. For $j = 1, \ldots, J$, we specify a binomial prior on z_j with prior probability 0.5 to allow each growth model to be equally likely *a priori*.

In contrast with Tolwinski-Ward et al. (2014), where they standardize the tree ring growth model output to have the same mean and standard deviation as the observed tree

ring chronology, we use (29) to calibrate the tree ring growth model output to the observed chronology. The growth model specific parameters β_{0j} and β_{1j} ($\tilde{\beta}_{0j}$ and $\tilde{\beta}_{1j}$) center and scale the synthetic tree ring growth model output to be coherent with the observed tree ring widths up to an error with variance σ_j^2 ($\tilde{\sigma}_j^2$), where the ($\sim$) distinguishes the growth model form. We specify priors for the VS growth calibration model parameters, with those for the Pro being defined similarly. For each species $j = 1, \ldots, J$, we specify a hierarchically pooled prior across species $\beta_{0j} \sim \mathrm{N}(\mu_{\beta_0}, \sigma^2_{\beta_0})$ and $\beta_{1j} \sim \mathrm{N}(\mu_{\beta_1}, \sigma^2_{\beta_1})$ with hyperpriors $\mu_{\beta_0} \sim \mathrm{N}(1,1)$, $\sigma^2_{\beta_0} \sim \mathrm{IG}(1,1)$, $\mu_{\beta_1} \sim \mathrm{N}(1,1)$, and $\sigma^2_{\beta_1} \sim \mathrm{IG}(1,1)$. For the $j = 1, \ldots, J$ calibration standard deviation parameters we define the prior $\sigma_j \sim \mathrm{logN}(\mu_{\sigma^2}, \sigma^2_{\sigma^2})$ with hierarchical pooling hyperparameters $\mu_{\sigma^2} \sim$N$(0, 10)$ and $\sigma^2_{\sigma^2} \sim$IG$(1,1)$. where N, logN, and IG refer to the normal, lognormal, and inverse Gamma distributions.

3.2.3. Process Model. The statistical learning about past climate is achieved through a deterministic tree ring growth model that uses monthly temperature and precipitation as inputs. Formation of tree ring widths occur periodically throughout the growing season, with the rate of growth influenced by the prevailing weather. However, the data and the forward model are in the form of monthly climate variables and annual tree growth increments. The monthly temporal scale of the temperature and precipitation presents a change of support problem because the observed tree ring data occur at annual, not monthly, resolution. Hence, prediction of climate at a monthly scale involves downscaling the annual resolution tree ring information into monthly increments. In years without climate observations, we use (27) and (28) to accomplish the downscaling by imposing the observed monthly climate patterns. The downscaling assumes that the monthly patterns and dynamics of temperature and precipitation within a given year are estimable from the observational period and that these patterns are representative of the reconstruction period. For example, because

temperature is strongly seasonal, the pattern of warm temperatures in summer and cold temperatures in winter will be the same regardless of the absolute magnitude of the annual or decadal temperature patterns. Precipitation is less consistent, but annual variability in these patterns allows for realistic downscaling. To align the different data sources occurring at different scales, we use a discrete time approximation of the continuous growth process on a monthly scale, thus aligning tree growth with the PRISM data. We then aggregate the monthly growth increments to an annual resolution, thereby upscaling the continuous growth from temperature and precipitation for year t, species j, and month s into the growth increments. Thus, the representation of annual tree ring growth under the VS model form (the Pro model form is defined similarly, replacing $\boldsymbol{\theta}_j^{VS}$ with $\boldsymbol{\theta}_j^{Pro}$, f with $\tilde{f}$ and g with $\tilde{g}$) is

$$f\left(\boldsymbol{w}_t, \boldsymbol{\theta}_j^{VS}\right) = \sum_{s=1}^{12} \chi_s \min\left(g\left(w_{T_{ts}}, \boldsymbol{\theta}_j^{VS}, \bar{T}_s, s_{T_S}\right), g\left(w_{P_{ts}}, \boldsymbol{\theta}_j^{vs}, \bar{P}_s, s_{P_S}\right)\right), \tag{30}$$

a weighted sum of monthly growth function responses to temperature and precipitation where the weights χ_s are the monthly average length of daylight scaled to the unit interval $(0, 1)$. Thus, χ_s scales growth to the known average amount of sunlight in a month, mimicking the physiology of tree growth. The monthly-scale function g ($\tilde{g}$) downscales the marginal annual climate anomaly to a monthly value and "grows" a monthly tree ring increment given the marginal climate. These marginal monthly growth functions then are combined by taking the minimum, allowing each month's growth to be either temperature or precipitation sensitive. Hence the tree ring growth model follows the "principle of limiting factors," which states that tree growth is constrained by the climatic variable that is limiting (Fritts, 1976). In the next sections, we describe the two forms of g and $\tilde{g}$ used in our model, which we call "VS-Lite," representing the growth model form used in Tolwinski-Ward et al. (2014), and "probit."

Biologically, all tree species in the tree ring network have a similar response to climate. For example, all tree species in the network have more similar climatological needs than, say, a tropical tree, which requires an entirely different climate. One explanation for what allows many tree species to grow in the same region are that different species occur in niche deviations from the overall mean response to climate. Hence, recent research in climate reconstruction methods demonstrates that inclusion of multiple tree species improves climate reconstruction skill (García-Suárez et al., 2009; Cook and Pederson, 2011). Allowing for species-specific climate responses ameliorates the difficulties of inverting the multivalued functional relationship between climate and tree ring widths by placing multiple constraints on the set of possible climate scenarios. Using multiple species is beneficial because it provides more constraints on climate and may allow for more precise estimation than in Tolwinski-Ward et al. (2014). However, separate growth parameters for each species increases the number of model parameters to be estimated, therefore we borrow strength by modeling the growth parameters hierarchically to improve parameter estimation and predictive skill (Gelman and Hill 2006, Chapter 12.2; Hooten and Hobbs 2015; Hobbs and Hooten 2015, Chapter 6.2). By treating each tree species' response to climate as a random draw from a pooled distribution, the model shares information among tree species. This borrowing of strength among species is easily incorporated into the probit growth model framework, but is not straightforward in the VS-Lite framework.

3.2.3.1. *VS-Lite tree ring growth model.* The "VS-Lite" model represents a statistical approximation to the Vaganov-Shashkin-Lite model that has been shown to create reasonable tree ring width chronologies given climate (Shashkin and Vaganov, 1993; Vaganov et al., 2006; Tolwinski-Ward et al., 2011). The VS-Lite tree ring growth model uses the linear

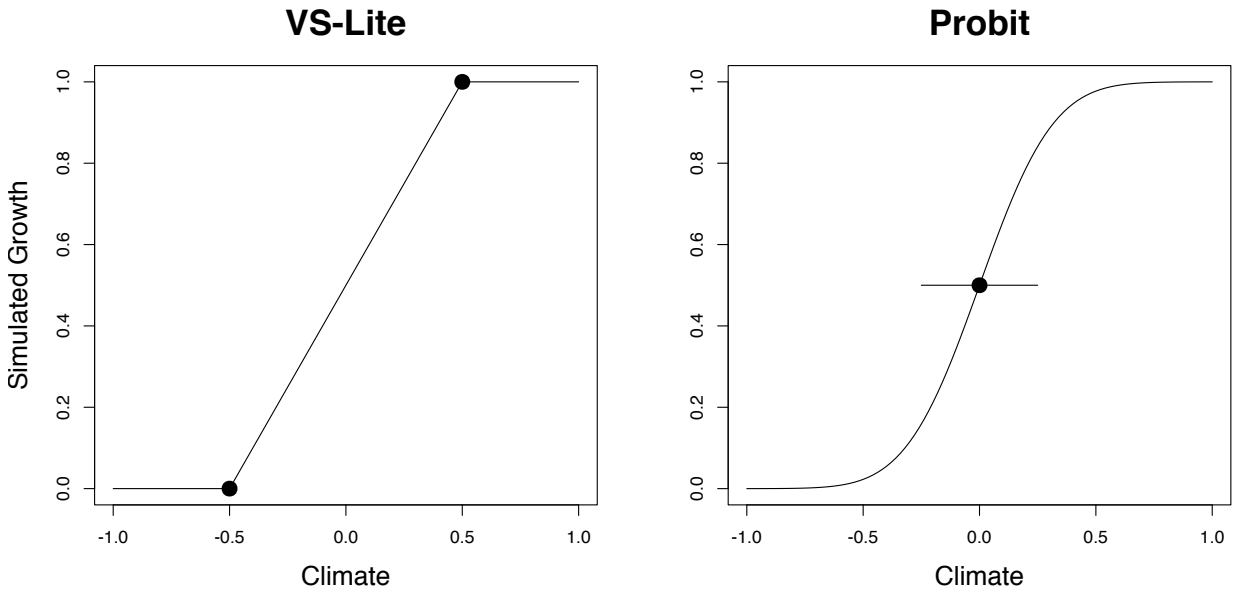


FIGURE 3.1. Example VS-Lite and probit ramp functions. The black dots on the VS-Lite plot represent the locations of T_{min_j} and T_{max_j}, the temperatures below which growth is zero or above which growth is optimal (equivalently P_{min_j} and P_{max_j} for precipitation). The black dot on the probit plot represents the probit mean growth response to temperature μ_{T_j} (μ_{P_j} for precipitation) and the line shows the probit standard deviation of growth response to temperature σ_{T_j} (σ_{P_j} for precipitation).

ramp function

$$\Psi(\eta) = \begin{cases} 0 & \text{if } \eta \leq 0 \\ \eta & \text{if } 0 < \eta < 1 \\ 1 & \text{if } \eta \geq 1, \end{cases}$$

as a link between climate and tree ring width growth shown in Figure 3.1. The monthly tree ring growth functions for temperature and precipitation using VS-Lite are

$$g\left(w_{T_{ts}}, \boldsymbol{\theta}_j^{VS}, \bar{T}_s, s_{T_s}\right) = \psi\left(\frac{w_{T_{ts}} s_{T_s} + \bar{T}_s - T_{min_j}}{T_{max_j} - T_{min_j}}\right) \tag{31}$$

$$g\left(w_{P_{ts}}, \boldsymbol{\theta}_j^{VS}, \bar{P}_s, s_{P_s}\right) = \psi\left(\frac{\exp\left\{w_{P_{ts}} s_{P_s} + \bar{P}_s\right\} - P_{min_j}}{P_{max_j} - P_{min_j}}\right), \tag{32}$$

and are combined according to the principle of limiting factors using (30). The growth parameters for the VS-Lite tree ring growth model are $\boldsymbol{\theta}_j^{VS} \equiv \left(T_{min_j}, T_{max_j}, P_{min_j}, P_{max_j}\right)'$ for each species $j = 1, \ldots J$. For temperatures below T_{min_j}, there is no tree ring growth. For monthly temperatures between T_{min_j} and T_{max_j}, tree ring growth increases linearly over the unit interval $(0, 1)$. When the monthly temperature exceeds T_{max_j}, growth occurs at a maximum rate, taking a value of 1. The interpretation for P_{min_j} and P_{max_j} is similar.

To complete the VS-Lite growth model parameterization, we specify priors for the growth parameters $\boldsymbol{\theta}_j^{VS}$ using a four parameter Beta(α, β, L, U) distribution, a Beta(α, β) distribution that has been shifted and scaled to have lower endpoint L and upper endpoint U. These prior models require expert knowledge to specify, hence we follow the recommendations of Tolwinski-Ward et al. (2013), recognizing that the previous work modeled soil moisture instead of precipitation under a different climate scenario than the Hudson Valley. We chose informative priors that place a majority of the probability mass in the center of the support that result in growth model priors that are reasonable given the climate of the Hudson Valley. For $j = 1, \ldots, J$, the VS-Lite parameter priors are $T_{min_j} \sim$ Beta(9, 5, 0, 9), $T_{max_j} \sim$ Beta(3.5, 3.5, 10, 24), $P_{min_j} \sim$ Beta(3.5, 3.5, 65, 85), and $P_{max_j} \sim$ Beta(3.5, 3.5, 85, 105). Despite existing knowledge of tree response to climate, the VS-Lite tree ring growth model could be sensitive to prior specification, especially if the true growth parameter values lie outside the range of prior support. In this case, the posterior probability of correctly estimating the true parameter value is exactly zero. Therefore, climate reconstruction using the VS-Lite growth model formulation has the potential to be highly influenced by misspecification of the prior support.

3.2.3.2. *Probit tree ring growth model.* An alternative to the VS-Lite tree ring growth model is the probit growth model, which was not examined in Tolwinski-Ward et al. (2014).

The probit growth model replaces the linear function $\psi(\eta)$ in the VS-Lite growth model by the infinitely differentiable inverse normal cumulative distribution function $\Phi^{-1}(\eta)$. The probit growth model parameters have infinite support and are parameterized as $\boldsymbol{\theta}_j^{Pro} \equiv \left(\mu_{T_j}, \sigma_{T_j}^2, \mu_{P_j}, \sigma_{P_j}^2\right)'$. There are only slight differences in the shape of the VS-Lite and probit growth model functions, as seen in Figure 3.1. The probit ramp function produces a smoother response to climate than the VS-Lite ramp function, but other than smoothness, the two shapes are quite similar, hence it seems likely that the shape of the growth function alone will not significantly improve predictive performance. The motivation for the probit growth function is that our model framework takes advantage of statistical properties not available in the VS-Lite framework. First, the prior support for the probit growth model is the real line, in comparison to the VS-Lite prior support that is restricted to compact support. In practice, if the true growth model in the VS-Lite framework is not in the range of prior support, the posterior probability of correctly estimating this parameter is exactly zero, regardless of the amount and quality of data. The probit growth function does not suffer from this problem. Additionally, the probit model can be easily extended to a hierarchical pooling framework. We propose the probit growth model form to evaluate the influences of these desirable statistical probabilities on the climate reconstruction. If the predictive performance of the two models is equivalent, the probit model will be preferred due to these properties.

The monthly probit growth increments due to temperature and precipitation are

$$\tilde{g}\left(w_{T_{ts}}, \boldsymbol{\theta}_j^{Pro}, \bar{T}_s, s_{T_s}\right) = \Phi^{-1}\left(\frac{w_{T_{ts}} s_{Ts} + \bar{T}_s - \mu_{Tj}}{\sigma_{Tj}}\right) \tag{33}$$

$$\tilde{g}\left(w_{P_{ts}}, \boldsymbol{\theta}_j^{Pro}, \bar{P}_s, s_{P_s}\right) = \Phi^{-1}\left(\frac{\exp\left\{w_{P_{ts}} s_{Ps} + \bar{P}_s\right\} - \mu_{Pj}}{\sigma_{Pj}}\right). \tag{34}$$

where the parameters μ_{T_j} and μ_{P_j} represent the species-specific temperature and precipitation probit mean growth rate, and the parameters σ_{T_j} and σ_{P_j} control the effective range where tree growth responds to climate, the probit standard deviation. A species with a higher value of μ_{T_j} will experience better growth under warmer weather than a species with a lower value of μ_{T_j} and a species with a larger σ_{T_j} will have a larger range of temperatures in which that tree species will grow than a species with a smaller σ_{T_j}. The interpretation of these growth model parameters is similar for precipitation.

For $j = 1, \ldots, J$, we specify the probit growth model parameter distributions for the species as: $\mu_{T_j} \sim \mathrm{N}\left(\mu_{\mu_T}, \sigma^2_{\mu_T}\right)$, $\sigma_{T_j} \sim \mathrm{logN}\left(\mu_{\sigma_T}, \sigma^2_{\sigma_T}\right)$, $\mu_{P_j} \sim \mathrm{N}\left(\mu_{\mu_P}, \sigma^2_{\mu_P}\right)$, and $\sigma_{P_j} \sim \mathrm{logN}\left(\mu_{\sigma_P}, \sigma^2_{\sigma_P}\right)$. The pooling of these effects occurs by adding one more level in the hierarchical model by defining a hyperprior model for each of the prior parameters above with $\mu_{\mu_T} \sim \mathrm{N}(13, 4)$, $\sigma^2_{\mu_T} \sim \mathrm{IG}(2, 0.5)$, $\mu_{\sigma_T} \sim \mathrm{N}(0, 1)$, $\sigma^2_{\sigma_T} \sim \mathrm{IG}(2, 0.5)$, $\mu_{\mu_P} \sim \mathrm{N}(85, 16)$, $\sigma^2_{\mu_P} \sim \mathrm{IG}(2, 0.5)$, $\mu_{\sigma_P} \sim \mathrm{N}(0, 2)$, $\sigma^2_{\sigma_P} \sim \mathrm{IG}(2, 0.5)$. These prior values represent likely values that cover the range of growing season temperature and precipitation values seen in the Hudson Valley while being highly flexible, thus allowing the model to estimate the growth parameters more flexibly than the VS-Lite growth model.

3.2.4. Dynamic multi-scale climate process. We model temperature and log precipitation anomalies jointly with a dynamic, multi-scale model, allowing prediction of unobserved temperature and precipitation when combined with the tree ring chronology data. The model is a temporal vector autoregressive process among years, given a propagator matrix $\boldsymbol{A}$, and a correlated autoregressive process among months determined by the structure of a covariance matrix $\boldsymbol{\Sigma}$. To account for trend in the temperature anomaly time series during the observational period 1895-2010, we include an intercept Δ_0 and slope Δ_1 in the model for years after $t^* = 1895$. In years before t^*, we do not have observational data and

thus we do not model a trend. Using these assumptions, we model the dynamic de-trended process

(35)
$$\begin{pmatrix} \boldsymbol{w}_{Tt} - \Delta_0 \boldsymbol{J} - (t - t^*)\Delta_1 \boldsymbol{J} \\ \boldsymbol{w}_{Pt} \end{pmatrix} \sim \mathrm{N}\left(\boldsymbol{A} \begin{pmatrix} \boldsymbol{w}_{Tt-1} - \Delta_0 \boldsymbol{J} - (t - t^*)\Delta_1 \boldsymbol{J} \\ \boldsymbol{w}_{Pt-1} \end{pmatrix}, \boldsymbol{\Sigma} \right) \quad \text{if } t \geq t^*,$$

(36)
$$\begin{pmatrix} \boldsymbol{w}_{Tt} - \Delta_0 \boldsymbol{J} \\ \boldsymbol{w}_{Pt} \end{pmatrix} \sim \mathrm{N}\left(\boldsymbol{A} \begin{pmatrix} \boldsymbol{w}_{Tt-1} - \Delta_0 \boldsymbol{J} \\ \boldsymbol{w}_{Pt-1} \end{pmatrix}, \boldsymbol{\Sigma} \right) \quad \text{if } t < t^*.$$

The propagator matrix $\boldsymbol{A} = \begin{pmatrix} \phi_1, 0 \\ 0, \phi_2 \end{pmatrix} \otimes \boldsymbol{I}$ defines the annual scale autocorrelation for the temperature and log precipitation anomalies, where ϕ_1 and ϕ_2 are the annual autocorrelation parameters. In these expressions, $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{J}$ is a 12×1 vector of ones. We model the inter-annual covariance $\boldsymbol{\Sigma}$ with a temporal multivariate conditionally autoregressive (MCAR) structure, a generalization of the conditionally autoregressive (CAR) structure in time that allows for the within year temperature and precipitation anomalies to have their own temporal autocorrelation parameters while also including a temporally explicit cross-correlation between the anomaly measurements (Mardia, 1988; Carlin and Banerjee, 2003; Gelfand and Vounatsou, 2003; Jin et al., 2005). We found the MCAR model to be the best predicting among a set of candidate models for temporal autocorrelation.

To construct the temporal MCAR covariance matrix $\boldsymbol{\Sigma}$, we first define a temporal CAR precision matrix that will be used as component in building $\boldsymbol{\Sigma}$. The particular precision matrix we use, $\boldsymbol{Q}(\omega)$, specifies a process identical to an autoregressive process of order one in the time series literature if $|\omega| < 1$ (Cressie and Wikle, 2011, p. 170). We define the monthly temperature CAR precision matrix $\boldsymbol{Q}(\omega_T)$ with autocorrelation parameter ω_T, and

the monthly log precipitation CAR precision matrix $\boldsymbol{Q}(\omega_P)$ with autocorrelation parameter ω_P allowing each climate variable to have its own monthly autocorrelation. We decompose $\boldsymbol{Q}(\omega_T) = \boldsymbol{L}_T\boldsymbol{L}_T'$ and $\boldsymbol{Q}(\omega_P) = \boldsymbol{L}_P\boldsymbol{L}_P'$ with a Cholesky decomposition (one could also use a spectral decomposition) and construct the MCAR precision matrix as

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_w^2}\begin{pmatrix} \mathbf{L}_T' & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_P' \end{pmatrix}(\boldsymbol{\Lambda} \otimes \mathbf{I}_{12})\begin{pmatrix} \mathbf{L}_T & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_P \end{pmatrix} \tag{37}$$

where the matrix $\boldsymbol{\Lambda} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ is a correlation matrix with ρ representing the atemporal cross-correlation between temperature and precipitation, σ_ω^2 is a global variance parameter, and $\otimes$ represents the Kronecker product. Thus using (37), we model intra-annual autocorrelation, monthly autocorrelation, and cross-correlation in the climate process.

To significantly reduce computation time, we employ an empirical Bayes approach to process the climate data (Casella, 1985). We estimate the propagator matrix $\boldsymbol{A}$, the intra-annual covariance matrix $\boldsymbol{\Sigma}$, and the trend parameters Δ_0 and Δ_1 offline using a hybrid Metropolis-Hastings and Gibbs MCMC sampling algorithm using uniform$(-1, 1)$ priors for the parameters ρ, ω_T, and ω_P and an Inverse Gamma(1, 1) prior for the variance σ_w^2. After fitting the model, we use posterior median values of each parameter as estimates of the true processes parameters, thus our posterior predictions do not include climate model parameter uncertainty and our corresponding credible intervals will be overly optimistic.

3.2.5. Posterior. We desire inference on the posterior distribution and quantities derived from the posterior distribution. The posterior we seek to approximate with our Markov

Chain Monte Carlo (MCMC) algorithm is

$$\begin{aligned}
&\left[\boldsymbol{w}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\theta}^{VS}, \boldsymbol{\theta}^{Pro}, \sigma^2, \boldsymbol{z} \middle| \boldsymbol{y}, \boldsymbol{T}, \boldsymbol{P}\right] \\
&\qquad \propto \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{t=1}^{\tau} \left[y_{itj} \middle| \beta_{0_j}, \beta_{1_j}, \boldsymbol{w}_t, \boldsymbol{\theta}_j^{VS}, \sigma_j^2\right]^{(1-z_j)} \\
&\qquad \times \left[y_{itj} \middle| \tilde{\beta}_{0_j}, \tilde{\beta}_{1_j}, \boldsymbol{w}_t, \boldsymbol{\theta}_j^{Pro} \tilde{\sigma}_j^2\right]^{z_j} \\
&\qquad \times \left[\boldsymbol{w}_t | \boldsymbol{w}_{t-1}, \boldsymbol{A}, \boldsymbol{\Sigma}, \Delta_0, \Delta_1\right] \left[\boldsymbol{\beta_0}\right] \left[\boldsymbol{\beta_1}\right] \\
&\qquad \times \left[\boldsymbol{\theta}^{VS}\right] \left[\boldsymbol{\theta}^{Pro}\right] \left[\boldsymbol{\sigma}^2\right] \left[\boldsymbol{z}\right],
\end{aligned}$$

where the parameter models for the VS-Lite and probit growth model are represented as $\left[\boldsymbol{\theta}^{VS}\right]$ and $\left[\boldsymbol{\theta}^{Pro}\right]$, respectively. Implementation of a hybrid Metropolis-Hastings and Gibbs MCMC algorithm allows for estimation of the posterior distribution (Banerjee et al., 2004; Carlin and Louis, 2011). Our model was implemented using R (R Core Team, 2015), while leveraging significant portions of C++ code using RcppArmadillo (Eddelbuettel and Sanderson, 2014) to increase computation speed. The MCMC algorithm for each candidate model was run for 15,000 iterations with the first half discarded as burn-in and thinning every 5 observations, for three parallel chains, resulting in 4,500 posterior samples for a total computation time of one hour on a dual core 2.6 GHz laptop with 8GB memory. Convergence was assessed using Gelman-Rubin's $\hat{R}$ statistic (Gelman and Rubin, 1992) and visual inspection of the trace plots.

3.3. Model Evaluation

Traditional paleoclimate reconstructions evaluate predictive performance with out-of-sample data using the coefficient of efficiency (CE) (Cook et al., 1994; Rutherford et al., 2005; Tingley and Huybers, 2010a,b). Despite the accepted use of this scoring statistic, Gneiting

and Raftery (2007) suggest that skill scores like CE are improper in general. Impropriety implies that it is possible to have predictions that, under expectation, have better CE skill scores than a model that is optimal. Therefore, use of improper scoring rules can lead to incorrect inference about predictive skill among a set of predictive models. Thus, we use a proper scoring rule. Proper scoring rules guarantee that, under expectation, the optimal predictive model will have the best predictive score (Gneiting et al., 2007; Gneiting, 2011; Hooten and Hobbs, 2015).

Our model produces a probabilistic forecast, hence, we use the continuous ranked probability score (CRPS), a proper scoring rule that accommodates both probabilistic and point forecasts. Several recent papers on late Holocene climate reconstructions have made use of the CRPS (Barboza et al., 2014; Werner and Tingley, 2015). Given a forecast with cumulative distribution function, F_t, at time t and out-of-sample observations $\boldsymbol{y}_{oos}$, the CRPS is defined as

$$CRPS(\{F_t\}_{t=1}^{\tau}, \boldsymbol{y}_{oos}) = -\sum_{t=1}^{\tau} \int_{-\infty}^{\infty} \left(F_t(y) - \boldsymbol{I}_{\{y \geq y_{oos,t}\}}\right)^2 dy. \tag{38}$$

Gneiting et al. (2007) show how (38) can be written alternatively as

$$CRPS(\{F_t\}_{t=1}^{\tau}, \boldsymbol{y}_{oos}) = \sum_{t=1}^{\tau} \left(E_{F_t} |y_t - y_{oos,t}| - \frac{1}{2} E_{F_t} |y_t - y_t'| \right), \tag{39}$$

where y_t and y_t' are independent copies of a linear random variable with distribution function F_t and the expectation E is with respect to the probability density induced by F_t. The first expectation in the above equation measures calibration, the absolute error of the prediction relative to the out-of-sample value and the second expectation rewards predictions that are

sharp (i.e., narrow prediction intervals). Hence, the CRPS rewards probabilistic predictions that are accurate and precise.

In a Bayesian context, the CRPS can be estimated after obtaining posterior samples. First, sample $\tilde{\boldsymbol{y}}^{(k)}$ from the posterior predictive distribution $\left[\tilde{\boldsymbol{y}}^{(k)} \middle| \boldsymbol{y}, \boldsymbol{\theta}^{(k)}\right]$ at each post burn-in iteration k. Then, the expression in (39) is approximated by

$$\widehat{CRPS}(\{\hat{F}_t\}_{t=1}^{\tau}, \boldsymbol{y}_{oos}) = \sum_{t=1}^{\tau} \left(\frac{1}{K} \sum_{k=1}^{K} \left| \tilde{y}_t^{(k)} - y_{oos,t} \right| - \frac{1}{2K^2} \sum_{k=1}^{K} \sum_{\ell=1}^{K} \left| \tilde{y}_t^{(k)} - \tilde{y}_t^{(\ell)} \right| \right). \tag{40}$$

We use the CRPS score in the simulation study to select the best model based on predictive performance. The CRPS is a negatively oriented scoring rule, therefore the model with the lowest $\widehat{CRPS}$ is the best scoring model, and under expectation, the best predicting model.

3.4. Simulation Study

We consider three variants of the model presented above to conduct a reconstruction experiment. First, we consider the VS-Lite tree ring growth model, modifying the work of Tolwinski-Ward et al. (2014) to be consistent with our modeling framework using our climate model and multi-species approach. Second, we use only the probit tree ring growth model. Lastly, we use the mixture model described in (29) that allows for choice of tree ring growth model. We evaluate predictive performance of these candidate models over nine total simulation scenarios, using each of the tree ring growth models to simulate pseudoproxy data and fitting each growth model to each of the three simulated datasets.

We simulate the data as follows. First, we estimate $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ from the instrumental climate data using (35). Using these estimates, we simulate a realization of the climate process with 446 years of pre-instrumental simulated climate variables while adding a trend with slope of $\frac{1}{110}$ to the 110 years of temperature observations in the observation period,

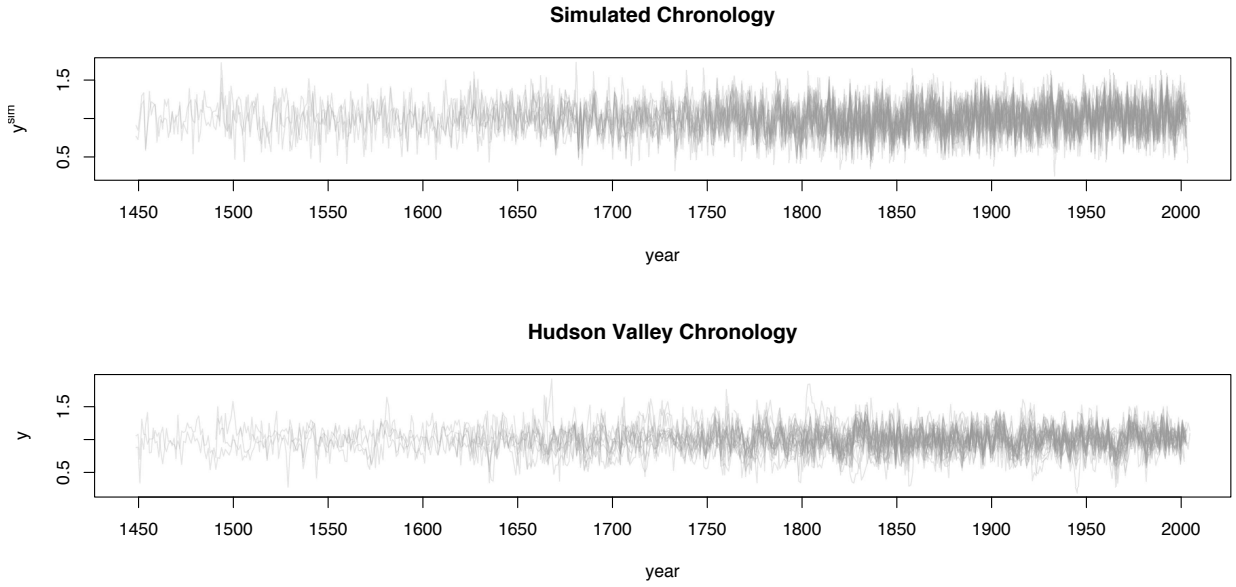


FIGURE 3.2. Simulated and observed tree ring width chronology for Hudson Valley

producing a current day increase of one degree Celsius. For the VS-Lite tree ring growth model simulation we sample the growth parameters from uniform distributions. For each species j, $T_{min_j} \sim \mathrm{U}(0, 9)$, $T_{max_j} \sim \mathrm{U}(10, 24)$, $P_{min_j} \sim \mathrm{U}(65, 85)$, and $P_{max_j} \sim \mathrm{U}(85, 105)$. For the probit tree ring growth model we sample the growth parameters $\mu_{T_j} \sim \mathrm{N}(16, 4)$, $\sigma_{T_j} \sim \log \mathrm{N}(\log(1), 0.5)$, $\mu_{P_j} \sim \mathrm{N}(85, 16)$, and $\sigma_{P_j} \sim \log \mathrm{N}(\log(2), 1)$ for each species $j = 1, \ldots, J$.

Next, we use each tree ring growth model to produce a noiseless tree ring chronology, standardizing each chronology to have mean 1 and standard deviation of 0.2, as in the Hudson Valley dataset. Adding in noise representing measurement and processing error, we simulate observed chronologies

$$y_{tj}^{sim} \sim \mathrm{N}\left(\sqrt{(1 - \sigma_{noise}^2)} f(\boldsymbol{w}_t, \boldsymbol{\theta}_j^{sim}), \sigma_{noise}^2\right). \tag{41}$$

The parameter σ^2_{noise} controls the signal to noise ratio in the simulated tree ring chronology, values of σ^2_{noise} near zero represent a high signal to noise ratio while values near one represent a low signal to noise ratio. We let $\sigma^2_{noise} = 0.75$, representing a signal to noise ratio of 0.58, which is at the high end of what is realistic for many tree ring chronologies (Smerdon, 2012). Figure 3.2 shows a realization of the simulated chronology and the observed chronology from Hudson Valley, demonstrating that our simulation methodology produces realistic tree ring chronologies. We apply the same structure of missingness to our simulated data as in the observed chronology, producing simulations as close as possible to the observed data.

TABLE 3.2. Table of CRPS scores for annual temperature, growing season temperature, annual precipitation, and growing season precipitation for data simulated with the VS-Lite growth model.

$\widehat{CRPS}$	Climatology	VS-lite	probit	mixture
Annual temperature	531.29	**455.71**	455.75	455.98
Growing season temperature	486.52	**462.71**	463.93	464.45
Annual log precipitation	98.53	85.20	85.16	**84.86**
Growing season log precipitation	130.92	91.65	92.20	**90.63**

TABLE 3.3. Table of CRPS scores for annual temperature, growing season temperature, annual precipitation, and growing season precipitation for data simulated with the probit growth model.

$\widehat{CRPS}$	Climatology	VS-lite	probit	mixture
Annual temperature	571.92	446.90	446.99	**446.72**
Growing season temperature	482.51	**450.02**	451.15	451.04
Annual log precipitation	101.06	90.61	**90.36**	90.69
Growing season log precipitation	129.34	**93.25**	93.59	93.61

TABLE 3.4. Table of CRPS scores for annual temperature, growing season temperature, annual precipitation, and growing season precipitation for data simulated with the mixture growth model.

$\widehat{CRPS}$	Climatology	VS-lite	probit	mixture
Annual temperature	540.23	438.65	**438.33**	438.80
Growing season temperature	489.36	440.66	441.17	**440.60**
Annual log precipitation	97.82	85.28	85.40	**84.49**
Growing season log precipitation	126.33	90.51	91.02	**90.04**

The model characteristic most important to us in climate reconstruction is predictive ability. The reconstruction temperature and precipitation $\widehat{CRPS}$ values are shown in Table 3.2 for data simulated with the VS-Lite growth model, Table 3.3 for data simulated with the probit growth model, and Table 3.4 for data simulated with the mixture growth model. For each of the simulated datasets $\widehat{CRPS}$ was estimated for both an annual and growing period reconstruction, with bold scores highlighting the model that performs best for each simulated dataset and time period. As a baseline comparison, the $\widehat{CRPS}$ scores for a climatological prediction are included. The $\widehat{CRPS}$ values suggest all three models are similar in predictive ability, although there might be a slight preference for the mixture growth model. Hence, we discuss the mixture growth model in what follows. Although the probit growth model is often not the best scoring model, the mixture growth model indicator variable, $\boldsymbol{z}$, suggests an even split between the VS-lite and probit tree ring growth model (the probit models is selected about 53% of the time) within the MCMC chain. An example reconstruction using the mixture tree ring growth model on the simulated data, Figure 3.3, demonstrates that the reconstruction performs quite well for log precipitation, but is rather uninformative for temperature. Note that, during approximately 1650-1700, the uncertainty intervals for log precipitation reconstruction increase in width and predictive skill decreases due to increasing numbers of missing chronologies as well as the amount of replication within a given chronology decreasing. This uncertainty provides information about when the reconstruction is performing well and over what time periods the predictions are no better than climatology without relying on calibration skill measures like CE.

In addition to producing accurate climate reconstructions, all growth model types recover the simulated growth model parameters when the model is applied to data simulated

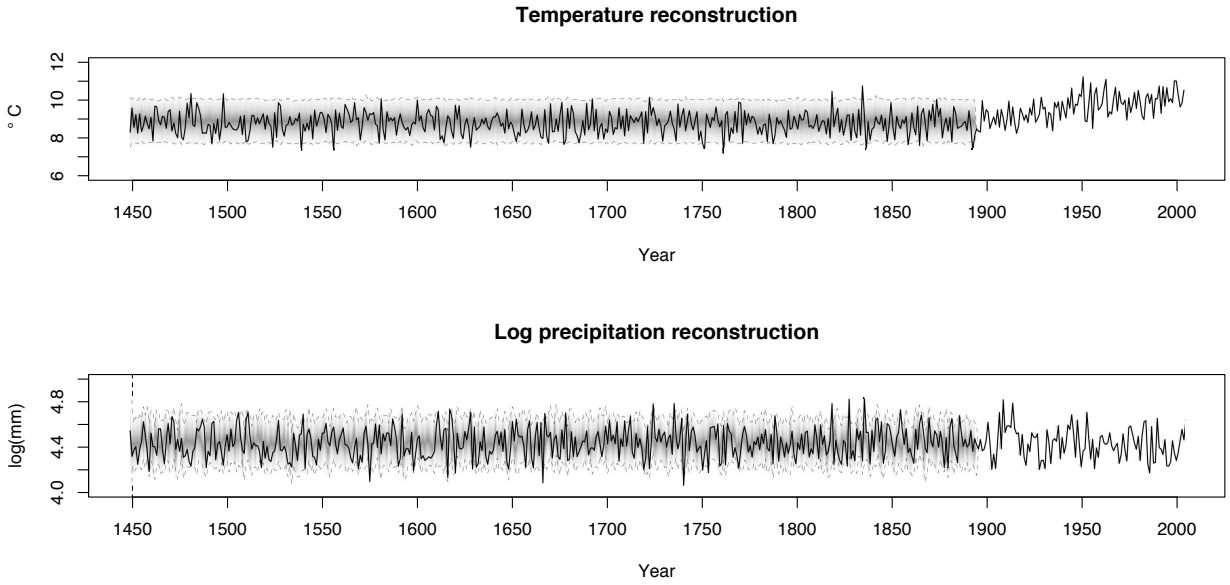


FIGURE 3.3. Reconstruction of Temperature and Log Precipitation from data simulated with the mixture growth model and fit with the mixture growth model. The gray shading is proportional to the posterior predictive density, the dotted gray lines show the 95% posterior predictive credible interval, and the black line is the simulation truth.

with the same structure, showing that all of the models are capable of estimating the simulated growth model parameters. This result validates the growth model's usefulness for investigation of climatological niches for different tree species as well as providing accurate predictions of climate. Yet, despite accurately estimating the temperature growth parameters, the temperature reconstruction is not much more informative than a climatological prediction. Hence, the simulation suggests the failure to predict temperature patterns is not due to poor estimation of the simulated growth parameters. To further illuminate why the model is unable to reconstruct temperature, Figure 3.4 shows the observed monthly temperature and precipitation patterns with the probit growth model parameters shaded showing the range over which the model is sensitive to climate. We obtain very little learning about temperature at an annual scale, as the model is sensitive to temperature for, at best, two to four months of the year and only in the spring/fall with no learning about temperature during

the winter (when temperature is most variable) and summer months. In fact, the effective learning about temperature occurs in much less than four months due to the multiplicative interaction between temperature and precipitation in (30). Therefore, any learning about temperature from the growth model is dominated by monthly variability in temperature. In contrast, the reconstruction of precipitation is quite accurate, especially in the near past, because the model is sensitive to precipitation values for all months of the year, illustrating that the change of support from monthly to annual scale causes few problems for precipitation. An investigation of climate scenarios with more significant overlap of the climate and tree growth sensitivities in simulations (not shown in these results) demonstrates potential for reconstruction of both temperature and precipitation patterns, presenting opportunities for further applications to other datasets.

3.5. Hudson Valley Results

Based on the predictive results in the simulation study, we applied the mixture tree ring growth model to the Hudson Valley data. Using our simulations and expert priors, the most obvious characteristic of the reconstruction is that we obtain only minimal learning about temperature, while the precipitation reconstruction performs well. To validate our predictive model, we translate and scale the reconstruction of Palmer Drought Severity Index (PDSI) in Pederson et al. (2013), a reconstruction that uses the same Hudson Valley dataset but very different methods, and compare this result to our reconstruction of log precipitation. Figure 3.5 shows that our log precipitation reconstruction strongly correlates ($r = 0.74$) with the previous PDSI reconstruction effort for the near past, while having the added benefit of explicitly accounting for uncertainty. It is reasonable that PDSI would be correlated with log precipitation because PDSI is a measure of drought severity and, because there is little

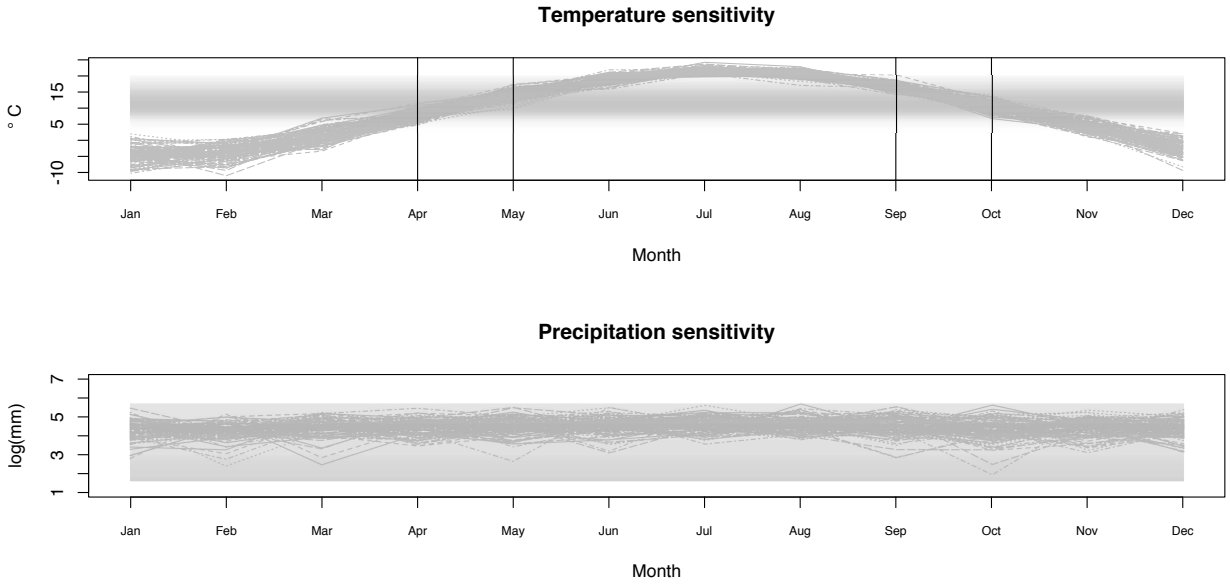


FIGURE 3.4. Reconstruction of annual temperature depends on the growth parameters. The shaded area shows where the growth parameters are sensitive to climate and the gray lines are observed climate. The vertical lines in the temperature plot show that annual reconstruction is only dependent on four or fewer months. Thus, annual scale information about temperature is hard to recover from only a few months while precipitation information can be extracted across all months.

inter-annual variability in temperature, the primary driver of drought (and tree growth) in the Hudson Valley is the amount of precipitation (Martin-Benito and Pederson, 2015).

The uncertainty estimates for the reconstruction of log precipitation provide insight for the prediction's reliability. As we backcast in time, the 95% credible interval size approaches the size of predictive intervals derived using the observational data only, suggesting that the reconstruction lacks predictive ability before approximately 1650-1700. This loss of predictive ability, as seen in the simulation study, occurs during a time period where many tree ring chronologies are lacking enough replication to be included in the ring network. As such, deviations between our log precipitation reconstruction and the previous reconstruction of PDSI in Figure 3.5 before 1650-1700 can be explained in the context of prediction uncertainty.

In addition to the climate reconstructions, it is possible to obtain inference about the climatological niches different tree species occupy from the growth parameters. Many tree species in the Hudson Valley dataset are represented in only one or two chronologies, therefore, inference is limited due to the relatively small sample size and associated overlap in uncertainty intervals. For the VS-lite growth model, the posterior samples show little ability to discern different ecological niches, but the posterior samples from the probit growth model posterior show many species have statistically different mean response to climate, demonstrating that the probit growth model can be easily adapted for ecological learning in a richer dataset where climatological prediction is not of interest. Also important is the posterior distribution of the indicator of model importance $\boldsymbol{z}$. We find a slight preference for the probit growth model using the Hudson Valley data, with a posterior probability $P(z = 1) = 0.54$. This is not substantially different than the prior probability of 0.5, which is expected because the two growth functions have very similar shapes. The slight preference for the probit growth occurs because of improved parameter estimation through the use of a hierarchically pooled prior model.

3.6. Discussion

We presented a methodological framework for reconstructing paleoclimate using a mechanistic Bayesian hierarchical model motivated by the work of Tolwinski-Ward et al. (2014). We proposed a novel probit tree ring growth model that takes advantage of the biology of tree growth and pools tree growth parameters to improve estimation and decrease sensitivity to prior specification. This new growth function was proposed in a framework that rigorously evaluates the growth model influence. Our extension to multispecies modeling of tree response to climate constrained predictive backcasts to a climate scenario consistent with

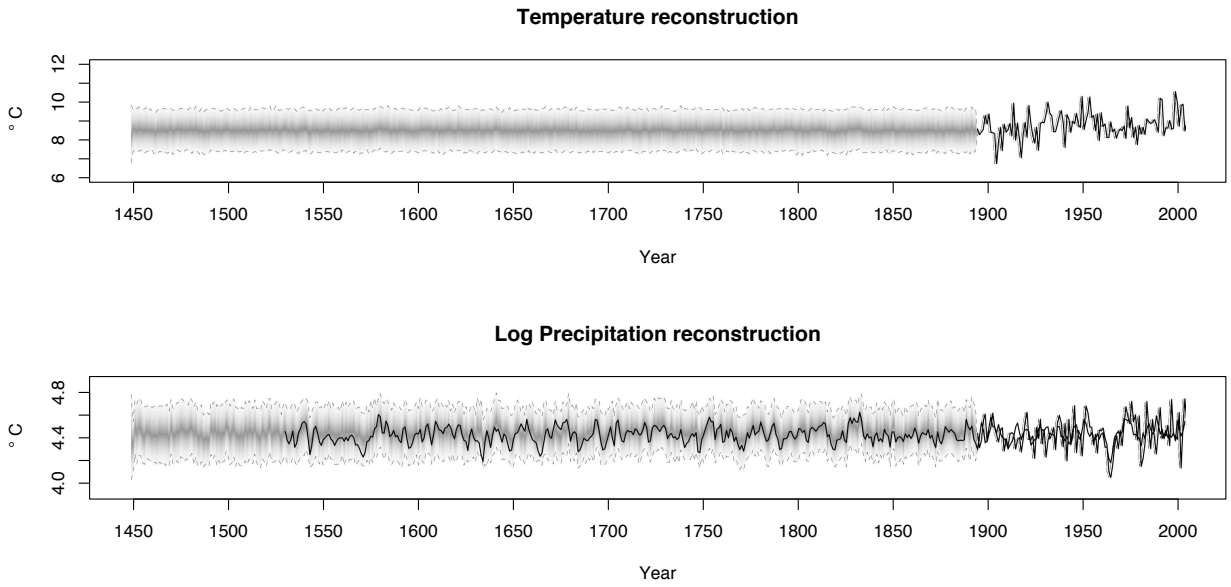


FIGURE 3.5. Climate reconstruction from the Hudson Valley chronology. The gray shading is proportional to posterior predictive density, the dotted lines show the 95% credible intervals and the solid black line in the precipitation plot is a translated and scaled reconstruction of drought (PDSI) from Pederson et al. (2013).

the data, thereby improving predictive skill over models that do not explicitly incorporate multiple species. The introduction of a calibration model reduced computation time and improved MCMC convergence over previous standardization methods. We developed an upscaling and downscaling model to align the different data sources on the same scale and proposed a novel dynamic joint process model for temperature and precipitation accounting for temporal correlation and cross-correlation. Our pseudoproxy simulation experiment evaluated predictions with a proper scoring rule that uses the full probabilistic forecast. The simulation study provided insight about model performance and provided feedback that can be used to interpret results from the Hudson Valley data.

By using a biologically motivated Bayesian hierarchical model for reconstructing climate processes from tree rings widths, we combine statistical techniques with scientific models developed in dendrochronology. Our modeling framework explicitly accounts for uncertainties,

in comparison to many previous climate reconstructions that use linear statistical methods and rely on asymptotic assumptions or bootstrap algorithms to estimate uncertainties, as discussed in Tingley and Huybers (2010b). An added benefit to using a biologically motivated model is the ability to adapt this modeling framework to make inference about ecological niches that different tree species occupy from the growth model parameter estimates.

By outlining the model shortcomings in detail, we gain a better understanding of the problem of multivariate climate reconstructions and can develop new ideas for improvement. Tree ring growth models are simplifications of the true biological process and more scientific realism can be added. For instance, at very warm temperatures, tree growth is inhibited due to water lost by transpiration. Including more accurate biology in the growth model will increase the number of months that are sensitive to the growth process and alleviates the change of support problem for the prediction of temperature. Another option to improve temperature predictions is to modify the limiting factor approach in (30) to allow more months to contribute to the reconstruction of temperature. Generalizing the multiscale climate process model will allow more complicated dynamics to be incorporated into the model, either through a statistical effort by allowing time-varying, non-stationarity processes, or by assimilating deterministic climate forcings in the mean of the reconstruction, as in Li et al. (2010) and Barboza et al. (2014). The introduction of a spatially explicit model could lead to spatially explicit climate field reconstructions, but would require a more computationally efficient method of fitting Bayesian models, like Hamiltonian Monte Carlo, variational Bayes, or integrated nested Laplace approximations. These and other potential improvements present opportunities for collaboration with dendrochronologists, climate scientists, ecologists, and statisticians to increase understanding of the relationship between climate and tree growth.

Recognition of Support

This material 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. We also thank Suz Tolwinski-Ward, Kristin Broms, and Ben Bird for their input on this manuscript.

CHAPTER 4

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

4.1. Introduction

There is a need for accurate estimates of paleoclimate to better understand the impacts of climate change, especially temperature and precipitation. Scientific measurements of temperature and precipitation have been recorded for the last few hundred years, and in many locations for a much shorter time. Historical weather data are often unreliable, sparse, and noisy because these data were recorded before widespread adoption of scientific measurement standards. As such, historical weather data are not widely used for rigorous statistical reconstructions of climate. To supplement the short record of direct human observations of temperature and precipitation, there are ongoing efforts to collect data that are indirect measurements of paleoclimate. These data, broadly classified as paleoclimate proxy data, consist of measurements of environmental, biological, geological, or other processes that incorporate climate information in an archive (Evans et al., 2013). Examples of paleoclimate proxy archives used for climate reconstruction include tree ring widths, ice cores, lake sediments, pollen, and compilations of historical records. After collection of these data, paleoclimate proxies are used for statistical learning about the relationships between the proxy data and climate, resulting in estimates of paleoclimate.

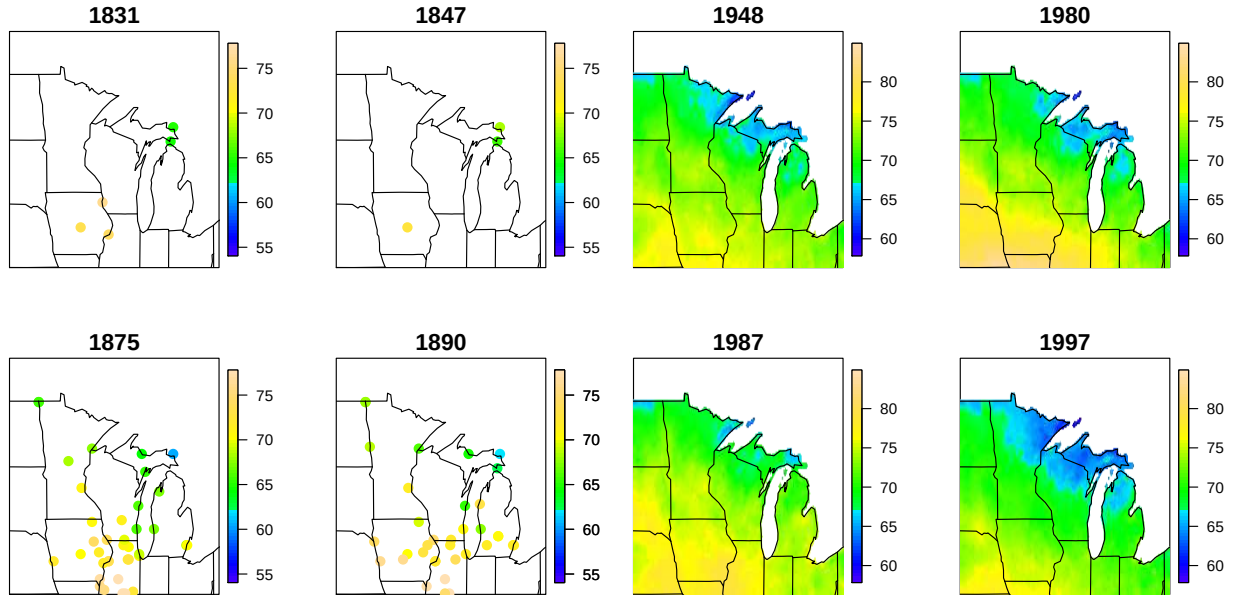
There are many challenges when modeling paleoclimate data. Each proxy archive has unique characteristics that present different modeling challenges. For example, uncertainty

in temporal resolution may vary among proxy archives. Tree ring widths are annually resolved (one tree ring is assumed to be one year) whereas the temporal resolution of lake sediments, created by landscape scale depositional processes that depend on precipitation, surrounding soil structure, fire events, and other processes that occur on varying timescales, present uncertainty in dating that must be accounted for (Blaauw and Christen, 2011; Paciorek and McLachlan, 2009; Blaauw and Christen, 2005). Proxy data can respond to slowly varying decadal or centennial time scales (pollen records or lake sediments), annual time scales (tree ring widths or annual coral growth), or daily time scales (compilations of historical measurements of temperature and precipitation). In addition, the observational data used to train statistical models can be hourly, daily, monthly, or even annual measurements of temperature and precipitation. Therefore, there is often a change of temporal support between the observed climate data and the paleoclimate proxy data. Thus, any statistical model must account for the change of support (Gotway and Young, 2002). Also, paleoclimate data are often sparse, contain many missing values, occur at irregular spatial and temporal locations, and are correlated in time. These challenges make it difficult to create generic statistical algorithms for analysis. Further complicating matters, the true target that one wishes to predict (the historical, unobserved climate) is never available to evaluate the predictive performance of our modeling efforts. To address the lack of data for predictive validation, skill scores to evaluate and rank model performance have been developed (Cook and Kairiukstis, 1990).

One source of paleoclimate data are historical records. Because humans have always been interested in weather, there are vast nonscientific records of weather. However, many reconstructions of paleoclimate using compiled historical records are not subject to direct statistical analysis because they often consist of less than precise measurements of weather

reported in letters, newspapers, books, and other documentary evidence (Bell and Ogilvie, 1978; Ogilvie, 1984; Kastellet et al., 1998; Brázdil et al., 2006). In addition, the data are often of unknown or varying reliability and are typically sparse, sometimes involving only a few locations per year. There is potential to extend our scientific understanding of climate back in time with the use of these historical weather records. Thus, there is a need for a statistical framework that can model historical data compiled from a variety of disparate sources by leveraging high quality climate data from the recent past.

For our reconstruction, we utilize two datasets we call the historical period dataset and the observational period dataset. The historical dataset consists of temperature observations from 1820-1893 from military records kept at United States forts in the Upper Midwestern United States. These data were compiled as part of the Climate Database Modernization Program (Andsager et al., 2004; Climate Database Modernization Program). At these forts, measurements were recorded with time and date, however, measures of time were often imprecise (e.g., "midday" or "early morning"). We focus on the month of July because tree growth in the study region has been shown to be highly influenced by July temperature, and the predictions from our work will be used for modeling forest stand dynamics in response to climate (Goring et al., 2013). As there was no standardized protocol for collection of these data, there are many data irregularities. July temperature measurements were obtained by a variety of methods: some records report daily minimum and maximum temperatures, others hourly measurements, and sometimes there are days or weeks without measurements. In addition, the number and locations of forts change through time, containing between two and 36 fort locations per year due to historical events including the Civil War and the westward expansion of the United States in the late 19^{th} century. Therefore, our model must align data sources through time and space, reconstructing temperature at approximately



(A) Historical period fort temperatures. (B) Observational period PRISM temperatures.

FIGURE 4.1. Plot of four representative years of the two data sources.

20,000 spatial locations. An example of four representative years of historical fort data is shown in Figure 4.1a.

Because the historical data are sparse, we cannot use traditional spatial statistical methods such as Kriging, because these methods require larger sample sizes to be computationally stable. Thus, we rely on an auxiliary dataset to facilitate our reconstruction. For the observational period data, we used the Parameter-elevation Relationships on Independent Slopes Model (PRISM) monthly average temperature surfaces created by interpolation of the United States Historical Climate Network data over the period 1895-2010 (PRISM Climate Group, Oregon State University). The PRISM data include 115 years of average July temperatures occurring over approximately 20,000 spatial locations of interest, as shown in Figure 4.1b. We use the observational data as a basis to learn about temperature in the historical period. Unlike the historical observations, the PRISM data are compiled from the United States Historical Climate Network (USHCN) and consist of high quality model interpolated temperature records.

To enable learning about climate in the historical period, we need to align the two data sources to common spatial and temporal scales. We align each fort location with the nearest grid cell in the PRISM data, thus accounting for any potential spatial misalignment. Aligning the data sources in time is more complicated because the fort data are highly irregular, whereas the PRISM data are monthly mean temperatures. We model the historical 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). The linear mixed model for daily July temperature is

$$y_{its} = \mathbf{B}(s)\boldsymbol{\upsilon}_{it} + \boldsymbol{\varepsilon}_{its}, \tag{42}$$

where y_{its} is the temperature observation at location i for year t and hour s. The matrix $\mathbf{B}(s)$ is a cyclic cubic spline basis expansion of order 4 over the 24 hour daily cycle with associated random effects $\boldsymbol{\upsilon}_{it}$ for each site i and year t in the historical period, imposing a diurnal pattern on temperature. The model is completed by the inclusion of independent, uncorrelated Gaussian error $\boldsymbol{\varepsilon}_{its}$, giving rise to interpolated daily temperature curves for July at each fort location i and year t. From the daily temperature curves, we are able to estimate mean July temperature, aligning the sparse, irregular historical data to the monthly time scale of the observational data. To facilitate parameter estimation in the presence of sparse data, our model borrows strength among days within the month of July, reducing the influence of measurement error and improving prediction of the mean diurnal temperature curve. By borrowing strength across sites within a year, our model produces a mean estimate that has less variability than the raw observations. We fit our model to the historical data using using the `R` package `mgcv` (Wood, 2011).

After aligning the two data sources to a common temporal scale, we need a modeling framework with which 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). 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. The use of PCR for the statistical reconstruction of climate has a long tradition, dating back to Lorenz (1956). To build our spatio-temporal predictive model, we use 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. We also develop robust versions of PCR and pPCR models that are flexible to the presence of potentially outlying measurements of mean July temperature that may arise from the non-standardized data collection.

We introduce traditional PCR in Section 4.2.1, develop a temporal extension that allows for flexibility between years while borrowing strength among years to improve estimation in Section 4.2.2, and define a probabilistic extension (pPCR) of the PCR that accounts for measurement error in Section 4.2.3. In Section 4.3, we introduce the robust specification of our PCR and pPCR models that better accommodate outlying observations, and in Section 4.4, 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 Section 4.5, and we describe the simulation study in Section 4.6, where we evaluate predictive performance in a synthetic data scenario. In Section 4.7, we apply our models to reconstruct historical average July temperature in the Upper Midwestern United States, choosing the model that performs best based on our scoring rules.

4.2. Model Statement

4.2.1. Principal Component Regression. A common statistical approach for reconstruction of historical climate using calibration period data is to regress the partially observed historical climate observations onto the observed climate. For a given reconstruction year, we start with the regression model

$$y_i = \mu + \mathbf{x}_i'\boldsymbol{\alpha} + \epsilon_i, \tag{43}$$

where y_i is an observation of the m-dimensional historic climate field at location i. The full observation vector $\mathbf{y}$ consists of the m observations of the temperature field, where we observe only n out of the m locations. The columns of the $m \times d$ matrix $\mathbf{X}$ contain d replicates of the observational period climate field at the m locations, forming a basis set of patterns for the regression. The d-dimensional vector of regression coefficients $\boldsymbol{\alpha}$ link the historical observation $\mathbf{y}$ with the set of possible climate patterns $\mathbf{X}$ in the observation period, allowing for climate fields that are linear combinations of observed climate patterns, up to uncorrelated model error. Thus, we can model temperatures that are warmer or cooler than the observational period. The uncorrelated model error ϵ_i is assumed to be independent and identically distributed Gaussian with variance τ^2. Because the likelihood is unaffected if μ is integrated out, we assume that the data $\mathbf{y}$ are centered and assume $\mu = 0$ in what follows.

In (43), the matrix $\mathbf{X}$ is 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. Because we are interested in prediction of the dependent variable $\mathbf{y}$ and not interpretation of the regression coefficients, we are free to manipulate the form of $\mathbf{X}$ to improve statistical learning. We begin by taking the singular value decomposition (SVD) of $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}'$, where the columns of $\mathbf{U}$ are the left singular vectors of $\mathbf{X}$, the diagonal matrix $\mathbf{\Lambda}$ has the singular values in descending order on the diagonal, and the columns of $\mathbf{V}$ are the right singular vectors. The principal components regression (PCR) model using the SVD is

$$\begin{aligned} y_i &= \mathbf{u}_i'\mathbf{\Lambda}\mathbf{V}'\boldsymbol{\alpha} + \epsilon_i \\ &= \mathbf{u}_i'\mathbf{\Lambda}^{\frac{1}{2}}\boldsymbol{\beta} + \epsilon_i \\ &= \mathbf{z}_i'\boldsymbol{\beta} + \epsilon_i, \end{aligned}$$

where $\mathbf{u}_i$ is the i^{th} row of $\mathbf{U}$. If the regression coefficient is given the usual prior model $\boldsymbol{\alpha} \sim \mathrm{N}(\mathbf{0}, \sigma_\alpha^2\mathbf{I})$, then $\boldsymbol{\beta} = \mathbf{\Lambda}^{\frac{1}{2}}\mathbf{V}'\boldsymbol{\alpha} \sim \mathrm{N}(\mathbf{0}, \sigma_\alpha^2\mathbf{\Lambda})$. The columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{X}'\mathbf{X}$, the diagonal elements of $\mathbf{\Lambda}$ are the eigenvalues of $\mathbf{X}'\mathbf{X}$, and $\mathbf{z}_i = \mathbf{u}_i\mathbf{\Lambda}^{\frac{1}{2}}$, the scaled principal component at location i. In practical applications of PCR, one often performs dimension reduction by retaining only the first p eigenvectors of $\mathbf{U}$ in the $m \times p$ 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 0.85 (or some other similarly high value). Because truncation removes the most variable eigenvectors, the truncation implies a prior belief that shrinks the regression coefficients and provides an implicit regularization

on the model (Hastie et al., 2009, Chapter 3.6). Performing truncation, (43) becomes

$$y_i = \mathbf{z}_i'\boldsymbol{\beta} + \epsilon_i, \tag{44}$$

where the regression coefficients $\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, \sigma_\beta^2 \boldsymbol{\Lambda}_p)$. 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 (43) and can improve prediction in practice.

Preexisting research suggests that truncation of the trailing principal components is not always appropriate because the higher variability 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 slow varying or uncommon processes appear in the lesser eigenvectors of the observational period data (which is likely because they are not the primary contributors to the annual-scale variability in climate), these signals would be discarded by truncation. 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) approaches 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

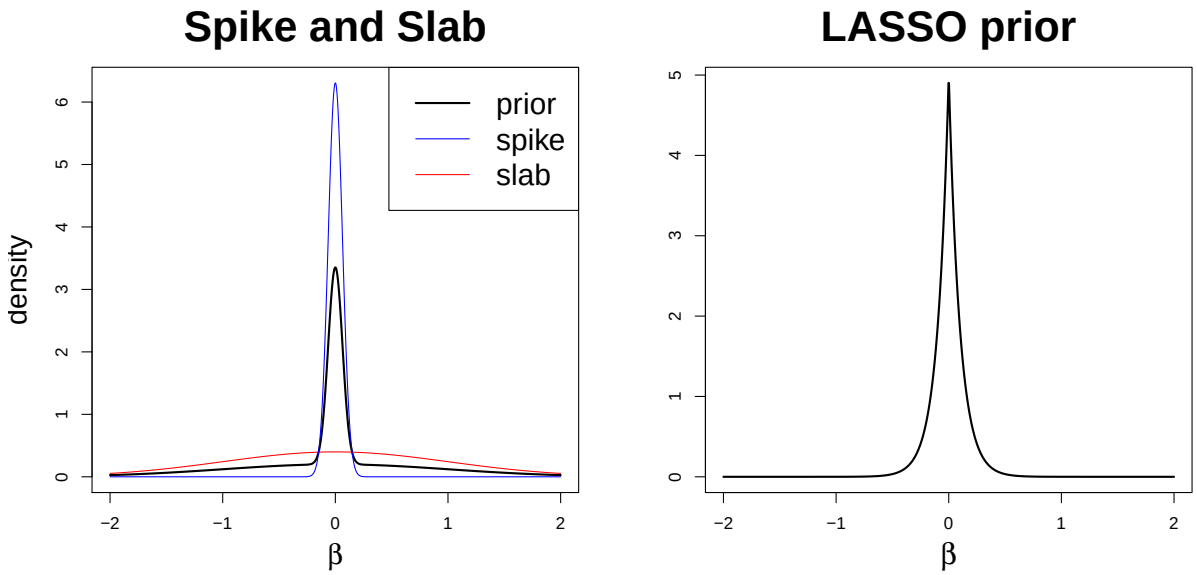


FIGURE 4.2. Plot of regularization priors. Both priors shrink the regression coefficients toward zero but the SSVS prior has heavier tails.

variable selection assumes the hierarchical prior on the j^{th} regression coefficient

$$\beta_j \sim \begin{cases} \mathrm{N}(0, \sigma^2_{\beta_j}) & \text{if } \xi_j = 1 \\ \mathrm{N}(0, \sigma^2_{\beta_j}\kappa^{-1}) & \text{if } \xi_j = 0, \end{cases}$$

where the variables ξ_j are indicators of the importance of the j^{th} latent principal component in the regression and have independent Bernoulli(Ψ_j) priors. The prior regression coefficient variance $\sigma^2_{\beta_j}$ could be assigned a prior if desired, but the shrinkage value $\kappa > 1$ must be fixed. A large κ produces a mixture distribution of a broad, relatively uninformative prior with "large" variance $\sigma^2_{\beta_j}$ (the "slab") and a highly informative prior at a small neighborhood around zero (the "spike"). This is commonly called the "spike and slab" prior and 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. An example prior with $\sigma^2_{\beta_j} = 1$, $\kappa = 250$, and $\Psi_i = 0.5$ is provided in Figure 4.2.

An alternative to variable selection methods like SSVS is penalized regression (Hastie et al., 2009). Common forms of penalized regression include ridge regression (Tikhonov or L_2 shrinkage (Hoerl and Kennard, 1970)), where one minimizes

$$\sum_{i=1}^{n}(y_i - \mathbf{z}_i\boldsymbol{\beta})^2 + \gamma\sum_{j=1}^{p}\beta_j^2$$

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

$$\sum_{i=1}^{n}(y_i - \mathbf{z}_i\boldsymbol{\beta})^2 + \gamma\sum_{j=1}^{p}|\beta_j|$$

with respect to $\boldsymbol{\beta}$ given the penalty term γ. These methods of variable selection work by penalizing large coefficients in the likelihood. 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 shrinks 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. Often, the optimal γ 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 a prior on the regression coefficients $\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, \gamma\mathbf{I})$ and the L_1 LASSO penalty assigns a Laplace (also called a Double Exponential) prior $\boldsymbol{\beta} \sim \prod_{j=1}^{d} \frac{\gamma}{2\sqrt{\tau^2}} \exp\{-\frac{\gamma|\beta_j|}{\sqrt{\tau^2}}\}$ (Figure 4.2). The LASSO prior can also be specified using the more computationally efficient hierarchical scale mixture of Gaussian distributions with exponential mixing distribution by assigning the prior

$$\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, \tau^2\mathbf{D}_\gamma)$$
$$\gamma_j \sim \mathrm{Exp}\left(\frac{\lambda^2}{2}\right),$$

where $\mathbf{D}_\gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_p)$ (Park and Casella, 2008). To perform a fully Bayesian regularization that properly accounts for parameter uncertainty, we assign the hyperprior $\lambda^2 \sim \mathrm{Gamma}(\alpha_{\lambda^2}, \beta_{\lambda^2})$ on the shrinkage parameter.

4.2.2. Temporal model. Thus far, we have considered reconstruction of only a single year in our model. In this section, we extend our model to account for the temporal nature of our data. We propose a model that is flexible and lets each year have its own parameters to capture temporal variation in climate, but also pools information hierarchically to improve parameter estimation in our sparse data scenario. Allowing the important principal components in the regression to vary with time, the model is capable of detecting slowly varying changes in temperature. For example, the irregular, but periodic, cycles of the Pacific Decadal Oscillation and the Atlantic Multidecadal Oscillation are slowly varying and likely do not explain a large portion of the variability in the temperature records (Schlesinger and Ramankutty, 1994; Mantua et al., 1997). Therefore, these and other similar climate signals could be removed through the truncation of the principal components. In fact, there are many such climate patterns that could be lost during truncation. Allowing the model

to learn about the important principal components on a year by year basis, we gain greater flexibility.

We extend the data model (44) by allowing the regression relationship to be time varying, giving the temporal regression model

$$y_{it} = \mathbf{z}_i' \boldsymbol{\beta}_t + \epsilon_{it} \tag{45}$$

where the observation y_{it} is the historical observation at site i for year t. The temporally varying regression coefficients $\boldsymbol{\beta}_t$ correspond to the influence of the latent principal components $\mathbf{Z}$ on the m_t historical observations $\mathbf{y}_t$. The time varying, independent Gaussian error $\boldsymbol{\epsilon}_t \sim \mathrm{N}(\mathbf{0}, \tau_t^2 \mathbf{I}_t)$ has variance τ_t^2 with $\mathbf{I}_t$ the m_t-dimensional identity matrix.

For the SSVS PCR model, we assign the "spike and slab" prior, where for each $j = 1, \ldots, p$,

$$\beta_{tj} \sim \begin{cases} \mathrm{N}(0, \sigma_{\beta_t}^2 \lambda_{p_j}) & \text{if } \xi_j = 1 \\ \mathrm{N}(0, \sigma_{\beta_t}^2 \kappa_j^{-1} \lambda_{p_j}) & \text{if } \xi_j = 0, \end{cases} \tag{46}$$

where λ_{p_j} is the j^{th} diagonal element of $\boldsymbol{\Lambda}_p$. The SSVS prior for the pPCR model is similar, but does not include the λ_{p_j} terms. We then pool across years by assuming the hierarchical pooling model $\sigma_{\beta_t} \sim \log\mathrm{N}(\mu_{\sigma_\beta}, \sigma_{\sigma_\beta}^2)$ with hyperpriors $\mu_{\sigma_\beta} \sim \mathrm{N}(0, 100)$ and $\sigma_{\sigma_\beta} \sim \mathrm{U}(0, 100)$.

For the LASSO PCR model, we hierarchically pool the error standard deviation by assigning the hyperprior $\tau_t \sim \log\mathrm{N}(\mu_\tau, \sigma_\tau^2)$ with hyperparameters $\mu_\tau \sim \mathrm{N}(0, 100)$ and $\sigma_\tau \sim \mathrm{U}(0, 100)$, where learning across years is achieved by updating μ_τ and σ_τ. Each year's regression coefficients $\boldsymbol{\beta}_t$ are assigned independent LASSO priors $\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, \tau_t^2 \mathbf{D}_{\gamma_t})$ where, for each $j = 1, \ldots, p$, $\gamma_{tj} \sim \mathrm{Exp}\left(\frac{\lambda_t^2}{2}\right)$. The shrinkage parameters are assigned $\lambda_t^2 \sim \mathrm{Gamma}(\alpha_\lambda, \beta_\lambda)$

priors 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 $\sigma^2_\lambda = \frac{\alpha_\lambda}{\beta^2_\lambda}$ and assigning the hyperpriors $\mu_\lambda \sim \text{logN}(0, 100)$ and $\sigma_\lambda \sim \text{U}(0, 100)$.

4.2.3. Probabilistic Principal Component Regression. PCR assumes the data, and therefore the principal components, to be observed without measurement error. The assumption that data are observed without error is not valid for our observational data. Our climate data are model interpolated, and therefore, the principal components have unaccounted for measurement error. Hence, the eigenvectors in $\mathbf{U}$ can be thought of as estimates of the true eigenvectors $\mathbf{Z}$ 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 $\mathbf{x}_i$ be the i^{th} row of $\mathbf{X}$, the model for the noisy observations is

$$\mathbf{x}_i = \mathbf{m} + \mathbf{K}\mathbf{z}_i + \boldsymbol{\eta}_i, \tag{47}$$

where $\mathbf{m}$ is d-vector of means, $\mathbf{K}$ is a $d \times p$ rotation matrix, $\mathbf{z}_i$ 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 $\mathbf{m}$ can be integrated out of the above equation without changing the likelihood, thus we assume that the data $\mathbf{X}$ have centered rows, and we set $\mathbf{m} = \mathbf{0}$. Because principal component vectors are orthonormal, we complete the principal component model specification by assigning independent priors $\mathbf{z}_i \sim \text{N}(\mathbf{0}, \mathbf{I})$, 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 choose to be $\mathbf{I}_p$). We set $\mathbf{K}$ at the MLE and rewrite (47) as

$$\mathbf{x}_i = \hat{\mathbf{K}}\mathbf{z}_i + \boldsymbol{\eta}_i. \tag{48}$$

After accounting for the measurement uncertainty in our predictor matrix $\mathbf{X}$, by estimating the unknowns $\mathbf{Z}$ and $\boldsymbol{\eta}$ in (48), we link the historical and modern observations by regressing $\mathbf{y}_t$ onto the latent eigenvectors $\mathbf{Z}$

$$y_{it} = \mathbf{z}_i'\boldsymbol{\beta}_t + \epsilon_{it}, \tag{49}$$

and estimate the unknown regression coefficients $\boldsymbol{\beta}_t$ in (49). We complete the model by assuming uncorrelated measurement error $\epsilon_{it} \sim \mathrm{N}(0, \tau_t^2)$ and assigning the hierarchical prior $\tau_t \sim \mathrm{logN}(\mu_\tau, \sigma_\tau^2)$ with hyperpriors $\mu_\tau \sim \mathrm{N}(0, 100)$ and $\sigma_\tau \sim \mathrm{U}(0, 100)$.

4.3. Robust Regression

Because the fort data were collected using non-standard methods, there is likely more variability in the data than can be explained by assuming a Gaussian distribution. Thus, we propose extending (49) to a model that is robust to outliers. The robust pPCR data model using the Student's-t distribution is

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

an over-dispersed regression model that accommodates outliers in the data. The ν_t parameter 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 $\mathrm{U}(0, 0.5)$ distribution. We generalize the prior to pool across years, assigning the inverse degrees of freedom the prior $\frac{1}{\nu_t} \sim \mathrm{Beta}(\alpha_\nu, \beta_\nu, 0, 0.5)$ where the four parameter $\mathrm{Beta}(\alpha, \beta, L, U)$ prior is a $\mathrm{Beta}(\alpha, \beta)$ prior scaled to the interval $[L, U]$. To hierarchically pool the prior model, we reparameterize the model with $\alpha_\nu = \mu_\nu\eta_\nu$ and $\beta_\nu = (1 - \mu_\nu)\eta_\nu$ for $\mu_\nu \in [0, 1]$ and $\eta_\nu \in [0, \infty)$. We complete the pooling by assigning the Jeffrey's hyperprior $\mu_\nu \sim \mathrm{Beta}(0.5, 0.5)$ and assigning the hyperprior $\eta_\nu \sim \mathrm{Gamma}(1, 1)$. 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

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

where the parameters have the same priors as introduced previously.

4.4. Posterior Distribution

The latent principal components $\mathbf{z}_i$ are high dimensional (approximately $20,000 \times p$), thus we aim to avoid the computational burden of sampling this parameter. Therefore, we integrate out the latent principal components

$$\int [y_{it}|\mathbf{z}_i, \boldsymbol{\beta}_t, \tau_t^2, \nu_t][\mathbf{x}_i|\mathbf{z}_i, \sigma^2][\mathbf{z}_i]\, d\mathbf{z}_i, \tag{50}$$

but this integral does not have an obvious analytical solution. We could attempt to numerically integrate out $\mathbf{z}_i$, but at great computational cost. Instead, we write our Student's-t data model as a scale mixture where

$$y_{it} \sim \mathrm{N}(\mathbf{z}_i'\boldsymbol{\beta}_t, v_{it}^2),$$

$$v_{it}^2 \sim \text{inv-}\chi^2(\nu_t, \tau_t^2),$$

and

$$[y_{it}|\mathbf{z}_i, \boldsymbol{\beta}_t, \tau_t^2, \nu_t] = \int [y_{it}|\mathbf{z}_i, \boldsymbol{\beta}_t, v_{it}^2][v_{it}^2|\tau_t^2, \nu_t]\, dv_{it}^2.$$

Then, we write the integral (50) as

$$\begin{aligned}
\int [y_{it}|\mathbf{z}_i, \boldsymbol{\beta}_t, \tau_t^2, \nu_t][\mathbf{x}_i|\mathbf{z}_i, \sigma^2][\mathbf{z}_i]\, d\mathbf{z}_i &= \int \left(\int [y_{it}|\mathbf{z}_i, \boldsymbol{\beta}_t, v_{it}^2][v_{it}^2|\tau_t^2, \nu_t]\, dv_{it}^2 \right) [\mathbf{x}_i|\mathbf{z}_i, \sigma^2][\mathbf{z}_i]\, d\mathbf{z}_i && (51) \\
&= \int \left(\int [y_{it}|\mathbf{z}_i, \boldsymbol{\beta}_t, v_{it}^2][\mathbf{x}_i|\mathbf{z}_i, \sigma^2][\mathbf{z}_i]\, d\mathbf{z}_i \right) [v_{it}^2|\tau_t^2, \nu_t]\, dv_{it}^2 && (52) \\
&= \int [y_{it}|\mu_{y_{it}}, \sigma_{y_{it}}^2][v_{it}^2|\tau_t^2, \nu_t]\, dv_{it}^2, && (53)
\end{aligned}$$

where the integral in (51) is evaluated in MCMC by first sampling $v_{it}^2 \sim \text{inv-}\chi^2(\nu_t, \tau_t^2)$ then evaluating the density $[y_{it}|\mu_{y_{it}}, \sigma_{y_{it}}^2]$. These modifications result in the integrated data model (see Appendix for details)

$$y_{it} \sim \text{N}(\mu_{y_{it}}, \sigma_{y_{it}}^2), \tag{54}$$

where

$$\mu_{y_{it}} = \frac{\mathbf{x}_i \hat{\mathbf{K}} \mathbf{M}_p^{-1} \boldsymbol{\beta}_t}{\sigma^2},$$

$$\sigma_{y_{it}}^2 = v_{it}^2 + \boldsymbol{\beta}_t' \mathbf{M}_p^{-1} \boldsymbol{\beta}_t,$$

with $\mathbf{M}_p^{-1} = (\frac{\boldsymbol{\Lambda}_p}{\sigma^2} + \mathbf{I}_p)^{-1}$, a diagonal matrix that can be inverted efficiently. For our data, integration results in a significant computational savings because we avoid sampling p vectors of length $m \approx 20,000$. The cost for not sampling the latent principal components $\mathbf{Z}$ is the loss of conjugacy for the regression coefficients $\boldsymbol{\beta}_t$ in the MCMC sampler, necessitating the use of Metropolis-Hastings updates. The posterior distribution from which we sample (for the robust pPCA model with LASSO regularization) using MCMC is

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

where H_t is the set of locations where there are observations for year t.

4.5. Scoring Rules

To evaluate model performance, we use scoring rules. 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 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 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. Because CE and RE are improper, it is possible that the optimal prediction can, on average, have a better 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). In general, MSPE is not proper, but because our data models are Gaussian and Student's-t, MSPE is proper for this case.

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

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

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

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m-m_t}\sum_{i\notin H_t}\left(\mathrm{E}(\tilde{y}_{it}|\mathbf{y}_t) - y_{oos,it}\right)^2.$$

The MSPE uses the posterior predictive mean, a point summary, instead of the full posterior distribution. Hence, 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. The smaller variance model should be preferred (if both predictions are equally accurate), 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. In our previous example, CRPS would score the prediction with smaller posterior predictive variance as the better predictive model when both models have the same posterior predictive mode (which is equivalent to the posterior predictive mean in symmetric distributions where MSPE is proper). Several recent papers on 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 cumulative distribution function, F_{it}, at location i and time t and out-of-sample observations $\mathbf{y}_{oos,t}$, the CRPS is defined as

$$CRPS(\{F_{it}\}_{t=1}^{T}, \mathbf{y}_{oos,t}) = -\sum_{t=1}^{T}\sum_{i\notin H_t}\int_{-\infty}^{\infty}\left(F_{it}(y) - I_{\{y\geq y_{oos,it}\}}\right)^2 dy. \tag{55}$$

Gneiting and Raftery (2007) show that (55) can be written alternatively as

$$CRPS(\{F_{it}\}_{t=1}^{T}, \mathbf{y}_{oos}) = \sum_{t=1}^{T}\frac{1}{m_{H_t}}\sum_{i\notin H_t}\left(E_{F_{it}}\left|y_{it} - y_{oos,it}\right| - \frac{1}{2}E_{F_{it}}\left|y_{it} - y_{it}^{*}\right|\right), \tag{56}$$

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 (56) 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 $\tilde{\mathbf{y}}_t^{(k)}$ from the posterior predictive distribution $\left[\tilde{\mathbf{y}}_t^{(k)}\middle|\mathbf{y}_t\right]$ at each post burn-in iteration k. Then, (56) is approximated by

$$\widehat{CRPS}(\{\hat{F}_{it}\}_{t=1}^{T}, \mathbf{y}_{oos}) = \sum_{t=1}^{T}\left(\frac{1}{m_{H_t}}\sum_{i\notin H_t}\left(\frac{1}{K}\sum_{k=1}^{K}\left|\tilde{y}_{it}^{(k)} - y_{oos,it}\right| - \frac{1}{2K^2}\sum_{k=1}^{K}\sum_{\ell=1}^{K}\left|\tilde{y}_{it}^{(k)} - \tilde{y}_{it}^{(\ell)}\right|\right)\right). \tag{57}$$

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, even though these methods can be costly computationally and

time consuming to implement. An alternative is to use the approximate leave-one-out cross-validation method LOO (Vehtari et al., 2016). 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). The target score to be approximated is the expected log pointwise predictive density for new data

$$elpd = \sum_{t=1}^{T} \sum_{i \in H_t} \int \left(\log[\tilde{y}_{it}|\mathbf{y}_t]\right) [\tilde{y}_{it}] \, d\tilde{y}_{it} \tag{58}$$

where $[\tilde{y}_{it}]$ is the unknown distribution of the true data generating process. Because the true data generating distribution in (58) is unknown in practice, we cannot use (58) as a measure of model performance. Instead, we can estimate *elpd* using the log pointwise predictive density (Celeux et al., 2006; Watanabe, 2010)

$$lpd = \sum_{t=1}^{T} \sum_{i \in H_t} \log[y_{it}|\mathbf{y}_t] = \sum_{t=1}^{T} \sum_{i \in H_t} \log \int [y_{it}|\boldsymbol{\theta}_t][\boldsymbol{\theta}_t|\mathbf{y}_t] \, d\boldsymbol{\theta}_t. \tag{59}$$

From our posterior samples, we approximate the integral in (59) using Monte Carlo integration (Gelman et al., 2014)

$$\widehat{lpd} = \sum_{t=1}^{T} \sum_{i \in H_t} \log \left(\frac{1}{K} \sum_{k=1}^{K} [y_{it}|\boldsymbol{\theta}_t^{(k)}] \right). \tag{60}$$

There is a problem with using (60) as a predictive score because the observed data $\mathbf{y}_t$ are used twice; to estimate a posterior density $[\boldsymbol{\theta}_t|\mathbf{y}_t]$ and to evaluate the density $[y_{it}|\boldsymbol{\theta}_t^{(k)}]$. Thus, use of (60) is an overestimate of the predictive ability and unsatisfactory as a predictive score (Gelman et al., 2014; Hobbs and Hooten, 2015). Instead, we can eliminate using the data

twice by defining the leave-one-out log pointwise predictive density

$$lpd_{loo} = \sum_{t=1}^{T} \sum_{i \in H_t} \log[y_{it}|\mathbf{y}_{(i)t}] \tag{61}$$

$$= \sum_{t=1}^{T} \sum_{i \in H_t} \log \int [y_{it}|\boldsymbol{\theta}_t][\boldsymbol{\theta}_t|\mathbf{y}_{(i)t}]\, d\boldsymbol{\theta}_t, \tag{62}$$

where $\mathbf{y}_{(i)t}$ are the data $\mathbf{y}_t$ at time t without the i^{th} location. One can calculate (61) directly by cross-validation at a high computational cost, or one can approximate (61) from post burn-in posterior samples using the full data as described below.

Following Gelfand et al. (1992) and Gelfand (1996) where each observation is conditionally independent given model parameters $\boldsymbol{\theta}_t$, we express the integral in (61) using importance ratios $r_{it} \equiv \frac{[\boldsymbol{\theta}_t|\mathbf{y}_{(i)t}]}{[\boldsymbol{\theta}_t|\mathbf{y}_t]} \propto \frac{1}{[y_{it}|\boldsymbol{\theta}_t]}$ as

$$\sum_{t=1}^{T} \sum_{i \in H_t} \log \int [y_{it}|\boldsymbol{\theta}_t] \frac{[\boldsymbol{\theta}_t|\mathbf{y}_{(i)t}]}{[\boldsymbol{\theta}_t|\mathbf{y}_t]} [\boldsymbol{\theta}_t|\mathbf{y}_t]\, d\boldsymbol{\theta}_t \propto \frac{\sum_{t=1}^{T} \sum_{i \in H_t} \log \int [y_{it}|\boldsymbol{\theta}_t] r_{it} [\boldsymbol{\theta}_t|\mathbf{y}_t]\, d\boldsymbol{\theta}_t}{\sum_{t=1}^{T} \sum_{i \in H_t} \log \int r_{it} [\boldsymbol{\theta}_t|\mathbf{y}_t]\, d\boldsymbol{\theta}_t} \tag{63}$$

Then, using MCMC samples we can approximate (63) using sampled importance ratios $r_{it}^{(k)}$ as

$$\widehat{lpd}_{loo} = \sum_{t=1}^{T} \sum_{i \in H_t} \log \left(\frac{\sum_{k=1}^{K} r_{it}^{(k)} [y_{it}|\boldsymbol{\theta}_t^{(k)}]}{\sum_{k=1}^{K} r_{it}^{(k)}} \right) = \sum_{t=1}^{T} \sum_{i \in H_t} \log \left(\frac{1}{\frac{1}{K} \sum_{k=1}^{K} \frac{1}{[y_{it}|\boldsymbol{\theta}_t^{(k)}]}} \right), \tag{64}$$

where the posterior using the full data $[\boldsymbol{\theta}_t|\mathbf{y}_t]$ is likely to have lower variance and thinner tails than the posterior using leave-one-out data $[\boldsymbol{\theta}_t|\mathbf{y}_{(i)t}]$, implying that the importance ratios could have large or potentially even infinite variances. Importance ratios with high variance can cause the estimate in (64) 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. The generalized Pareto distribution function

$$[x|\mu,\sigma,\xi] = \frac{1}{\sigma}(1+\xi(x-\mu))^{-\frac{1}{\xi+1}}, \tag{65}$$

is commonly used to model the tails of distributions to investigate extreme behavior by fitting (65) to the largest M of the K posterior importance sampling weights $r_{it}^{(k)}$. If the estimated tail parameter $\hat{\xi}_{it}$ is less than $\frac{1}{2}$, the variance of the importance ratio is finite and the importance ratios approximating the log posterior score holding out y_{it} can be used directly to approximate LOO by setting the importance weights $w_{it}^{(k)} = r_{it}^{(k)}$. If the estimated tail parameter $\frac{1}{2} < \hat{\xi}_{it} < 1$, the variance of the importance ratios is infinite, but the mean of the importance ratios exists. Hence, Vehtari and Gelman (2015) propose smoothing the importance ratios by fitting the generalized Pareto distribution to the largest $M = 0.2K$ importance weights $\{r_{it}^{(k)}, k = 1,\ldots,K\}$ and generating the smoothed importance ratios

$$\tilde{w}_{it}^{(k)} = \begin{cases} r_{it}^{(k)} & \text{if } r_{it}^{(k)} \text{ is in the bulk of the distribution} \\ F^{-1}\left(\frac{q-\frac{1}{2}}{M}\right) & \text{if } r_{it}^{(k)} \text{ is in the tail of the distribution,} \end{cases}$$

where F^{-1} is the inverse CDF of the generalized Pareto distribution (65) and $q = \{1,\ldots,M\}$ are the order statistics of the M largest importance ratios. To guarantee finite variance of the importance sampling weights, Vehtari and Gelman (2015) propose truncating any smoothed importance weights $\tilde{w}_{it}^{(k)}$ larger than $K^{\frac{3}{4}}\bar{w}_{it}$, where $\bar{w}_{it} = \frac{1}{K}\sum_{k=1}^{K}\tilde{w}_{it}^{(k)}$ is the mean of the smoothed importance weights $\{\tilde{w}_{it}^{(k)}, k = 1,\ldots,K\}$, resulting in the smoothed and truncated importance weights $w_{it}^{(k)}$. The smoothed importance ratios now have finite variance and can be used to approximate the leave-one-out cross validation, although the approximation LOO

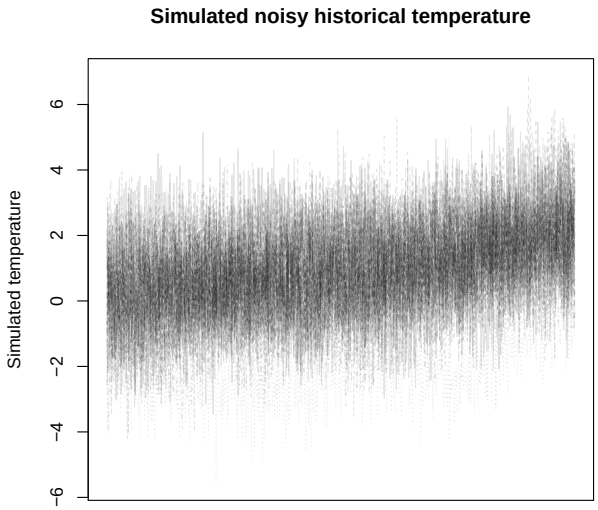
converges to the true cross-validation at a slower rate than if not smoothed. If the estimated tail parameter $\hat{\xi}_{it} > 1$, the mean and variance of the importance ratios likely do not exist but the variance of the smoothed importance ratios is finite, but large, and the use of LOO is sensitive to the held out observation. Using the smoothed importance weights, we obtain the Pareto-smoothed importance sampling approximation

$$\widehat{elpd}_{PSIS} = \sum_{t=1}^{T} \sum_{i \in H_t} \log \left(\frac{\sum_{k=1}^{K} w_{it}^{(k)} [y_{it} | \boldsymbol{\theta}_t^{(k)}]}{\sum_{k=1}^{K} w_{it}^{(k)}} \right). \tag{66}$$

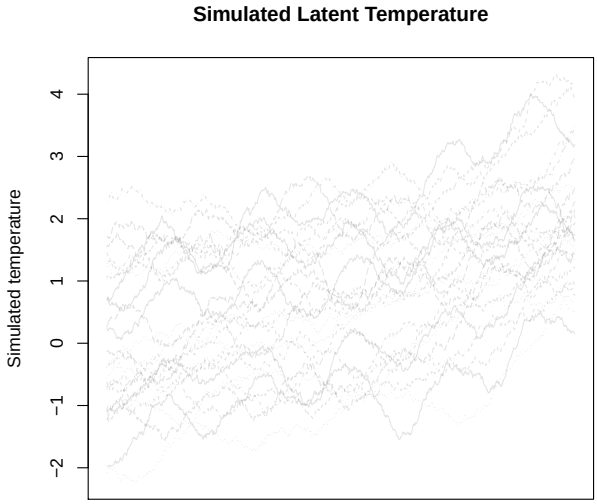
We use the deviance scale and set $\widehat{LOO} = -2\widehat{elpd}_{PSIS}$ to make LOO a negatively oriented score (the best model is the one with the lowest score).

4.6. Simulation

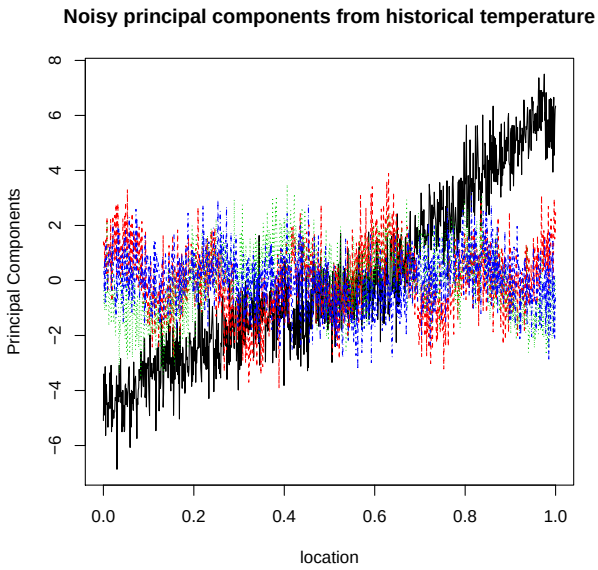
The nature of paleoclimate data makes it difficult to verify the predictive ability of our model using cross-validation. With only a handful of observations available for each year we aim to reconstruct, cross-validation techniques could be highly biased due to the effects of unusual observations in small sample sizes. The potential for bias is especially important given we expect noisy and potentially outlying observations due to the data collection procedure. Additionally, the high dimensionality of the field we aim to reconstruct, and the use of computationally intensive MCMC estimation, make cross-validation costly. Instead of evaluating model performance with cross-validation, we conduct a simulation study to explore different models for the historical fort data and evaluate model performance using the scoring rules above. Although we do not simulate from the model that is used for estimation, we believe the simulated data provide a reasonable approximation to the true generating process for average July temperature, providing an environment for model testing and exploration of empirical properties.



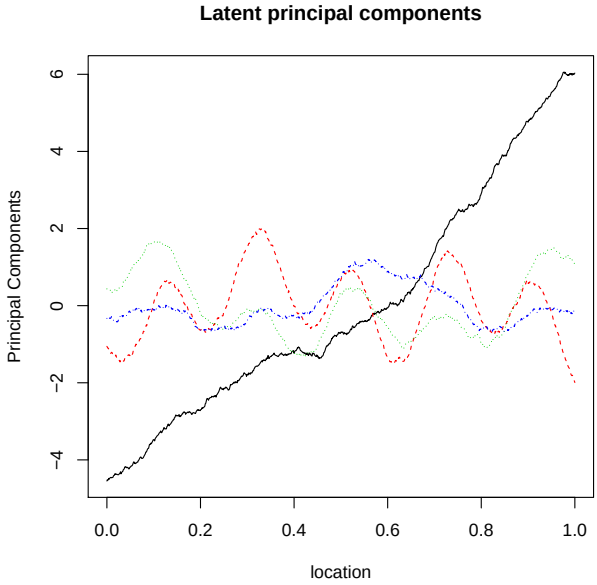
(A) Noisy temperature patterns.



(B) Latent temperature patterns.



(C) Noisy principal components.



(D) Latent principal components.

FIGURE 4.3. Plot of simulated data showing latent climate process, observed noisy data, latent principal components, and noisy, observed principal components.

We simulate mean July temperature in one spatial dimension (this is easily extended into two dimensions using the real data), allowing for faster computation and easier graphical exploration of the spatio-temporal process. We simulate $T = 50$ realizations from the model

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

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 and Atlantic Multidecadal Oscillation that influence mean July temperature. We do not claim our simulation behaves like any climatological process, only that the toy problem facilitates exploration of interesting patterns potentially seen in climatological data.

We include a spatially correlated random effect $\boldsymbol{\eta}_t \sim \mathrm{N}\left(\mathbf{0}, \sigma^2_{\eta_t}\mathbf{R}\left(\boldsymbol{\phi}\right)\right)$ that smooths the patterns, generating realizations of a one-dimensional climate field. A common choice for the form of $\mathbf{R}\left(\boldsymbol{\phi}\right)$ 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, j^{th} element $\mathbf{R}_{ij}(\boldsymbol{\phi}) = \exp\left\{-d_{ij}\phi\right\}$ where d_{ij} represents the Euclidean distance between the i^{th} and j^{th} spatial locations and ϕ is the spatial range parameter.

To create observations that match the temporal irregularities and spatial clustering behavior in the historical fort data, we sample the one-dimensional spatial field using weighted probabilities that generate clustered observations in space, storing the simulated temperature observations at the m_t locations in the vector $\mathbf{y}_t$. Using our sampling design, we generate noisy samples for the historical period years $t = 1, \ldots, 25$ with

$$y_{it} = s_{it} + \tilde{\epsilon}_{it}, \tag{67}$$

where $\tilde{\epsilon}_{it}$ is independent Gaussian error with variance $\tilde{\sigma}^2 = 1.5$ that represents the uncertainty in historical temperature measurements at the fort locations and s_{t_i} is the latent

temperature at location i in year t. To generate the set of patterns $\mathbf{X}$ that will be used in our regression model, we sample for $t = 26, \ldots, 50$

$$\mathbf{x}_t = \mathbf{s}_t + \ddot{\boldsymbol{\epsilon}}_t, \tag{68}$$

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 observational period. The measurements from the observational period are PRISM model interpolated data and have measurement error, but the measurement error should be less than that of the historical period, hence $\tilde{\sigma}^2 > \ddot{\sigma}^2$. We then combine the simulated values into the noisy pattern matrix $\mathbf{X} \equiv (\mathbf{x}_{26}, \ldots \mathbf{x}_{50})$ that is used in our model framework. The noiseless matrix of temperature patterns $\mathbf{S} \equiv (\mathbf{s}_1, \ldots \mathbf{s}_{25})$ is 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 $\mathbf{y}_t$, limiting the ability to improve reconstruction skill. Figure 4.3a shows the simulated noisy temperature realizations and Figure 4.3b shows the simulated latent temperature field that is the target of our reconstruction, with the x-axis representing spatial location. The noisy principal components are plotted in Figure 4.3c with the latent principal components shown in Figure 4.3d showing the effect of measurement error on the principal components. Thus, we have a situation that is analogous to the errors-in-covariates framework, where noisy observations of the covariates (in our case the principal components) can lead to bias in the regression coefficients, inflated residual variance, and a reduction in prediction skill (Fuller, 2009; Buonaccorsi, 2010; Carroll et al., 2006).

We compare the performance of each model specification using MSPE, CRPS, and LOO scoring rules in our simulation where the best model is the one with the smallest score. We fit the PCR and pPCR models using SSVS and LASSO methods of regularization with both

TABLE 4.1. Simulation experiment scores. Smaller values indicated better model performance.

(A) MSPE.

	Gaussian		Robust	
	PCR	pPCR	PCR	pPCR
SSVS	0.8430	0.8450	0.8433	0.8441
LASSO	0.8383	0.8469	0.8386	0.8469

(B) CRPS.

	Gaussian		Robust	
	PCR	pPCR	PCR	pPCR
SSVS	271.5	272.8	271.8	272.9
LASSO	270.8	273.6	270.9	273.4

(C) LOO.

	Gaussian		Robust	
	PCR	pPCR	PCR	pPCR
SSVS	9897	9910	9898	9911
LASSO	9889	9908	9888	9907

the Gaussian and robust Student's-t data models. Thus, we compare eight models, with the results displayed in Table 4.1. Across scores, the models perform similarly, with the PCR and robust PCR models slightly outperforming the pPCR and robust pPCR models. The Pareto tail parameter estimates of the LOO importance sampling weights provide an additional diagnostic of model performance. Pareto tail parameter estimates for a data point greater than 0.5 show the model is not adequately predicting that data point, and tail parameters greater than one show severe problems with the model. Examining the Pareto tail parameters for each of the models fit to the simulated data shows the robust models have slightly less evidence of model misspecification (fewer tail parameters greater than 0.5) than the traditional Gaussian data model. Visual examination of reconstructions not shown in this manuscript for therobust PCR and robust pPCR show that, like the tables of scoring rules (Table 4.1), there is little difference among the methods. All models do not reconstruct the fine scale, within year variability, which is expected given the small sample sizes. However, the models are able to reconstruct the large-scale, inter-annual differences, which are one of the main features of the simulated data.

4.7. Fort Data Reconstruction

Now that we have explored our model framework using a simulation study, we apply our models to the historical measurements of temperature. We fit the eight models used in the simulation study to the data and present the results in Table 4.2a. With respect to the LOO scores, the PCR models outperform the pPCR models, which is expected given the smaller sample sizes in the reconstruction than in the simulation. In addition, the Gaussian data model outperforms the robust Student's-t data model. When we examine the LOO Pareto tail parameter plots in Figure 4.4a we see evidence of misspecification because many of the tail parameter estimates are greater than 0.5, and every model has some tail parameter estimates greater than one. The LOO Pareto tail parameter plots give us the added benefit of identifying a severe outlier, occurring at data point 549 in Figure 4.4a, corresponding to an average July temperature measurement of 47 degrees F, an unrealistic observation. After removing the outlier and refitting the models, we see improved fit because the LOO values have decreased (Table 4.2b), and the plots of the Pareto tail parameter values in Figure 4.4b have fewer values greater than 0.5. After removing the outlier, the best performing model is the robust PCR model, with the Gaussian PCR model also performing well. The performance of the robust pPCR model also improves with removal of the outlier, while the Gaussian pPCR model shows evidence of misspecification due to the presence of two large Pareto tail parameter estimates, possibly additional outlying observations the Gaussian model cannot accommodate. All models still show evidence of misspecification because some Pareto tail parameter estimates are greater than 0.5, but the removal of the outlier has improved model fit.

To visualize our results, we plot reconstructions of four representative years of the historical temperature surfaces using the robust PCR model (Figure 4.5) and the robust pPCR

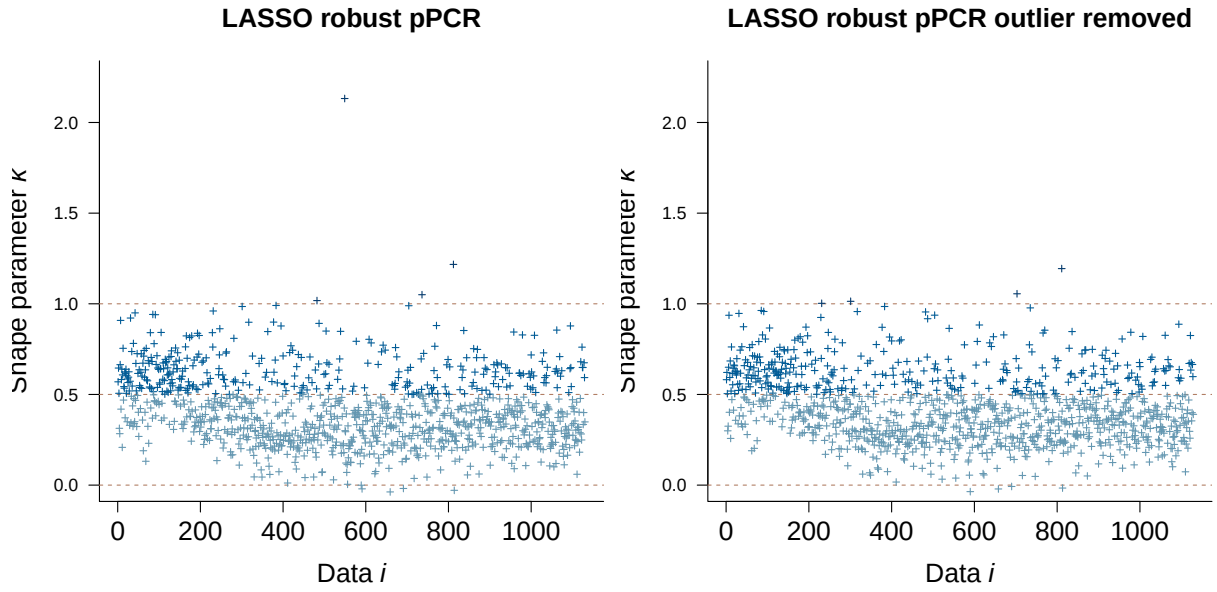
TABLE 4.2. Fort historical reconstruction scores. Smaller values indicate better model performance

(A) LOO.

	Gaussian		Robust	
	PCR	pPCR	PCR	pPCR
SSVS	5117	5231	5122	5250
LASSO	5163	5226	5181	5250

(B) LOO with outlier removed.

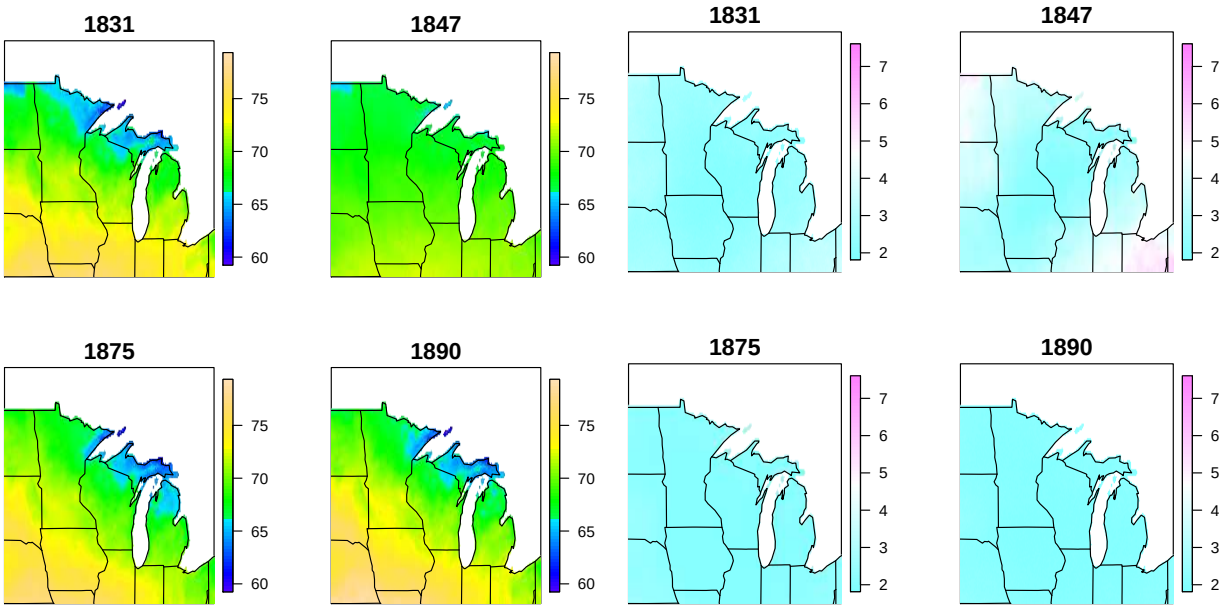
	Gaussian		Robust	
	PCR	pPCR	PCR	pPCR
SSVS	5093	5325	5008	5156
LASSO	5138	5318	5050	5150



(A) With outlier. (B) With outlier removed.

FIGURE 4.4. Plot of fort data LOO Pareto shape estimates. Values less than 0.5 show good model performance and values over 1.0 show poor model performance.

model (Figure 4.6). Visual comparison of these reconstructions illustrates the differences in the two models. The robust PCR model assumes the climate patterns in the observational data are without error; the influence of these patterns is seen in the posterior predictive mean surface Figure 4.5a. In comparison, the robust pPCR model shows less influence of these climate patterns in the posterior predictive mean (Figure 4.6a) because the model includes the assumption that some of the pattern is noise, producing reconstructions with less spatial variability. Differences in posterior predictive standard deviations are also evident (Figures



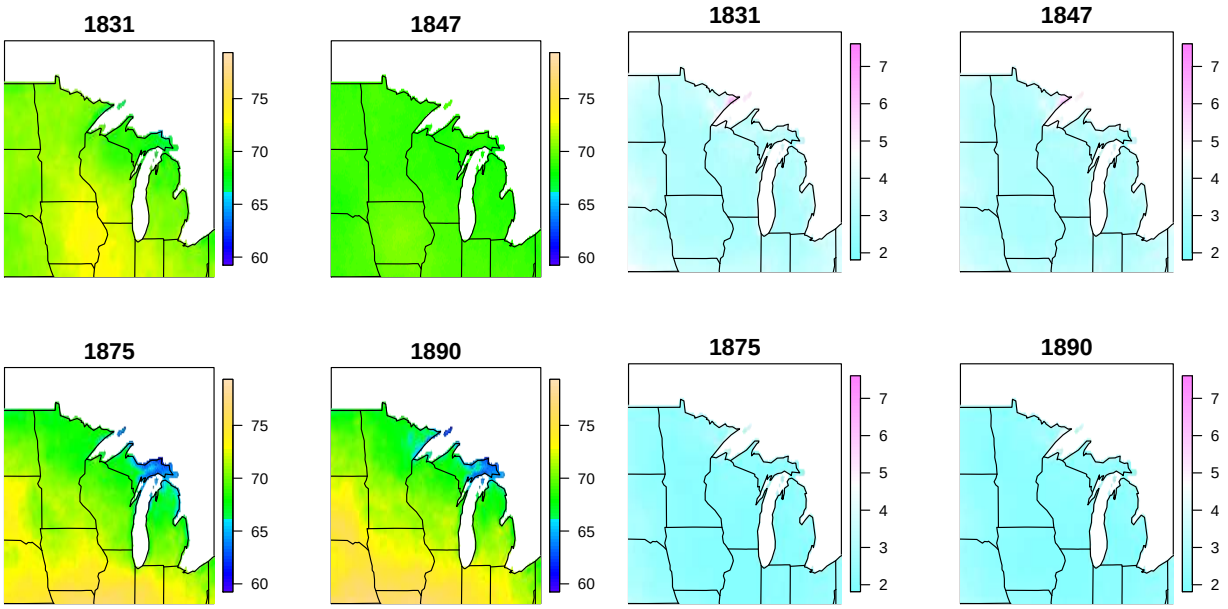
(A) Posterior predictive mean. (B) Posterior predictive standard deviation.

FIGURE 4.5. Plot of fort temperature reconstruction using robust PCR model.

4.5b and 4.6b). The robust PCR model shows reduction of uncertainty in the spatial locations near observations and higher uncertainty away from the observations while the robust pPCR model has posterior predictive standard deviations that are more spatially diffuse and perhaps more realistic. Overall, both the robust PCR and the robust pPCR models perform well, but the best performing model according to our scoring rule is the robust PCR model, due to low LOO scores and the least evidence of model misspecification in the estimated Pareto tail parameter plots.

4.8. Conclusion

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 proxy measurements of climate. The nuances of paleoclimate data often require specialized modeling techniques and careful investigation into modeling assumptions and performance.



(A) Posterior predictive mean. (B) Posterior predictive standard deviation.

FIGURE 4.6. Plot of fort temperature reconstruction using robust probabilistic PCR model.

In addition, careful investigation is needed to validate paleoclimate reconstruction skill. We presented a framework to explore rigorous climate reconstruction methods. We extended principal component regression methods, applying regularization techniques to choose important principal components, generating robust models to account for the presence of outliers, and exploring the use of a probabilistic principal component model to account for measurement uncertainty in the observed temperature patterns. By evaluating the predictive skill of our models in a statistically rigorous manner, we were able to explore our extensions of PCR for climate reconstruction, laying the groundwork for exploration of richer datasets better suited to these techniques. Future research using more complex climate data than the PRISM monthly mean temperatures is warranted, especially for climate variables that are far from stationary Gaussian processes (e.g., wind speed or precipitation).

We also presented a framework for evaluating paleoclimate reconstructions using simulation studies and proper scoring rules. By using proper scoring rules and exploring model

performance in a simulation framework, we have stronger support for our predictions. We presented three scoring rules developed in the statistics literature 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 novel and computationally efficient LOO approximation to leave-one-out cross-validation. Our use of LOO to score the historical fort model predictions not only enabled us to perform model selection, but also aided in diagnosing an outlying observation and refining model fit. Thus, based on our results in simulation and application, the use of LOO in uncorrelated Bayesian models is promising.

Ultimately, our temperature reconstructions extend the climatological record in the upper Midwestern United States 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. Because previous work has suggested that average July temperature is a good predictor of tree species assemblage, our climate reconstruction can be used to learn about forest ecology in the region. By extending the average July temperature records into the past with our models, we gain the potential to better understand how forest ecology has changed with climate change and improve future climate reconstructions.

Recognition of Support

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.

CHAPTER 5

Conclusion

5.1. Overview

In this dissertation, I presented several statistical models applied to ecological and environmental data. In each of these applications, a primary goal was the evaluation and improvement of estimation and/or predictive skill. Each modeling effort is unique to its respective application, thus I developed model frameworks and evaluated predictive performance using scoring rules appropriate for the data and question at hand. In the previous chapters, I investigated estimator performance for survey sampling estimators, developed predictive hierarchical models with non-linear, mechanistic components, and constructed a Bayesian probabilistic principal component model with data driven variable selection for a pattern matching predictive model.

In Chapter 2, I explored the empirical properties of the EPSE estimator under a variety of situations where the theoretical properties are unknown. Using United States Forest Service Forest Inventory Analysis data to examine the empirical performance of the EPSE variance estimator, I demonstrated that the EPSE variance estimator is biased when the strata are chosen in such a way as to minimize the variance estimate. In addition, the resampling study showed that setting the EPSE stratum boundaries at fixed quantiles *a priori* produces variance estimates that are empirically unbiased, thus demonstrating that use of fixed quantile EPSE stratum boundaries can serve as a default choice. The simulation study provided valuable insights into which stratum boundary scenarios deserve further theoretical exploration and have been used as default settings for improving predictions for the 2011 National Land Cover Database (Fry et al., 2011).

In Chapter 3, I developed a novel method for reconstructing climate using a deterministic model that approximates tree ring growth given monthly climate. By conducting a simulation study in which data were generated to be as similar to the observed data as possible using the mechanistic model, I showed that the model accurately reconstructs paleoclimate with appropriate uncertainties, especially precipitation. The simulation study also revealed strengths and limitations of the model, especially a lack of precision in the temperature reconstruction. Because the model accurately estimated the associated temperature growth parameters, this demonstrates that the lack of learning about temperature arises from an interaction between the change of temporal support and the mechanistic growth model. The use of proper scoring rules allowed exploration into the statistical consequences of different modeling assumptions on the reconstruction skill. Additionally, an unexpected benefit arising from the mechanistic model framework is the ability to estimate ecological climate niches (i.e., which tree species are drought tolerant, etc.), providing a tool for future exploration of ecology through dendrochronological samples.

In Chapter 4, I presented a framework for reconstructing historical climate using compiled historical data. The historical data are sparse in both time and space and include potentially outlying observations that result from non-standardized data collection. Because of the challenges that the data present for modeling, it was important to explore the consequences of modeling assumptions on predictive ability through a simulation study. Comparing predictive ability in the simulation study using proper scoring rules demonstrated that the computationally efficient approximation to leave-one-out cross-validation (LOO) using the proper log score can be applied for efficient model evaluation within our study framework. In addition, the use of LOO to score model predictions using the fort data identified outlying observations, resulting in improved predictions and better model fit.

5.2. Extensions of Current Work

The research initiated in this dissertation is ongoing; there is potential for improvement, especially in improving fidelity to the underlying ecological process. For example, the growth functions used in Chapter 3 are monotonically increasing in both temperature and precipitation. Clearly, this is only a simple approximation of how trees respond to climate; for example, at high temperatures evapotranspiration results in water loss and reduced growth. Additionally, there is no reason to suspect that trees respond to climate uniformly throughout the year. In fact, it seems likely that the ideal climate for tree growth changes throughout the year. Generalizing the growth functions to accommodate more realistic ecology should improve the model and possibly produce more precise temperature reconstructions. Another potential improvement in the tree ring model presented in Chapter 3 involved constructing a better algorithm for sampling the climate state-space model. Sampling the latent state-space jointly with a Kalman filter (Cressie and Wikle, 2011, Chapter 8) or approximating the latent state-space with a particle filter (Andrieu and Roberts, 2009) embedded within the MCMC algorithm could improve mixing, reduce computation cost, and allow for parallelizing the model for fitting in a high performance computing environment.

Currently, the bivariate climate reconstruction from tree ring widths in Chapter 3 is not spatially explicit, thus the model would need to be extended to properly account for spatial autocorrelation if this model is to be applied to a large spatial domain. By extending the model to have a spatio-temporally correlated precipitation and temperature reconstruction, the model can borrow strength among nearby locations and possibly improve reconstruction skill, but at an increased computational cost. One way to reduce the computational burden is to allow each subregion to have independent mechanistic growth model parameters and to fit each of these growth models in parallel given the spatio-temporal multivariate climate

field. The proposed model would thereby produce a spatio-temporal multivariate climate reconstruction without significantly increasing computational burden in a high performance computing environment.

In Chapter 4, I considered a reconstruction of mean July temperature. A natural extension of this work is to generalize the model framework to model all twelve months, resulting in a state-space model for the latent climate. Within the state-space context, I could better explore the effects of slowly varying climate patterns on the reconstruction effort. Also, the addition of temporal information could improve model performance by borrowing strength across months in the year. Finally, I could exploit state-space methods like including a Kalman filter or particle filter embedded within the MCMC algorithm to improve mixing and parallelize model fitting.

5.3. Concluding Remarks

In this chapter, I have discussed ways to improve the ecological applications presented in the dissertation and proposed statistical extensions for future work. Fundamentally, extension of the work in this dissertation requires close collaboration between statisticians and scientists, not only to include better ecological realism within the statistical models, but also to disseminate statistical ideas to the wider scientific community. By bridging the gap between the statistical and scientific communities, both disciplines benefit and many interesting questions can be addressed.

BIBLIOGRAPHY

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723.

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, pages 697–725.

Andsager, K., M.C., K., and Spinar, M. (2004). 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.

Bahadur, R. R. (1966). A note on quantiles in large samples. *Annals of Mathematical Statistics*, 37(3):577–580.

Banerjee, S., Gelfand, A. E., and Carlin, B. P. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press.

Barboza, L., Li, B., Tingley, M. P., Viens, F. G., et al. (2014). Reconstructing past temperatures from natural proxies and estimated climate forcings using short-and long-memory models. *The Annals of Applied Statistics*, 8(4):1966–2001.

Bechtold, W. A. and Patterson, P. L. (2005). *The enhanced forest inventory and analysis program: national sampling design and estimation procedures*. US Department of Agriculture Forest Service, Southern Research Station Asheville, North Carolina.

Bell, W. and Ogilvie, A. (1978). Weather compilations as a source of data for the reconstruction of European climate during the medieval period. *Climatic Change*, 1(4):331–348.

Berger, J. O. (2013). *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media.

Bernardo, J. M. and Smith, A. (2009). *Bayesian Theory*, volume 405. John Wiley & Sons.

Billingsley, P. (2008). *Probability and Measure*. John Wiley & Sons.

Blaauw, M. and Christen, J. (2005). Radiocarbon peat chronologies and environmental change. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(4):805–816.

Blaauw, M. and Christen, J. (2011). Flexible paleoclimate age-depth models using an autoregressive gamma process. *Bayesian Analysis*, 6(3):457–474.

Bradley, R. S. (2011). High-resolution paleoclimatology. In *Dendroclimatology*, pages 3–15. Springer.

Brázdil, R., Kundzewicz, Z., and Benito, G. (2006). Historical hydrology for studying flood risk in Europe. *Hydrological Sciences Journal*, 51(5):739–764.

Breidt, F. J. and Opsomer, J. D. (2008). Endogenous post-stratification in surveys: classifying with a sample-fitted model. *The Annals of Statistics*, pages 403–427.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

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

Carbone, M. S., Czimczik, C. I., Keenan, T. F., Murakami, P. F., Pederson, N., Schaberg, P. G., Xu, X., and Richardson, A. D. (2013). Age, allocation and availability of nonstructural carbon in mature red maple trees. *New Phytologist*, 200(4):1145–1155.

Carlin, B. P. and Banerjee, S. (2003). Hierarchical multivariate CAR models for spatio-temporally correlated survival data. *Bayesian Statistics*, 7:45–63.

Carlin, B. P. and Louis, T. A. (2011). *Bayesian Methods for Data Analysis*. CRC Press.

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

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician*, 39(2):83–87.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*, volume 2. Duxbury Pacific Grove, CA.

Celeux, G., Forbes, F., Robert, C. P., Titterington, D. M., et al. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, 1(4):651–673.

Christiansen, B. and Ljungqvist, F. C. (2011). Reconstruction of the extratropical NH mean temperature over the last millennium with a method that preserves low-frequency variability. *Journal of Climate*, 24(23):6013–6034.

Climate Database Modernization Program. http://mrcc.isws.illinois.edu/research/cdmp/cdmp.html.

Cook, E. R., Briffa, K., and Jones, P. (1994). Spatial regression methods in dendroclimatology: A review and comparison of two techniques. *International Journal of Climatology*, 14(4):379–402.

Cook, E. R. and Kairiukstis, L. A. (1990). *Methods of Dendrochronology: Applications in the Environmental Sciences*. Springer.

Cook, E. R. and Pederson, N. (2011). Uncertainty, emergence, and statistics in dendrochronology. In *Dendroclimatology*, pages 77–112. Springer.

Coulston, J. W., Moisen, G. G., Wilson, B. T., Finco, M. V., Cohen, W. B., Brewer, C. K., et al. (2012). Modeling percent tree canopy cover: A pilot study. *Photogrammetric Engineering and Remote Sensing*, 78(7):715–727.

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. John Wiley & Sons.

Crist, E. P. and Cicone, R. C. (1984). A physically-based transformation of Thematic Mapper data—The TM Tasseled Cap. *Geoscience and Remote Sensing, IEEE Transactions on*, (3):256–263.

Dahlke, M., Breidt, F. J., Opsomer, J. D., Van Keilegom, I., et al. (2013). Nonparametric endogenous post-stratification estimation. *Statistica Sinica*, 23:189–211.

D'Arrigo, R. D., Kaufmann, R. K., Davi, N., Jacoby, G. C., Laskowski, C., Myneni, R. B., and Cherubini, P. (2004). Thresholds for warming-induced growth decline at elevational tree line in the Yukon Territory, Canada. *Global Biogeochemical Cycles*, 18(3).

Dawid, A. P. (1984). Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, pages 278–292.

Eddelbuettel, D. and Sanderson, C. (2014). Rcpparmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis*, 71:1054–1063.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

Erdös, P. and Rényi, A. (1959). On the central limit theorem for samples from a finite population. *Publications of the Mathematics Institute of the Hungarian Academy of Science*, 4:49–61.

Evans, M. N., Tolwinski-Ward, S., Thompson, D., and Anchukaitis, K. J. (2013). Applications of proxy system modeling in high resolution paleoclimatology. *Quaternary Science Reviews*, 76:16–28.

Fritts, H. (1976). *Tree Rings and Climate*. Elsevier.

Fry, J. A., Xian, G., Jin, S., Dewitz, J. A., Homer, C. G., LIMIN, Y., Barnes, C. A., Herold, N. D., and Wickham, J. D. (2011). Completion of the 2006 national land cover database for the conterminous United States. *Photogrammetric Engineering and Remote Sensing*, 77(9):858–864.

Fuller, W. A. (2009). *Measurement Error Models*, volume 305. John Wiley & Sons.

García-Suárez, A., Butler, C., and Baillie, M. (2009). Climate signal in tree-ring chronologies in a temperate climate: A multi-species approach. *Dendrochronologia*, 27(3):183–198.

Geisser, S. and Eddy, W. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160.

Gelfand, A. E. (1996). Model determination using sampling-based methods. *Markov chain Monte Carlo in Practice*, pages 145–161.

Gelfand, A. E., Dey, D., and Chang, H. (1992). *Bayesian Statistics 4*, chapter Model determination using predictive distributions with implementation via sampling based methods (with discussion), pages 147–167. Oxford University Press.

Gelfand, A. E. and Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4(1):11–15.

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

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, pages 457–472.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Gislason, P. O., Benediktsson, J. A., and Sveinsson, J. R. (2006). Random forests for land cover classification. *Pattern Recognition Letters*, 27(4):294–300.

Gneiting, T. (2008). Editorial: probabilistic forecasting. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(2):319–321.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151.

Gneiting, T. and Raftery, A. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Goring, S., Williams, J. W., Ruid, M., MacLachlan, J. S., Jackson, S. T., Paciorek, C. J., Thurman, A., Zhu, J., Brooks, W., Mladenoff, D., Cogbill, C., Record, S., and Dietz, M. C. (2013). Estimating pre-settlement vegetation in the american midwest: Exploring climate relationships and links to proxy data for robust data assimilation. Miami, FL. International Biogeography Society Biennial Meeting.

Gotway, C. A. and Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458):632–648.

Griffin, D., Woodhouse, C. A., Meko, D. M., Stahle, D. W., Faulstich, H. L., Carrillo, C., Touchan, R., Castro, C. L., and Leavitt, S. W. (2013). North American monsoon precipitation reconstructed from tree-ring latewood. *Geophysical Research Letters*, 40(5):954–958.

Grimmett, G. and Stirzaker, D. (2001). *Probability and Random Processes*. Oxford University Press.

Guillot, D., Rajaratnam, B., and Emile-Geay, J. (2015). Statistical paleoclimate reconstructions via Markov random fields. *The Annals of Applied Statistics*, 9(1):324–352.

Hadi, A. S. and Ling, R. (1998). Some cautionary notes on the use of principal components regression. *The American Statistician*, 52(1):15–19.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematics Institute of the Hungarian Academy of Science*, 5:361–74.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, pages 1491–1523.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*. Springer series in statistics. Springer, Berlin, 2nd edition.

Hobbs, N. T. and Hooten, M. B. (2015). *Bayesian Models: A Statistical Primer for Ecologists*. Princeton University Press.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Holst, L. (1973). Some limit theorems with applications in sampling theory. *The Annals of Statistics*, pages 644–658.

Hooten, M. and Hobbs, N. (2015). A guide to Bayesian model selection for ecologists. *Ecological Monographs*, 85(1):3–28.

Hooten, M. B. and Wikle, C. K. (2007). Shifts in the spatio-temporal growth dynamics of shortleaf pine. *Environmental and Ecological Statistics*, 14(3):207–227.

Horowitz, J. L. and Manski, C. F. (2006). Identification and estimation of statistical functionals using incomplete data. *Journal of Econometrics*, 132(2):445–459.

Huber, P. J. and Ronchetti, E. (2011). *Robust statistics*. Springer.

Jackson, T. A., Moisen, G. G., Patterson, P. L., and Tipton, J. R. (2012). Repeatability in photo-interpretation of tree canopy cover and its effect on predictive mapping. *In: McWilliams, Will; Roesch, Francis A. eds. 2012. Monitoring Across Borders: 2010 Joint Meeting of the Forest Inventory and Analysis (FIA) Symposium and the Southern Mensurationists. e-Gen. Tech. Rep. SRS-157. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station*, pages 189–192.

Jin, X., Carlin, B. P., and Banerjee, S. (2005). Generalized hierarchical multivariate CAR models for areal data. *Biometrics*, 61(4):950–961.

Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Applied Statistics*, pages 300–303.

Jones, P., Briffa, K., Barnett, T., and Tett, S. (1998). High-resolution palaeoclimatic records for the last millennium: Interpretation, integration and comparison with General Circulation Model control-run temperatures. *The Holocene*, 8(4):455–471.

Jones, P., Briffa, K., Osborn, T., Lough, J., Van Ommen, T., Vinther, B., Luterbacher, J., Wahl, E., Zwiers, F., Mann, M., et al. (2009). High-resolution palaeoclimatology of the last millennium: a review of current status and future prospects. *The Holocene*, 19(1):3–49.

Kastellet, E., Nesje, A., and Pedersen, E. (1998). Reconstructing the palaeoclimate of Jæren, Southwestern Norway, for the period 1821–1850, from historical documentary records. *Geografiska Annaler: Series A, Physical Geography*, 80(1):51–65.

Koopman, S. J., Shephard, N., and Creal, D. (2009). Testing the assumptions behind importance sampling. *Journal of Econometrics*, 149(1):2–11.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*, volume 31. Springer Science & Business Media.

Li, B., Nychka, D. W., and Ammann, C. M. (2010). The value of multiproxy reconstruction of past climate. *Journal of the American Statistical Association*, 105(491):883–895.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

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

Madow, W. G. et al. (1948). On the limiting distributions of estimates based on samples from finite universes. *The Annals of Mathematical Statistics*, 19(4):535–545.

Mann, M. E., Zhang, Z., Hughes, M. K., Bradley, R. S., Miller, S. K., Rutherford, S., and Ni, F. (2008). Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia. *Proceedings of the National Academy of Sciences*, 105(36):13252–13257.

Mantua, N. J., Hare, S. R., Zhang, Y., Wallace, J. M., and Francis, R. C. (1997). A Pacific interdecadal climate oscillation with impacts on salmon production. *Bulletin of the American Meteorological Society*, 78(6):1069–1079.

Mardia, K. (1988). Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. *Journal of Multivariate Analysis*, 24(2):265–284.

Martin-Benito, D. and Pederson, N. (2015). Convergence in drought stress, but a divergence of climatic drivers across a latitudinal gradient in a temperate broadleaf forest. *Journal of Biogeography*, 42(5):925–937.

McRoberts, R. E. (2010). Probability-and model-based approaches to inference for proportion forest using satellite imagery as ancillary data. *Remote Sensing of Environment*, 114(5):1017–1025.

McRoberts, R. E., Gobakken, T., and Næsset, E. (2012). Post-stratified estimation of forest area and growing stock volume using lidar-based stratifications. *Remote Sensing of Environment*, 125:157–166.

McRoberts, R. E., Holden, G. R., Nelson, M. D., Liknes, G. C., and Gormanson, D. D. (2005). Using satellite imagery as ancillary data for increasing the precision of estimates for the Forest Inventory and Analysis program of the USDA Forest Service. *Canadian Journal of Forest Research*, 35(12):2968–2980.

McRoberts, R. E., Næsset, E., and Gobakken, T. (2013). Inference for lidar-assisted estimation of forest growing stock volume. *Remote Sensing of Environment*, 128:268–275.

Melvin, T. M. and Briffa, K. R. (2008). A "signal-free" approach to dendroclimatic standardisation. *Dendrochronologia*, 26(2):71–86.

Moberg, A., Sonechkin, D. M., Holmgren, K., Datsenko, N. M., and Karlén, W. (2005). Highly variable Northern Hemisphere temperatures reconstructed from low-and high-resolution proxy data. *Nature*, 433(7026):613–617.

Moisen, G. G. and Frescino, T. S. (2002). Comparing five modelling techniques for predicting forest characteristics. *Ecological Modelling*, 157(2):209–225.

Murphy, A. and Daan, H. (1985). *Probability, Statistics, and Decision Making in the Atmosperic Sciences*, pages 379–437. Westview Press, Boulder, CO.

Noorbaloochi, S. and Meeden, G. (1983). Unbiasedness as the dual of being Bayes. *Journal of the American Statistical Association*, 78(383):619–623.

Ogilvie, A. E. (1984). The past climate and sea-ice record from Iceland, Part 1: Data to AD 1780. *Climatic Change*, 6(2):131–152.

Paciorek, C. J. and McLachlan, J. (2009). Mapping ancient forests: Bayesian inference for spatio-temporal trends in forest composition using the fossil pollen proxy record. *Journal of the American Statistical Association*, 104(486):608–622.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Pederson, N., Bell, A. R., Cook, E. R., Lall, U., Devineni, N., Seager, R., Eggleston, K., and Vranes, K. P. (2013). Is an epic pluvial masking the water insecurity of the greater New York City region? *Journal of Climate*, 26(4):1339–1354.

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

PRISM Climate Group, Oregon State University.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rosen, B. (1972). Asymptotic theory for successive sampling with varying probabilities without replacement, I. *The Annals of Mathematical Statistics*, pages 373–397.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Number 12. Cambridge University Press.

Rutherford, S., Mann, M., Delworth, T., and Stouffer, R. (2003). Climate field reconstruction under stationary and nonstationary forcing. *Journal of Climate*, 16(3):462–479.

Rutherford, S., Mann, M., Osborn, T., Briffa, K., Jones, P., Bradley, R., and Hughes, M. (2005). Proxy-based Northern Hemisphere surface temperature reconstructions: Sensitivity to method, predictor network, target season, and target domain. *Journal of Climate*, 18(13):2308–2329.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model Assisted Survey Sampling*. Springer Science & Business Media.

Schlesinger, M. E. and Ramankutty, N. (1994). An oscillation in the global climate system of period 65-70 years. *Nature*, 367(6465):723–726.

Scott, C. T., Bechtold, W. A., Reams, G. A., Smith, W. D., Westfall, J. A., Hansen, M. H., and Moisen, G. G. (2005). Sample-based estimators used by the forest inventory and analysis national information management system. *The Enhanced Forest Inventory and Analysis Program—National Sampling Design and Estimation Procedures*, pages 43–67.

Shashkin, A. and Vaganov, E. (1993). Simulation-model of climatically determined variability of conifers annual increment (on the example of common pine in the steppe zone). *Russian Journal of Ecology*, 24(5):275–280.

Smerdon, J. E. (2012). Climate models as a test bed for climate reconstruction methods: Pseudoproxy experiments. *Wiley Interdisciplinary Reviews: Climate Change*, 3(1):63–77.

Steig, E. J., Schneider, D. P., Rutherford, S. D., Mann, M. E., Comiso, J. C., and Shindell, D. T. (2009). Warming of the Antarctic ice-sheet surface since the 1957 international geophysical year. *Nature*, 457(7228):459–462.

Stocker, T., Qin, D., Plattner, G., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, B., and Midgley, B. (2013). IPCC, 2013: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tingley, M. P., Craigmile, P. F., Haran, M., Li, B., Mannshardt, E., and Rajaratnam, B. (2012). Piecing together the past: Statistical insights into paleoclimatic reconstructions. *Quaternary Science Reviews*, 35:1–22.

Tingley, M. P. and Huybers, P. (2010a). A Bayesian algorithm for reconstructing climate anomalies in space and time. Part I: Development and applications to paleoclimate reconstruction problems. *Journal of Climate*, 23(10):2759–2781.

Tingley, M. P. and Huybers, P. (2010b). A Bayesian algorithm for reconstructing climate anomalies in space and time. Part II: Comparison with the regularized expectation-maximization algorithm. *Journal of Climate*, 23(10):2782–2800.

Tipping, M. E. and Bishop, C. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622.

Tipton, J., Hooten, M., Pederson, N., Tingley, M., and Bishop, D. (2016). Reconstruction of late Holocene climate based on tree growth and mechanistic hierarchical models. *Environmetrics*, 27(1):42–54.

Tipton, J., Moisen, G., Patterson, P., Jackson, T. A., and Coulston, J. (2012). Sampling intensity and normalizations: Exploring cost-driving factors in nationwide mapping of tree canopy cover. *In: McWilliams, Will; Roesch, Francis A. eds. 2012. Monitoring Across Borders: 2010 Joint Meeting of the Forest Inventory and Analysis (FIA) Symposium and the Southern Mensurationists. e-Gen. Tech. Rep. SRS-157. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station.*

Tipton, J., Opsomer, J., and Moisen, G. (2013). Properties of endogenous post-stratified estimation using remote sensing data. *Remote Sensing of Environment*, 139:130–137.

Tolwinski-Ward, S., Tingley, M., Evans, M., Hughes, M., and Nychka, D. (2014). Probabilistic reconstructions of local temperature and soil moisture from tree-ring data with potentially time-varying climatic response. *Climate Dynamics*, 44(3-4):791–806.

Tolwinski-Ward, S. E., Anchukaitis, K. J., and Evans, M. N. (2013). Bayesian parameter estimation and interpretation for an intermediate model of tree-ring width. *Climate of the Past*, 9(4):1481–1493.

Tolwinski-Ward, S. E., Evans, M. N., Hughes, M. K., and Anchukaitis, K. J. (2011). An efficient forward model of the climate controls on interannual variation in tree-ring width. *Climate Dynamics*, 36(11-12):2419–2439.

Vaganov, E. A., Hughes, M. K., and Shashkin, A. V. (2006). *Growth Dynamics of Conifer Tree Rings: Images of Past and Future Environments*, volume 183. Springer.

Vehtari, A. and Gelman, A. (2015). Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646v2*.

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

Wang, L. (2012). Bayesian principal component regression with data-driven component selection. *Journal of Applied Statistics*, 39(6):1177–1189.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, 11:3571–3594.

Werner, J. P. and Tingley, M. (2015). Technical note: Probabilistically constraining proxy age–depth models within a Bayesian hierarchical reconstruction model. *Climate of the Past*, 11(3):533–545.

Williams, P. J. and Hooten, M. B. (2016). Combining statistical inference and decisions in ecology. *In Prep*.

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

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36.

Zhang, Z., Mann, M. E., and Cook, E. R. (2004). Alternative methods of proxy-based climate field reconstruction: Application to summer drought over the conterminous United States back to AD 1700 from tree-ring data. *The Holocene*, 14(4):502–516.

APPENDIX A

Supplementary Material for Chapter 3

Supplementary Material, including code, can be found at GitHub at http://github.com/jtipton25/Mechanistic-Tree-Ring.

APPENDIX B

Supplementary Material for Chapter 4

B.1. Marginal distribution

Conditional on the unobserved latent principal components, we can write the robust Student's-t pPCR likelihood as

$$
\begin{aligned}
&\left[y_{it}\middle|\mathbf{Z}_i,\boldsymbol{\beta}_t,v_{it}^2\right]\left[\mathbf{X}_i|\mathbf{Z}_i,\sigma\right]\left[\mathbf{Z}_i\right] \\
&\qquad\propto\exp\left\{-\frac{(y_{it}-\mathbf{Z}_i\boldsymbol{\beta}_t)^2}{2v_{it}^2}-\frac{\left(\mathbf{X}_i-\hat{\mathbf{K}}\mathbf{Z}_i\right)'\left(\mathbf{X}_i-\hat{\mathbf{K}}\mathbf{Z}_i\right)}{2\sigma^2}-\frac{\mathbf{Z}_i'\mathbf{Z}_i}{2}\right\} \\
&\qquad\propto\exp\left\{-\frac{1}{2}\mathbf{Z}_i'\left(\frac{\boldsymbol{\beta}_t\boldsymbol{\beta}_t'}{v_{it}^2}+\frac{\hat{\mathbf{K}}'\hat{\mathbf{K}}}{\sigma^2}+\mathbf{I}\right)\mathbf{Z}_i+\mathbf{Z}_i'\left(\frac{\boldsymbol{\beta}_t y_{it}}{v_{it}^2}+\frac{\hat{\mathbf{K}}'\mathbf{X}_i}{\sigma^2}\right)\right\} \\
&\qquad\qquad\times\exp\left\{-\frac{y_{it}^2}{2v_{it}^2}-\frac{\mathbf{X}_i'\mathbf{X}_i}{2\sigma^2}\right\} \\
&\qquad\propto\exp\left\{-\frac{1}{2}\left(\mathbf{Z}_i-\boldsymbol{\mu}_{Z_{it}}\right)'\boldsymbol{\Sigma}_{Z_{it}}^{-1}\left(\mathbf{Z}_i-\boldsymbol{\mu}_{Z_{it}}\right)-\frac{y_{it}^2}{2v_{it}^2}-\frac{\mathbf{X}_i'\mathbf{X}_i}{2\sigma^2}+\frac{\boldsymbol{\mu}_{Z_{it}}'\boldsymbol{\Sigma}_{Z_{it}}^{-1}\boldsymbol{\mu}_{Z_{it}}}{2}\right\}
\end{aligned}
$$

where $\boldsymbol{\Sigma}_{Z_{it}}=\left(\frac{\boldsymbol{\beta}_t\boldsymbol{\beta}_t'}{v_{it}^2}+\frac{\hat{\mathbf{K}}'\hat{\mathbf{K}}}{\sigma^2}+\mathbf{I}\right)^{-1}$ and $\boldsymbol{\mu}_{Z_{it}}=\boldsymbol{\Sigma}_{Z_{it}}\left(\frac{\boldsymbol{\beta}_t y_{it}}{v_{it}^2}+\frac{\hat{\mathbf{K}}'\mathbf{X}_i}{\sigma^2}\right)$. Integrating the above equation with respect to $\mathbf{Z}_i$ we get

$$
\begin{aligned}
\left[y_{it}\middle|\boldsymbol{\beta}_t,v_{it}^2,\sigma\right]&\propto\exp\left\{-\frac{y_{it}^2}{2v_{it}^2}+\frac{\boldsymbol{\mu}_{Z_{it}}'\boldsymbol{\Sigma}_{Z_{it}}^{-1}\boldsymbol{\mu}_{Z_{it}}}{2}\right\} \\
&\propto\exp\left\{-\frac{y_{it}^2}{2v_{it}^2}+\frac{\left(\boldsymbol{\Sigma}_{Z_{it}}\left(\frac{y_{it}\boldsymbol{\beta}_t}{v_{it}^2}+\frac{\hat{\mathbf{K}}'\mathbf{X}_i}{\sigma^2}\right)\right)'\boldsymbol{\Sigma}_{Z_{it}}^{-1}\left(\boldsymbol{\Sigma}_{Z_{it}}\left(\frac{y_{it}\boldsymbol{\beta}_t}{v_{it}^2}+\frac{\hat{\mathbf{K}}'\mathbf{X}_i}{\sigma^2}\right)\right)}{2}\right\} \\
&\propto\exp\left\{-\frac{1}{2}y_{it}^2\left(\frac{1}{v_{it}^2}-\frac{\boldsymbol{\beta}_t'\boldsymbol{\Sigma}_{Z_{it}}\boldsymbol{\beta}_t}{(v_{it}^2)^2}\right)+y_{it}\left(\frac{\boldsymbol{\beta}_t'\boldsymbol{\Sigma}_{Z_{it}}\hat{\mathbf{K}}'\mathbf{X}_i}{v_{it}^2\sigma^2}\right)\right\}
\end{aligned}
$$

which is $\mathrm{N}\left(\mu_{y_{it}}, \sigma^2_{y_{it}}\right)$ with

$$\begin{aligned}
\sigma^2_{y_{it}} &= \left(\frac{1}{v^2_{it}} - \frac{\boldsymbol{\beta}_t' \boldsymbol{\Sigma}_{Z_{it}} \boldsymbol{\beta}_t}{(v^2_{it})^2}\right)^{-1} \\
\mu_{y_{it}} &= \left(\frac{1}{v^2_{it}} - \frac{\boldsymbol{\beta}_t' \boldsymbol{\Sigma}_{Z_{it}} \boldsymbol{\beta}_t}{(v^2_{it})^2}\right)^{-1} \frac{\boldsymbol{\beta}_t' \boldsymbol{\Sigma}_{Z_{it}} \hat{\mathbf{K}}' \mathbf{X}_i}{v^2_{it} \sigma^2}
\end{aligned}$$

To further simplify this equation, we use the Morrison-Woodbury matrix inversion formula

$$(\mathbf{A} + \mathbf{UDV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{D}^{-1} + \mathbf{VA}^{-1}\mathbf{U}\right)^{-1}\mathbf{VA}^{-1}.$$

Then, defining $\mathbf{A} = v^2_{it}\mathbf{I}$, $\mathbf{U} = \boldsymbol{\beta}_t'$, $\mathbf{V} = \boldsymbol{\beta}_t$, and $\mathbf{D} = \left(\frac{\hat{\mathbf{K}}'\hat{\mathbf{K}}}{\sigma^2} + \mathbf{I}\right)^{-1}$,

$$\begin{aligned}
\sigma^2_{y_{it}} &= \left(\frac{\mathbf{I}}{v^2_{it}} - \frac{\boldsymbol{\beta}_t' \boldsymbol{\Sigma}_{Z_{it}} \boldsymbol{\beta}_t}{(v^2_{it})^2}\right)^{-1} \\
&= \left(\frac{\mathbf{I}}{v^2_{it}} - \frac{\boldsymbol{\beta}_t'}{v^2_{it}}\left(\frac{\boldsymbol{\beta}_t \boldsymbol{\beta}_t'}{v^2_{it}} + \frac{\hat{\mathbf{K}}'\hat{\mathbf{K}}}{\sigma^2} + \mathbf{I}\right)^{-1} \frac{\boldsymbol{\beta}_t}{v^2_{it}}\right)^{-1} \\
&= v^2_{it} + \boldsymbol{\beta}_t'\left(\frac{\hat{\mathbf{K}}'\hat{\mathbf{K}}}{\sigma^2} + \mathbf{I}\right)^{-1}\boldsymbol{\beta}_t \\
&= v^2_{it} + \boldsymbol{\beta}_t' \mathbf{M}_p^{-1} \boldsymbol{\beta}_t
\end{aligned}$$

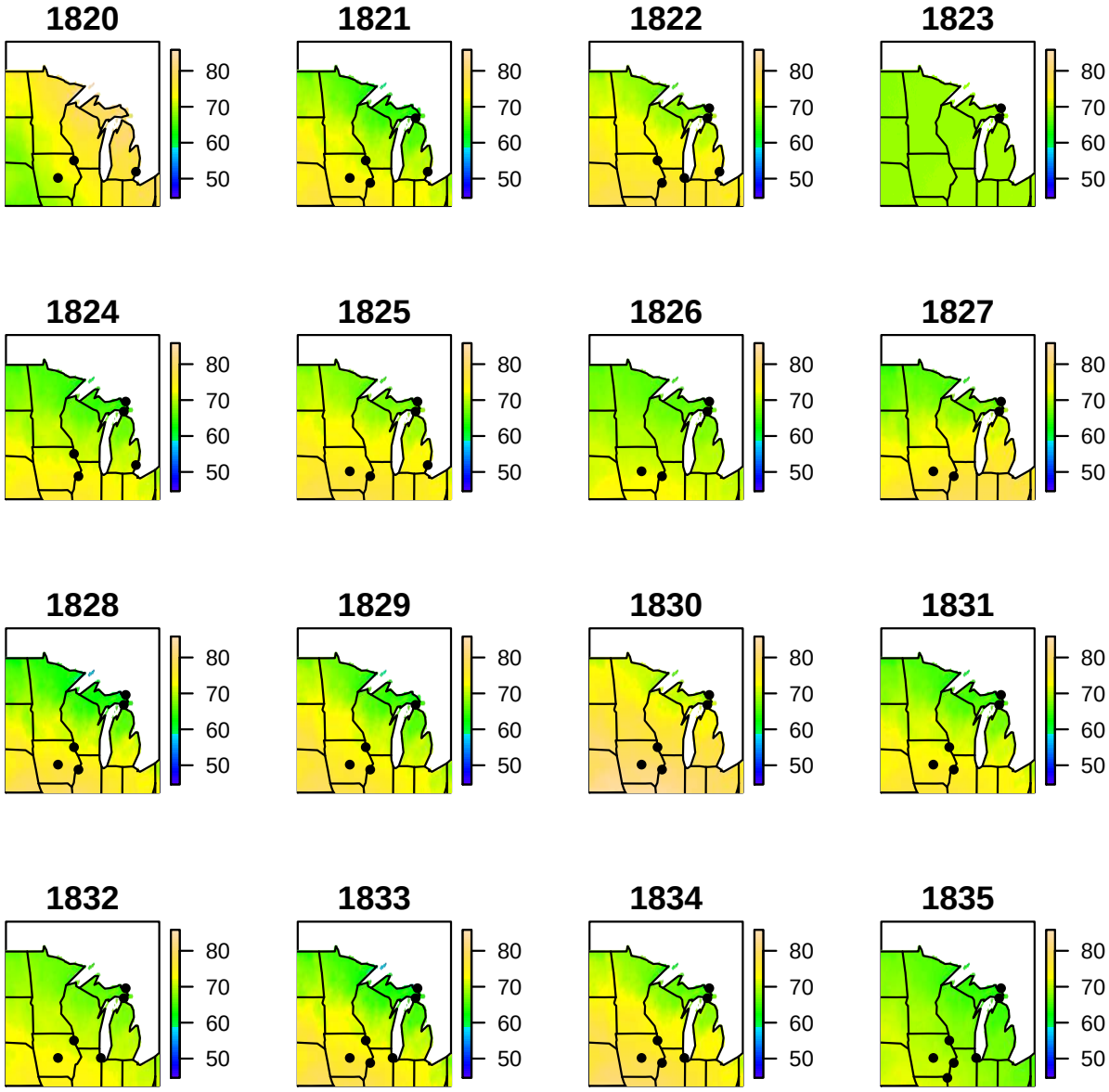
with $\mathbf{M}_p = \frac{\hat{\mathbf{K}}'\hat{\mathbf{K}}}{\sigma^2} + \mathbf{I}$.

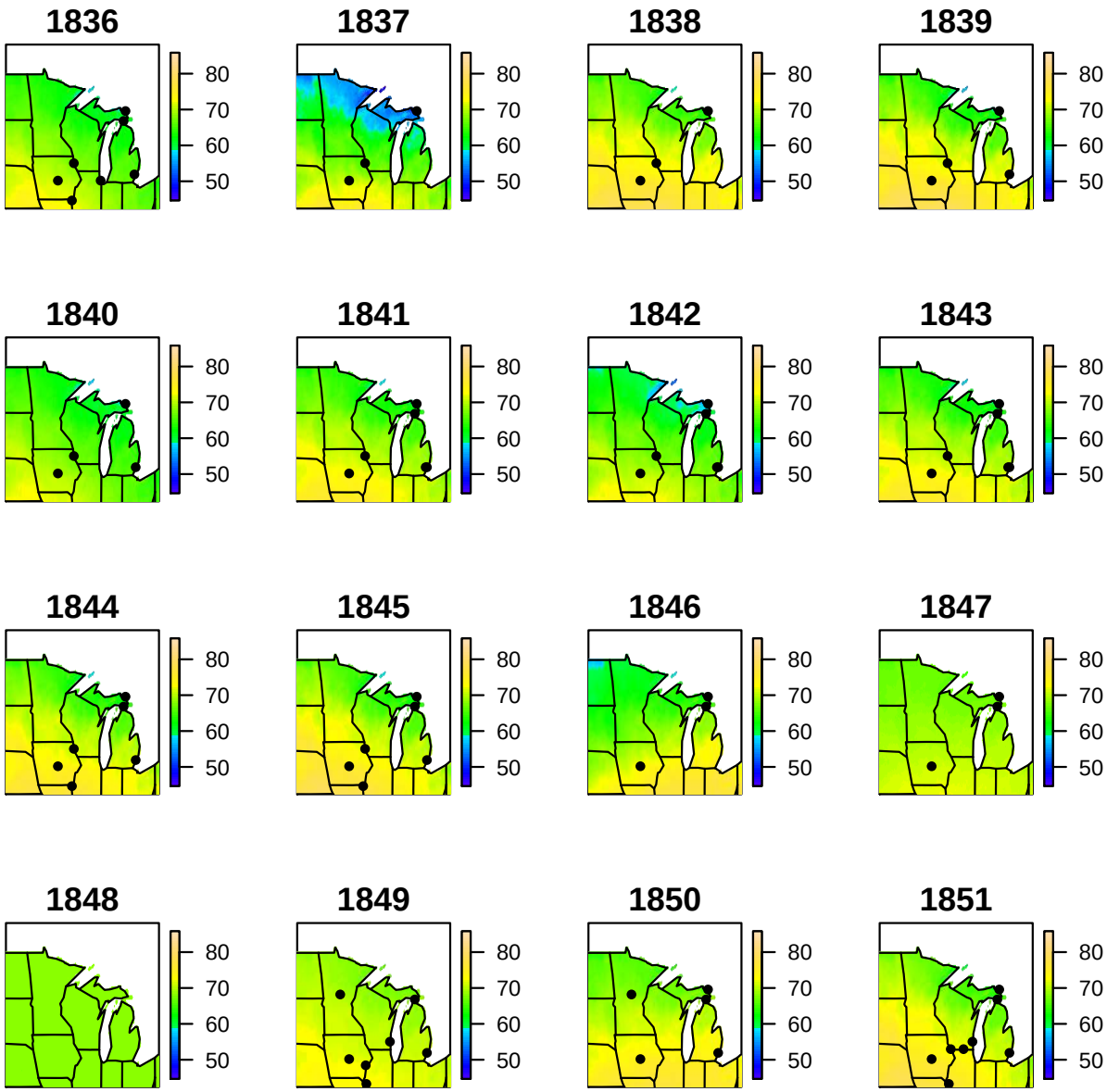
Then, letting $\mathbf{B}_t = \mathbf{M}_p^{-1/2}\boldsymbol{\beta}_t$ the conditional mean $\mu_{y_{it}}$ is

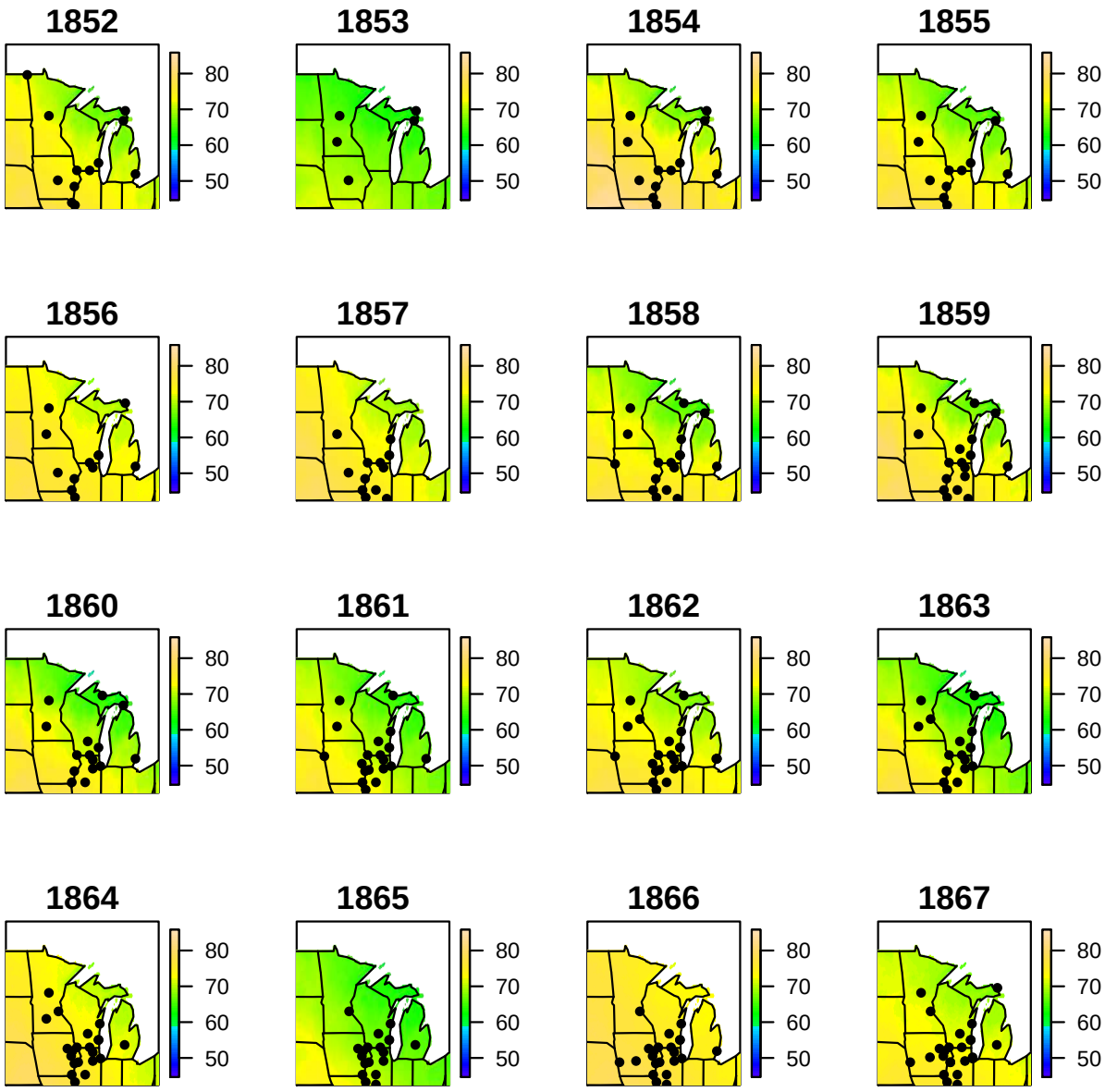
$$
\begin{aligned}
\mu_{y_{it}} &= \left(\frac{1}{v_{it}^2} - \frac{\boldsymbol{\beta}_t'\boldsymbol{\Sigma}_{Z_{it}}\boldsymbol{\beta}_t}{(v_{it}^2)^2}\right)^{-1}\left(\frac{\boldsymbol{\beta}_t'\boldsymbol{\Sigma}_{Z_{it}}\hat{\mathbf{K}}'\mathbf{X}_i}{v_{it}^2\sigma^2}\right) \\
&= \left(v_{it}^2 + \boldsymbol{\beta}_t'\mathbf{M}_p^{-1}\boldsymbol{\beta}_t\right)\left(\frac{\boldsymbol{\beta}_t'}{v_{it}^2}\left(\frac{\boldsymbol{\beta}_t\boldsymbol{\beta}_t'}{v_{it}^2} + \frac{\hat{\mathbf{K}}'\hat{\mathbf{K}}}{\sigma^2} + \mathbf{I}\right)^{-1}\frac{\hat{\mathbf{K}}'\mathbf{X}_i}{\sigma^2}\right) \\
&= \left(v_{it}^2 + \mathbf{B}_t'\mathbf{B}_t\right)\left(\frac{\boldsymbol{\beta}_t'}{v_{it}^2}\left(\frac{\boldsymbol{\beta}_t\boldsymbol{\beta}_t'}{v_{it}^2} + \mathbf{M}_p\right)^{-1}\frac{\hat{\mathbf{K}}'\mathbf{X}_i}{\sigma^2}\right) \\
&= \left(v_{it}^2 + \mathbf{B}_t'\mathbf{B}_t\right)\left(\frac{\boldsymbol{\beta}_t'}{v_{it}^2}\left(\mathbf{M}_p^{\frac{1}{2}}\left(\frac{\mathbf{M}_p^{-\frac{1}{2}}\boldsymbol{\beta}_t\boldsymbol{\beta}_t'\mathbf{M}_p^{-\frac{1}{2}}}{v_{it}^2} + \mathbf{I}\right)\mathbf{M}_p^{\frac{1}{2}}\right)^{-1}\frac{\hat{\mathbf{K}}'\mathbf{X}_i}{\sigma^2}\right) \\
&= \left(v_{it}^2 + \mathbf{B}_t'\mathbf{B}_t\right)\frac{\mathbf{B}_t'}{v_{it}^2}\left(\frac{\mathbf{B}_t\mathbf{B}_t'}{v_{it}^2} + \mathbf{I}\right)^{-1}\mathbf{M}_p^{-\frac{1}{2}}\frac{\hat{\mathbf{K}}'\mathbf{X}_i}{\sigma^2} \\
&= \left(\mathbf{B}_t' + \frac{\mathbf{B}_t'\mathbf{B}_t\mathbf{B}_t'}{v_{it}^2}\right)\left(\frac{\mathbf{B}_t\mathbf{B}_t'}{v_{it}^2} + \mathbf{I}\right)^{-1}\mathbf{M}_p^{-\frac{1}{2}}\frac{\hat{\mathbf{K}}'\mathbf{X}_i}{\sigma^2} \\
&= \mathbf{B}_t'\left(\mathbf{I} + \frac{\mathbf{B}_t\mathbf{B}_t'}{v_{it}^2}\right)\left(\frac{\mathbf{B}_t\mathbf{B}_t'}{v_{it}^2} + \mathbf{I}\right)^{-1}\mathbf{M}_p^{-\frac{1}{2}}\frac{\hat{\mathbf{K}}'\mathbf{X}_i}{\sigma^2} \\
&= \mathbf{B}_t'\mathbf{M}_p^{-\frac{1}{2}}\frac{\hat{\mathbf{K}}'\mathbf{X}_i}{\sigma^2} \\
&= \frac{\boldsymbol{\beta}_t'\mathbf{M}_p^{-1}\hat{\mathbf{K}}'\mathbf{X}_i}{\sigma^2}
\end{aligned}
$$

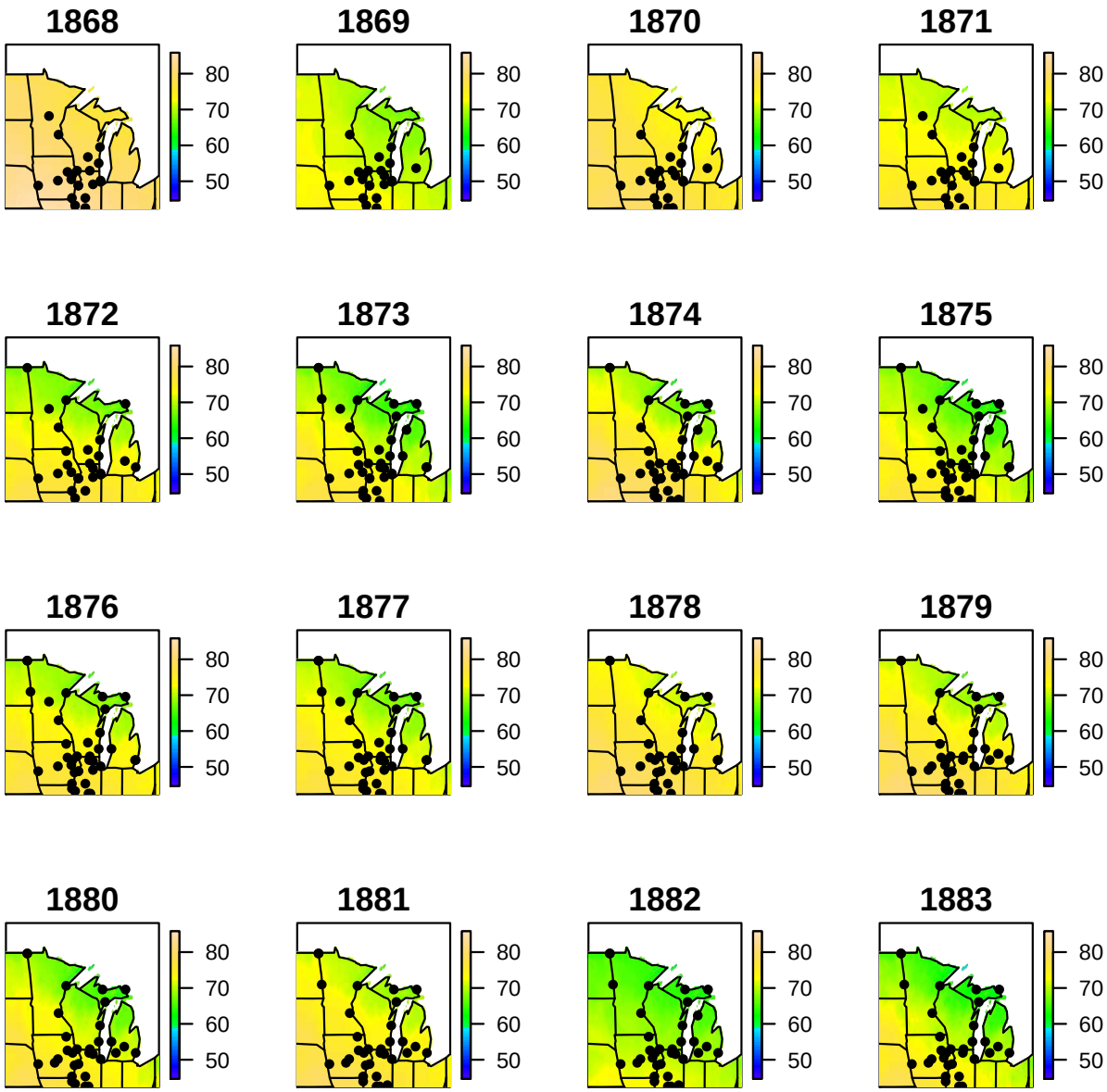
$$
y_{it}|\boldsymbol{\beta}_t, v_{it}^2, \sigma \sim \mathrm{N}\left(\frac{\mathbf{X}_i'\hat{\mathbf{K}}\mathbf{M}_p^{-1}\boldsymbol{\beta}_t}{\sigma^2}, v_{it}^2 + \boldsymbol{\beta}_t'\mathbf{M}_p^{-1}\boldsymbol{\beta}_t\right).
$$

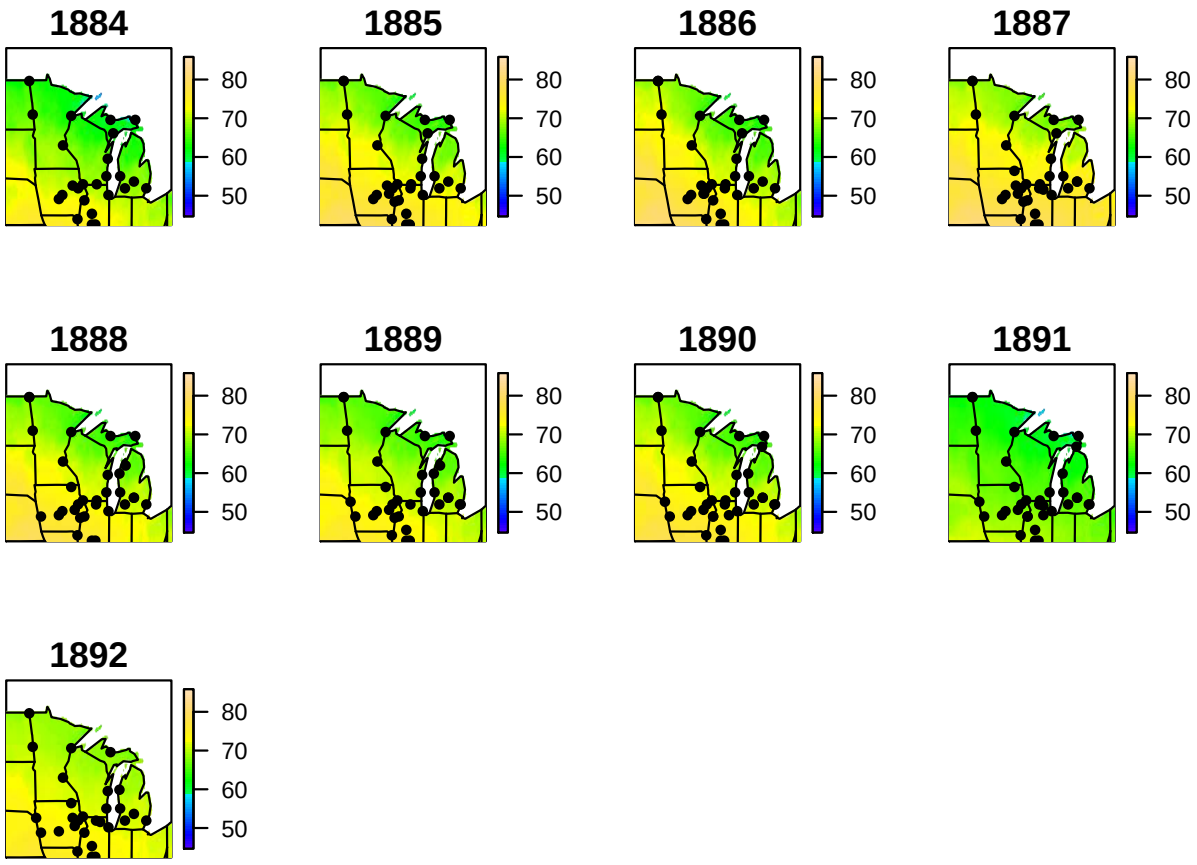
B.2. Robust PCR Model Posterior Mean July Temperature



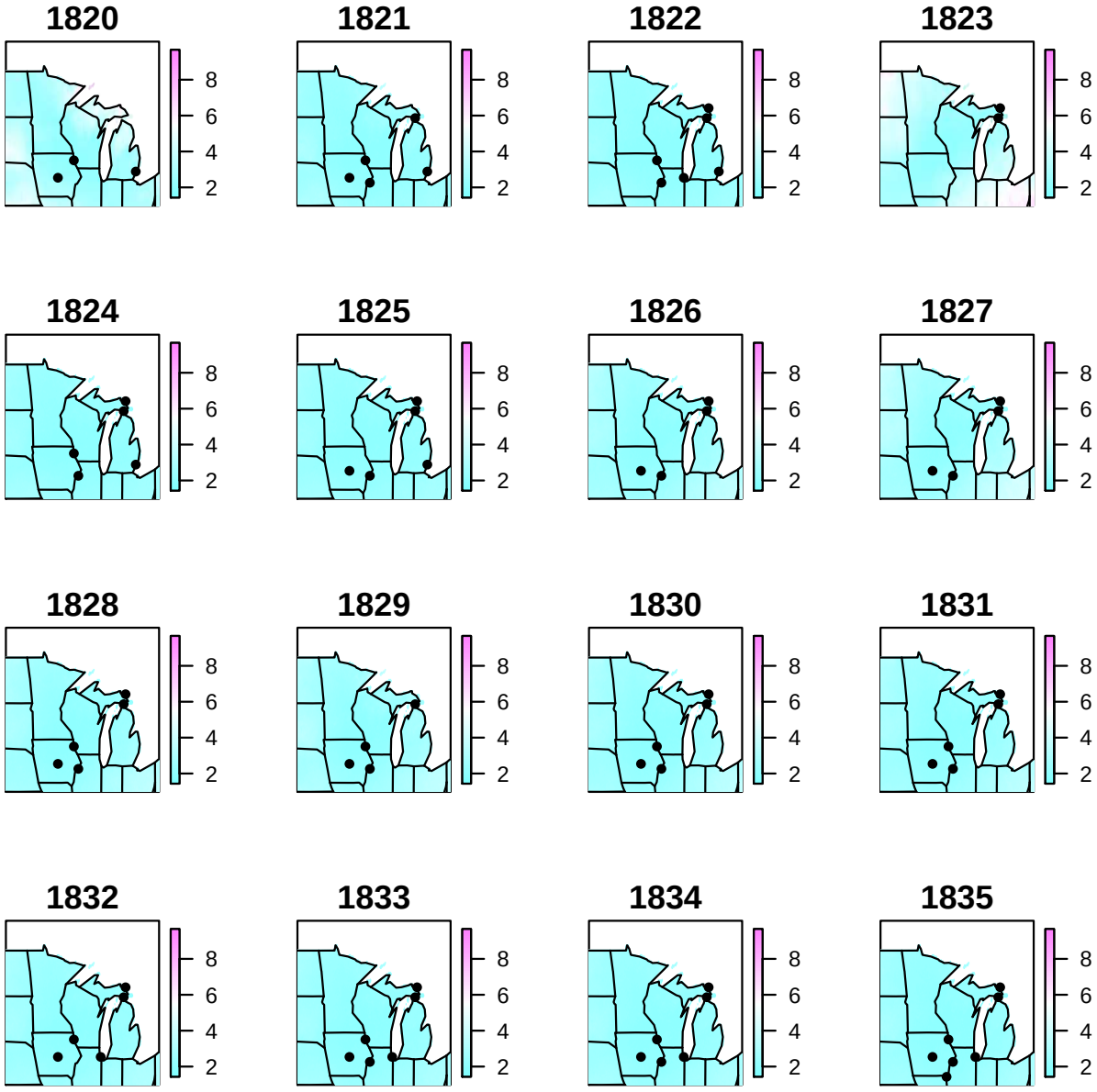

1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
80
70
60
50

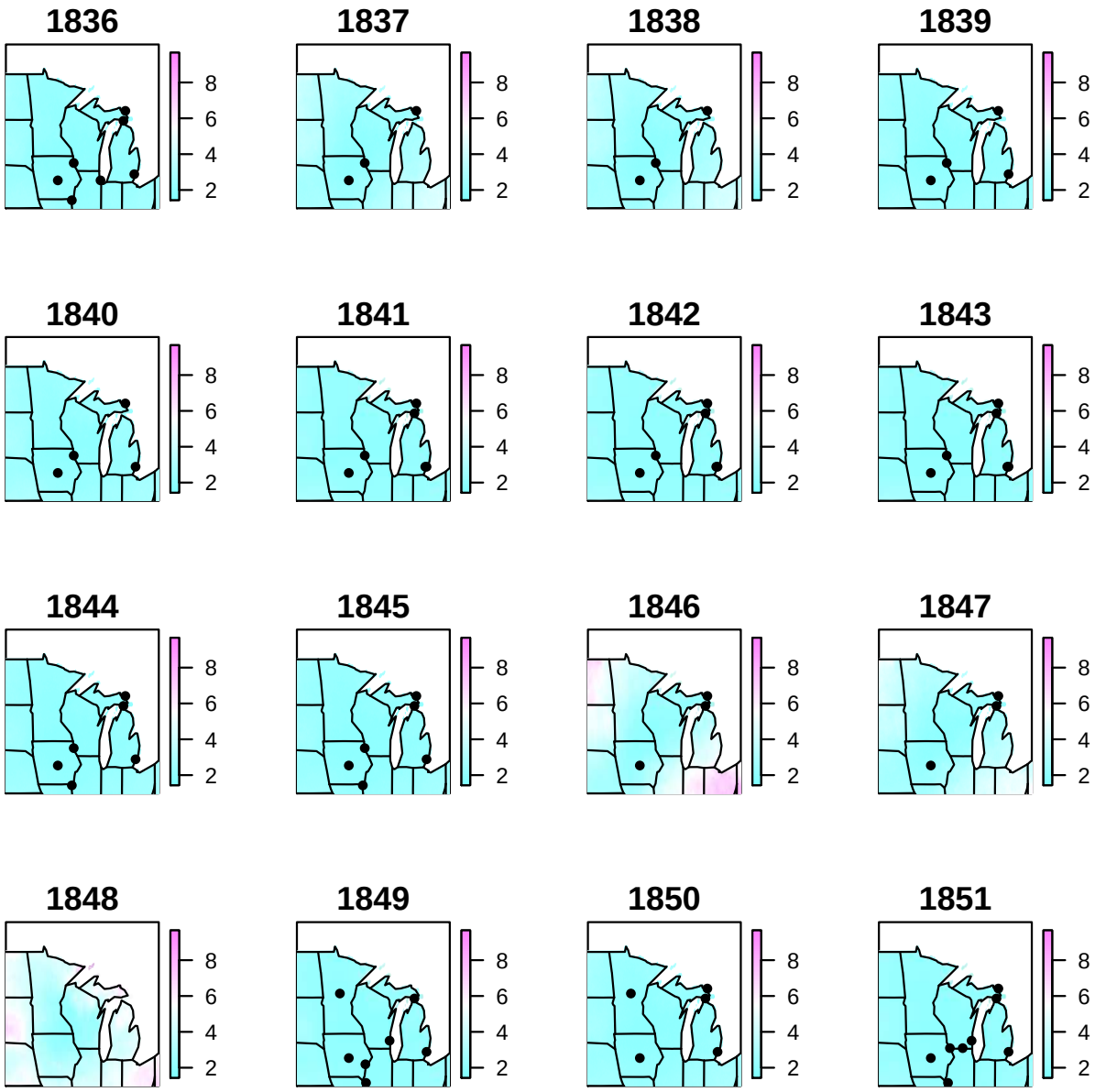

1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
80
70
60
50

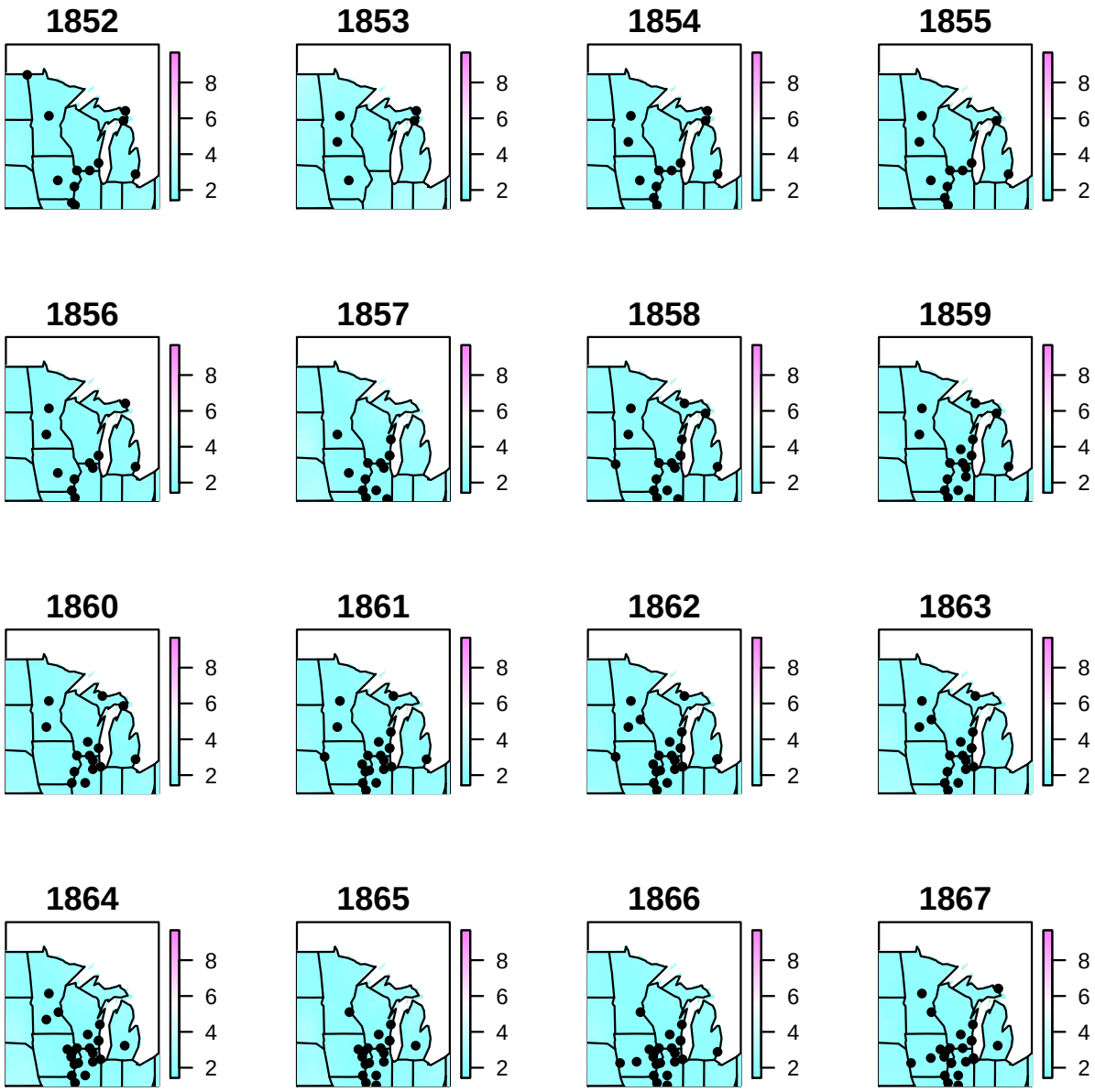

1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
80
70
60
50

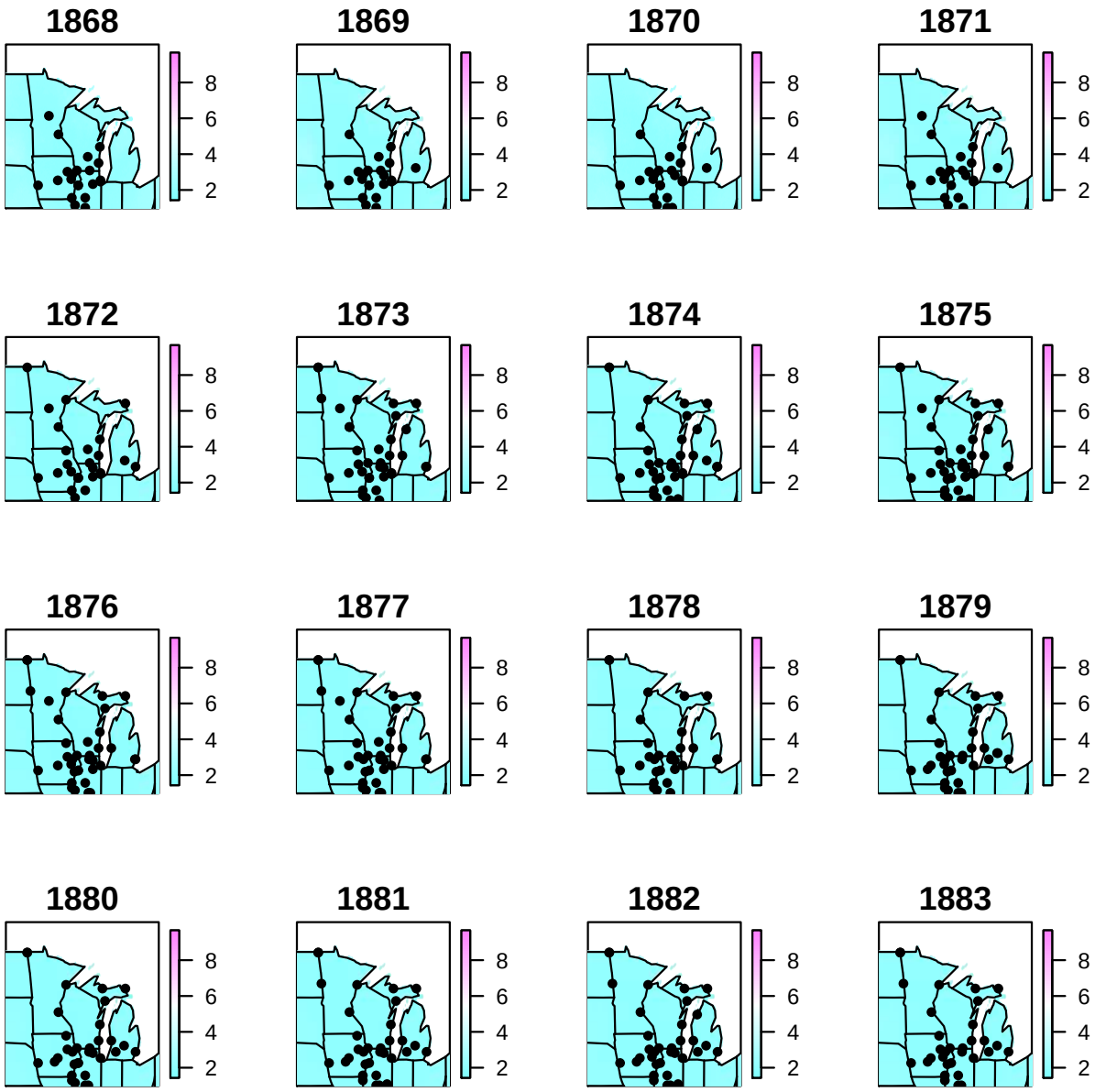

1884
80
70
60
50
1885
80
70
60
50
1886
80
70
60
50
1887
80
70
60
50
1888
80
70
60
50
1889
80
70
60
50
1890
80
70
60
50
1891
80
70
60
50
1892
80
70
60
50

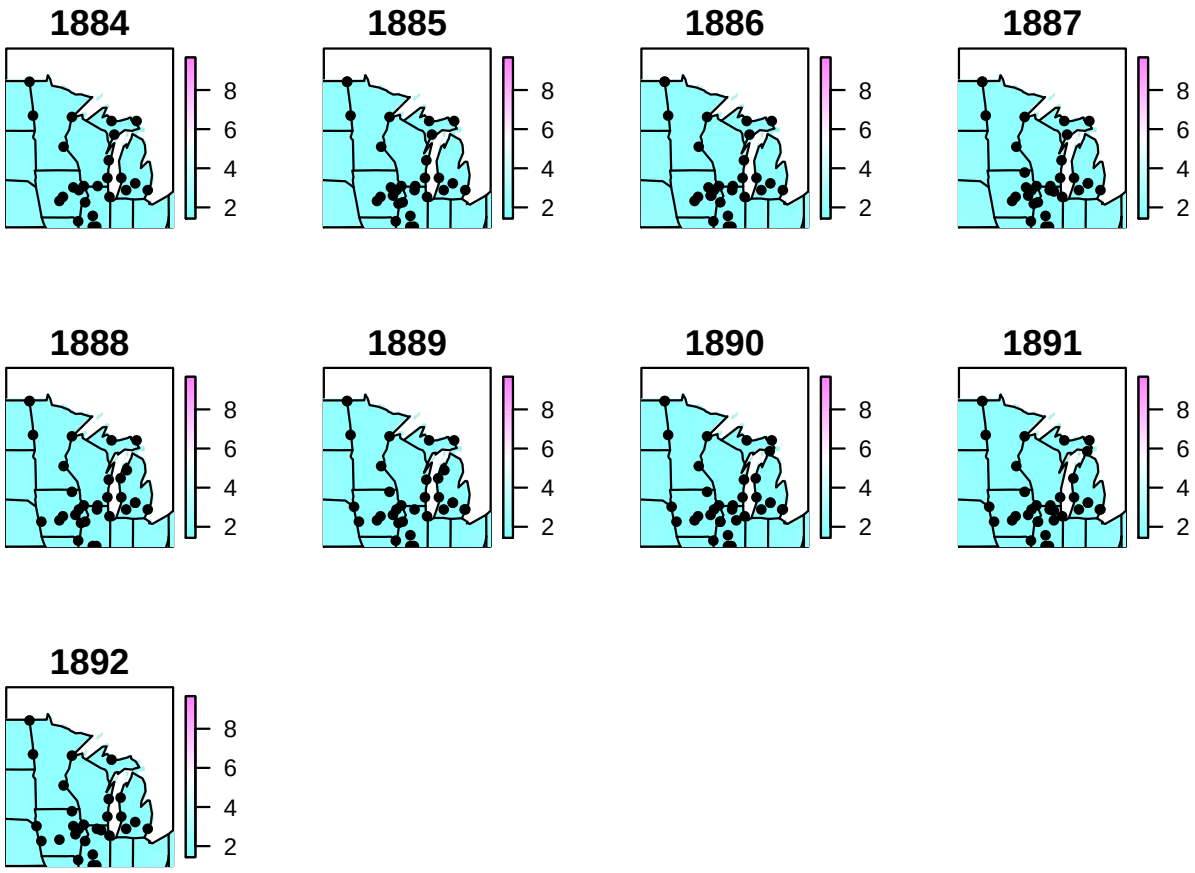
B.3. Robust PCR Model July Temperature Posterior Standard Deviations



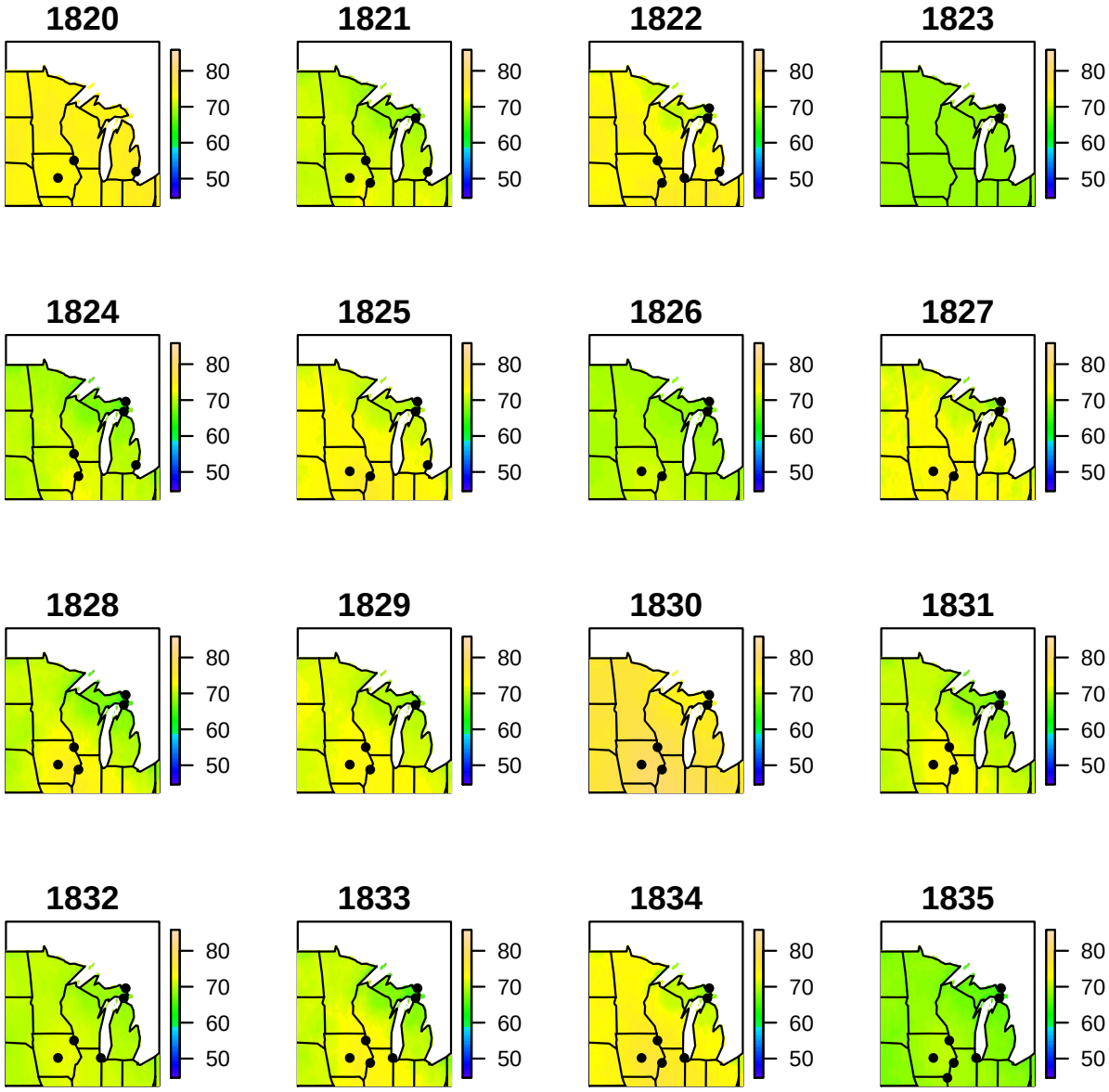

1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
8
6
4
2

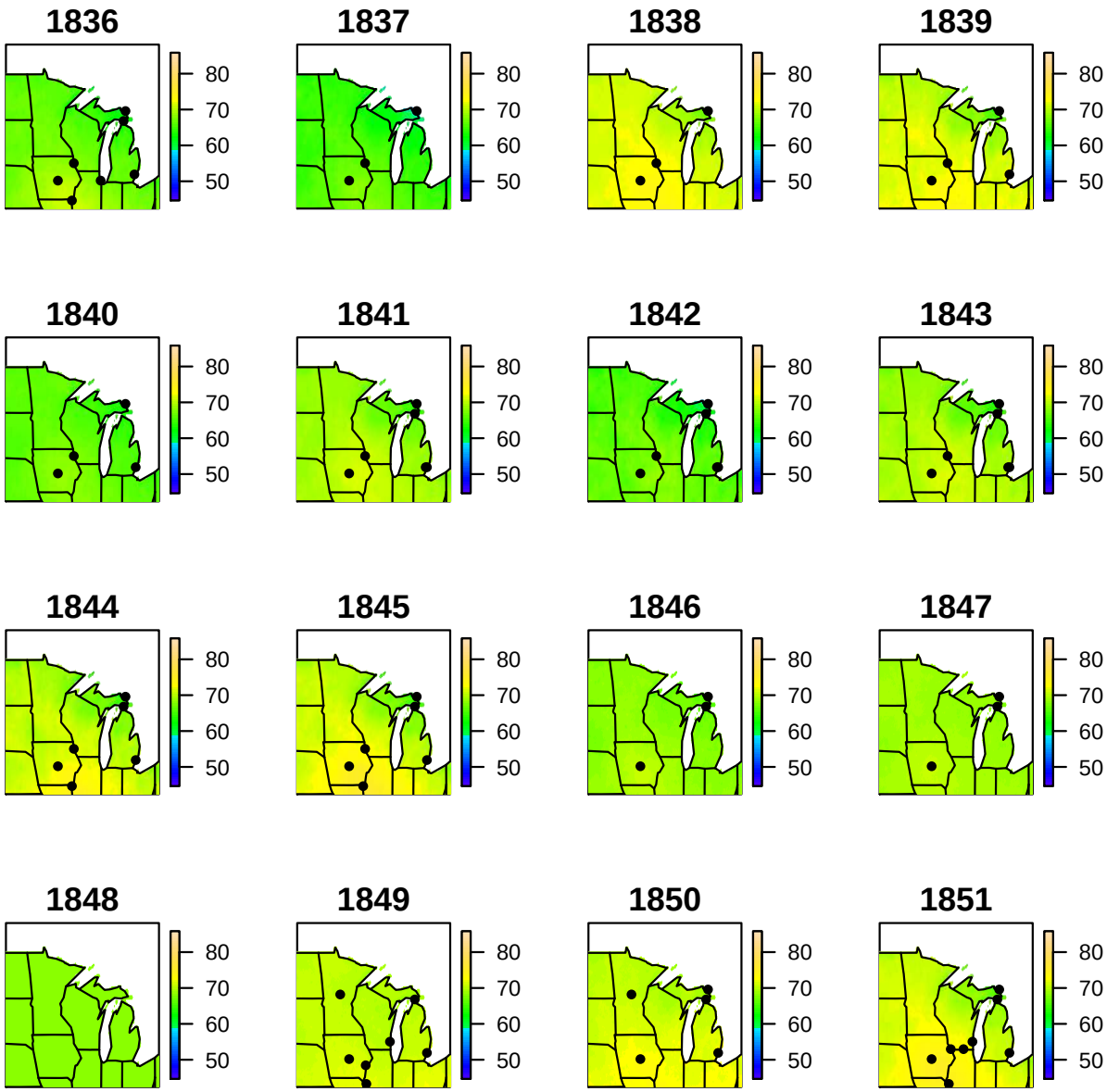

1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
8
6
4
2

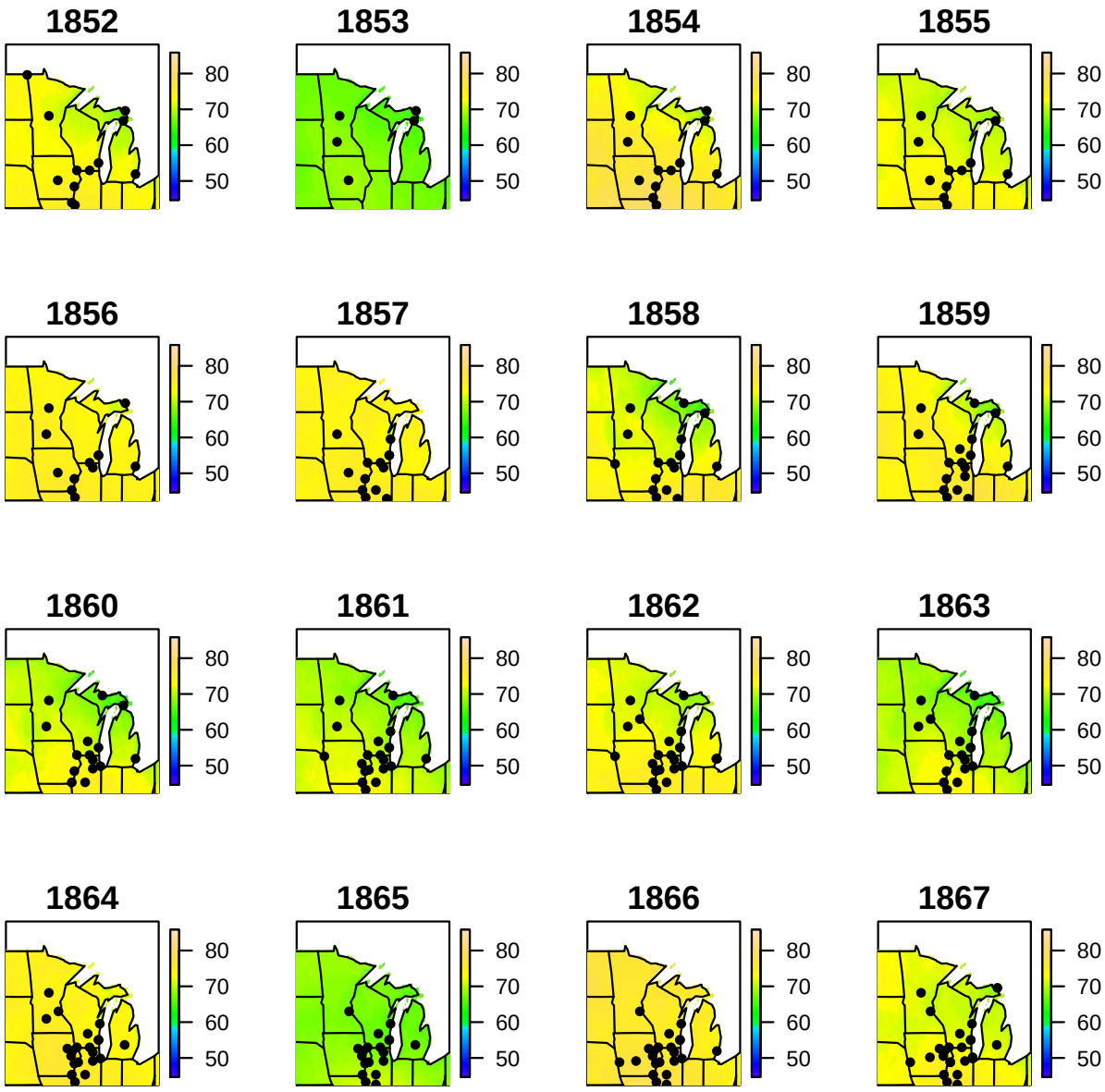

1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
8
6
4
2

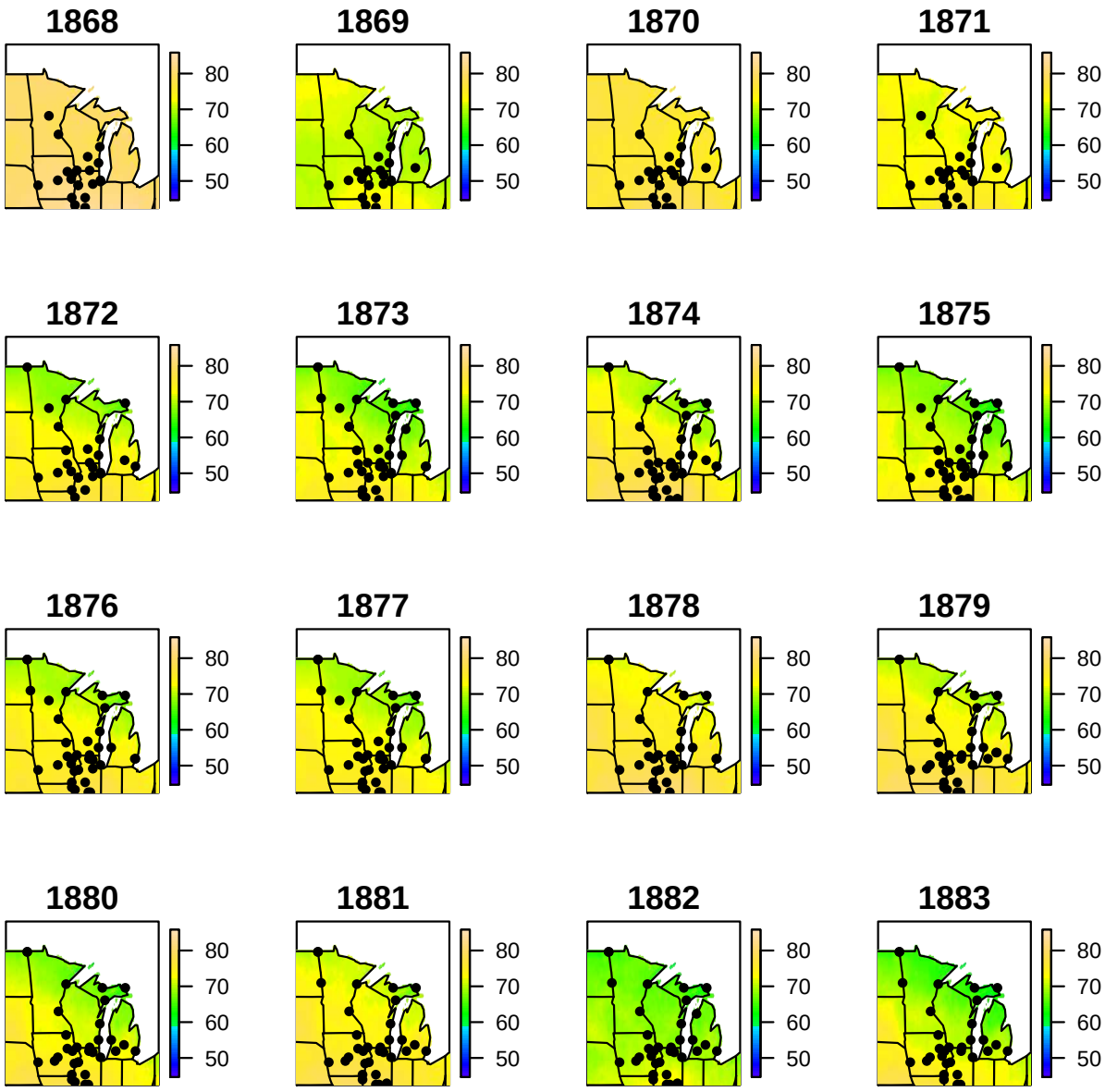

1884
8
6
4
2
1885
8
6
4
2
1886
8
6
4
2
1887
8
6
4
2
1888
8
6
4
2
1889
8
6
4
2
1890
8
6
4
2
1891
8
6
4
2
1892
8
6
4
2

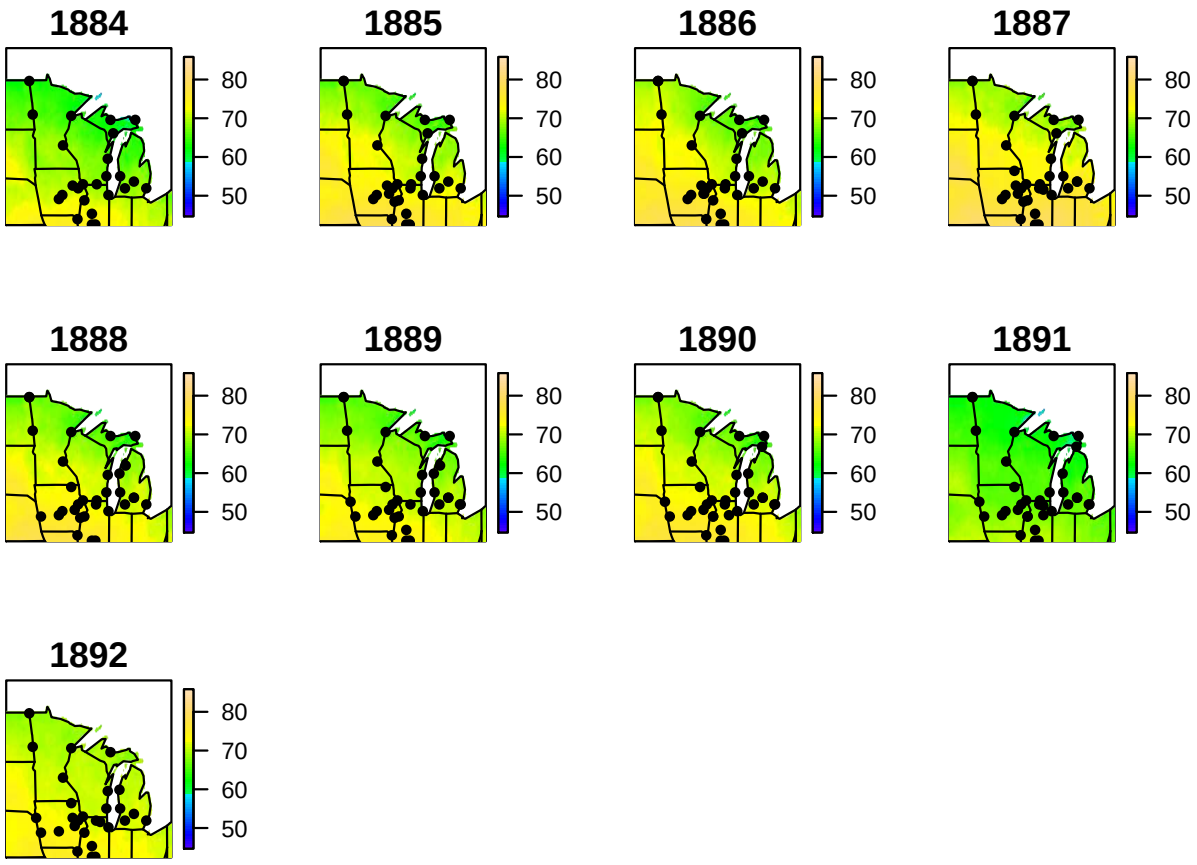
B.4. Robust Probabilistic PCR Model Posterior Mean July Temperature




1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
80
70
60
50

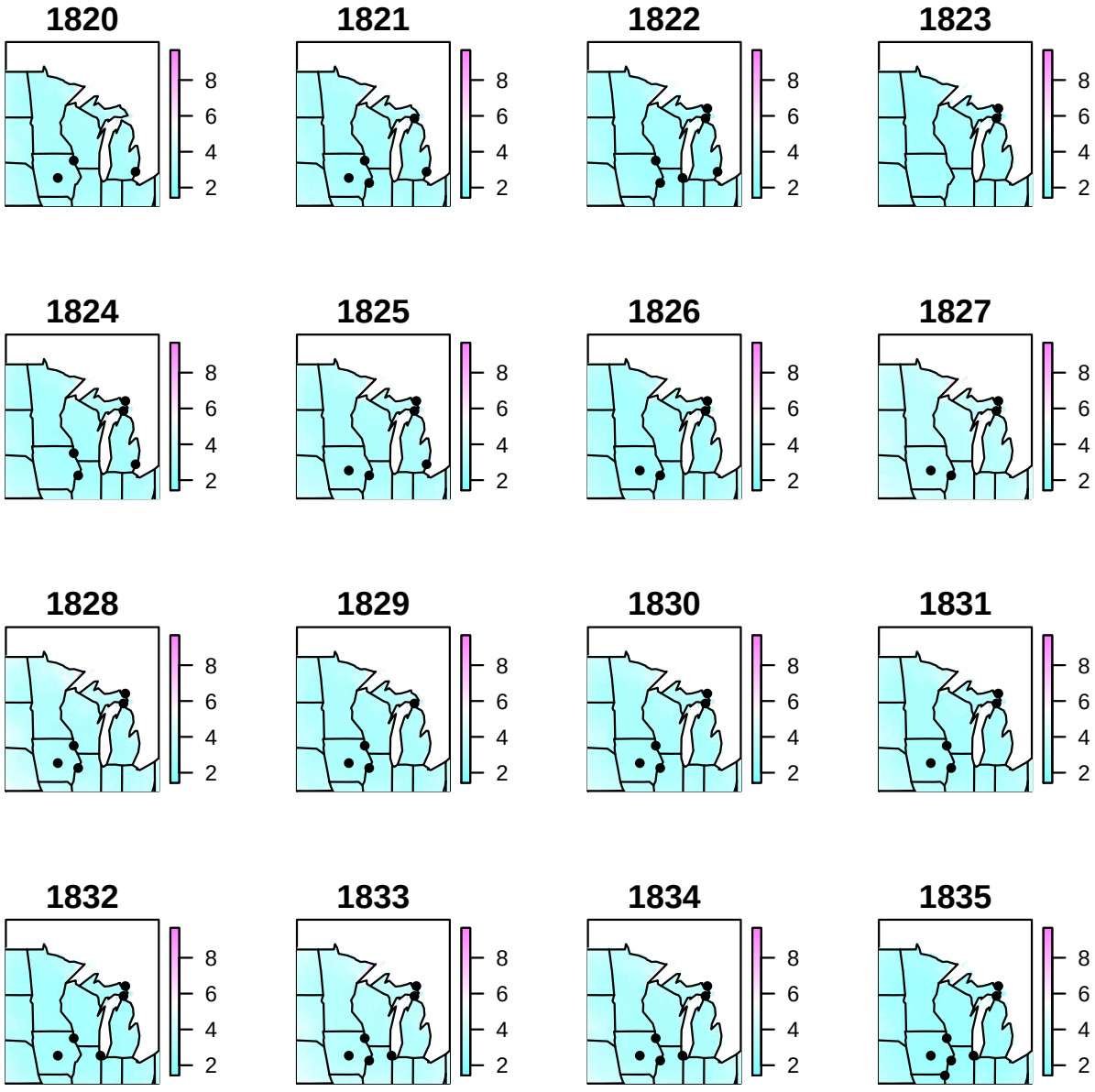

1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
80
70
60
50

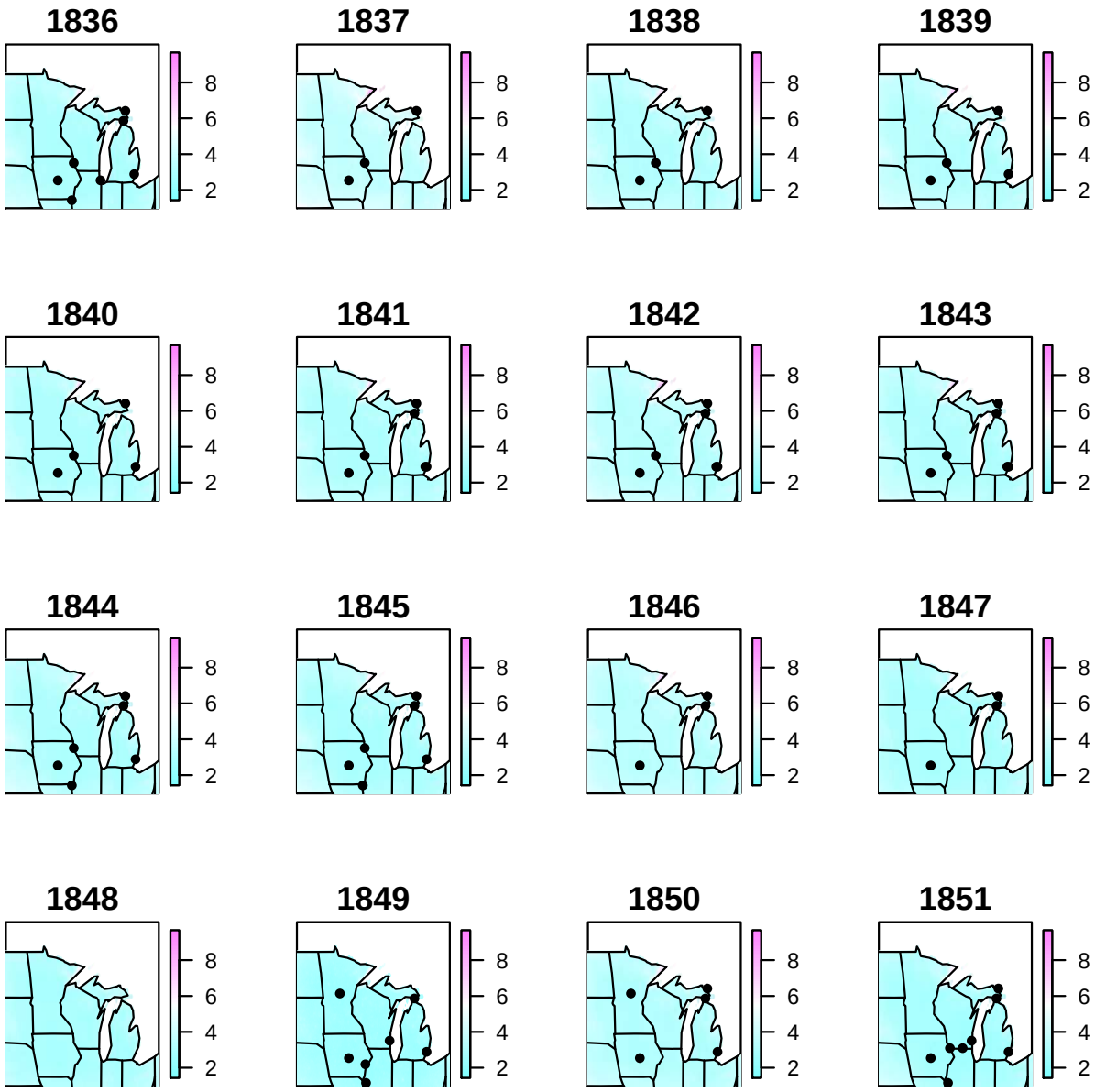

1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
80
70
60
50

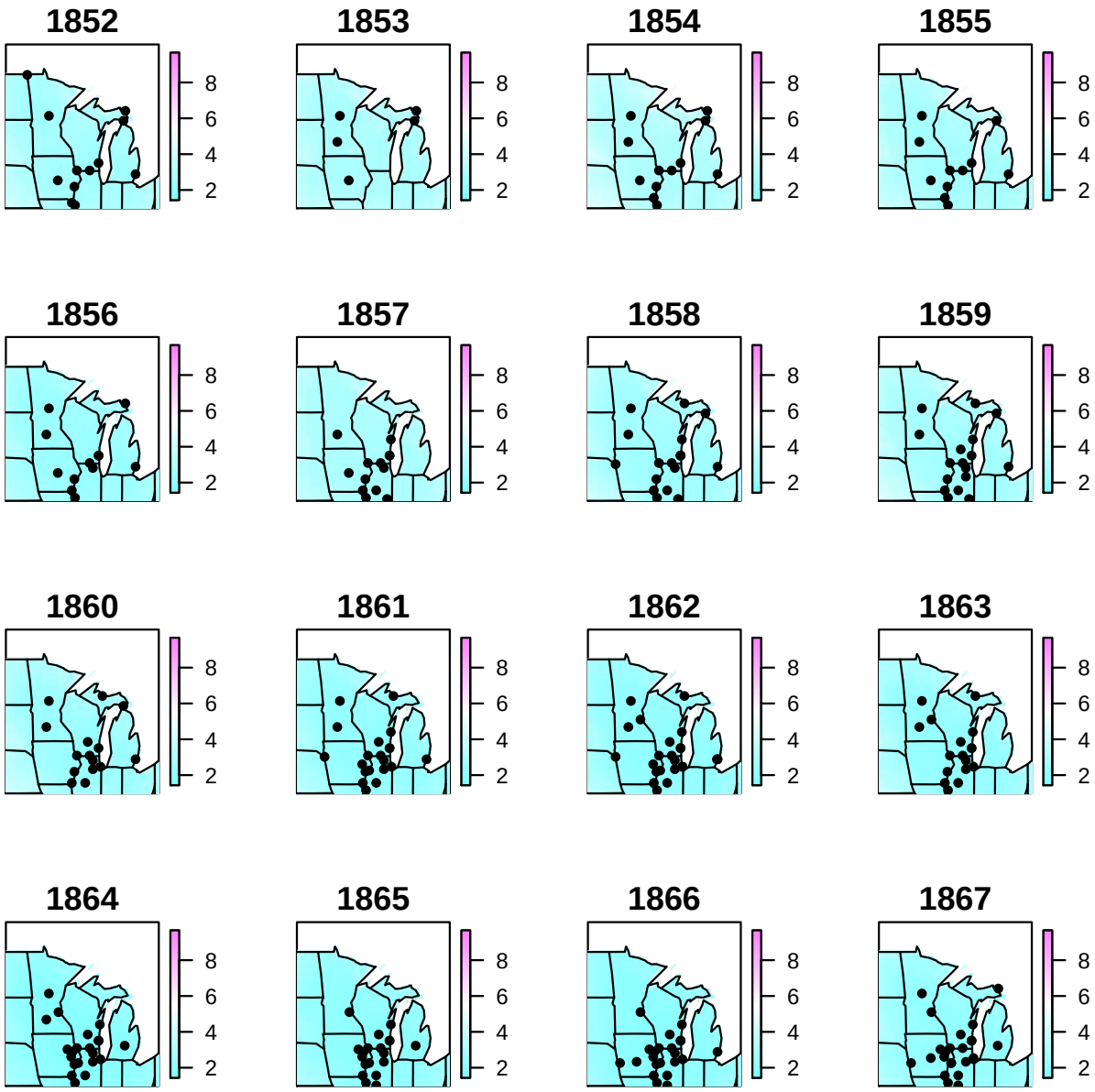

1884
80
70
60
50
1885
80
70
60
50
1886
80
70
60
50
1887
80
70
60
50
1888
80
70
60
50
1889
80
70
60
50
1890
80
70
60
50
1891
80
70
60
50
1892
80
70
60
50

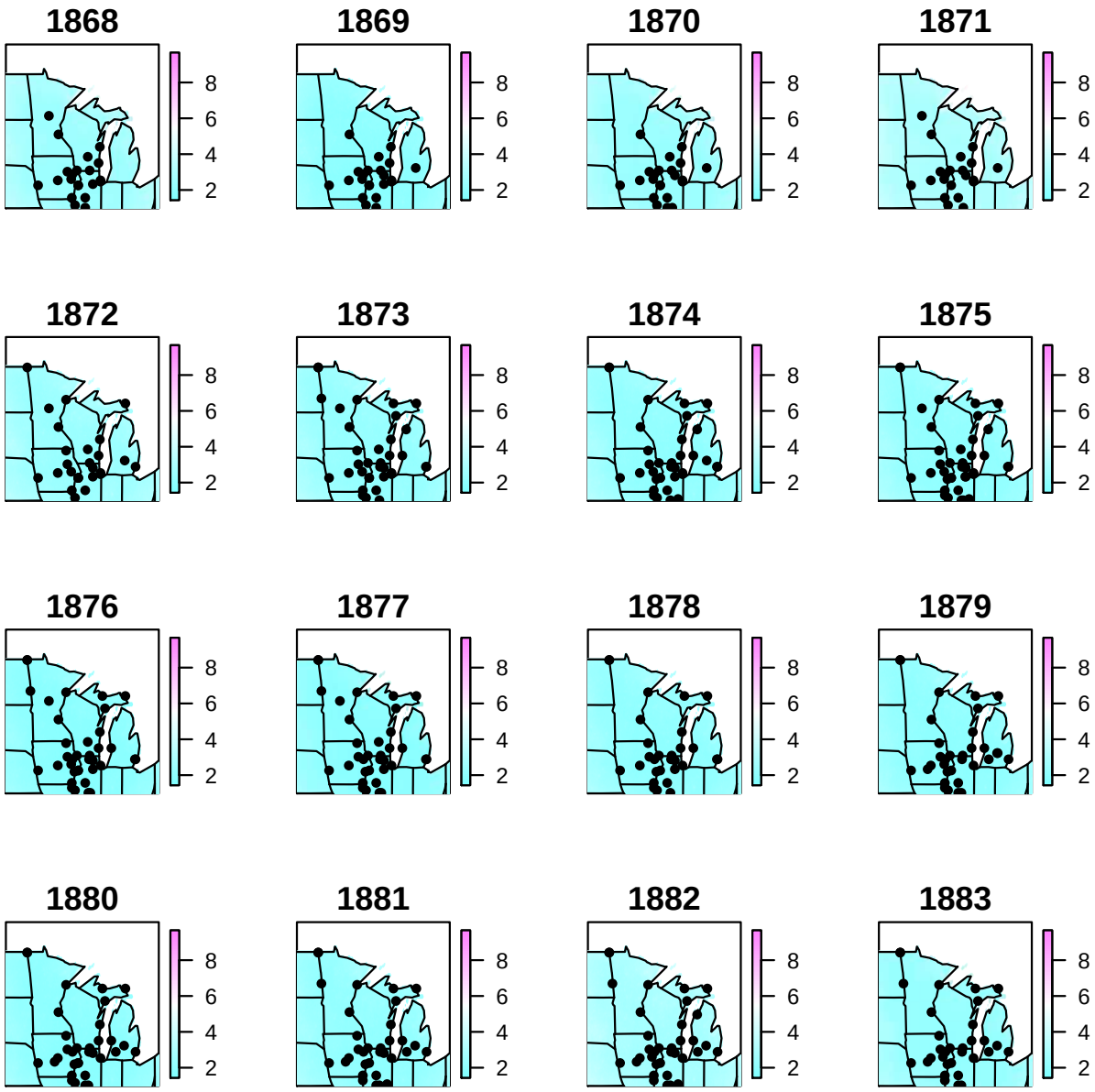
B.5. Robust Probabilistic PCR Model July Temperature Posterior

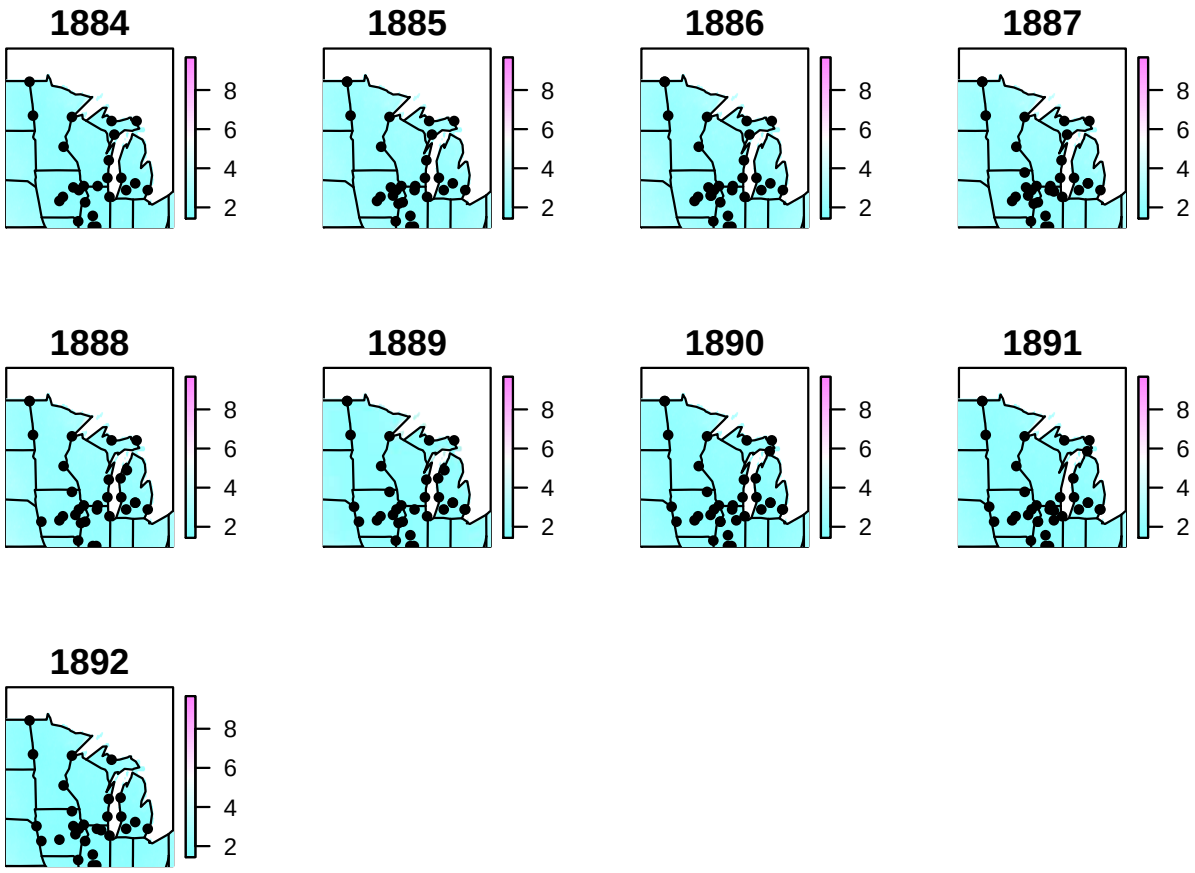
Standard Deviations




1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
8
6
4
2


1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
8
6
4
2


1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
8
6
4
2


1884
8
6
4
2
1885
8
6
4
2
1886
8
6
4
2
1887
8
6
4
2
1888
8
6
4
2
1889
8
6
4
2
1890
8
6
4
2
1891
8
6
4
2
1892
8
6
4
2