Harvard University
Harvard University Biostatistics Working Paper Series

Year 2008 *Paper* 88

Expanded Technical Report: Mapping Ancient Forests: Bayesian Inference for Spatio-temporal Trends in Forest Composition Using the Fossil Pollen Proxy Record

Christopher J. Paciorek* Jason S. McLachlan†

*Harvard School of Public Health, paciorek@hsph.harvard.edu

†University of Notre Dame

Expanded Technical Report: Mapping Ancient Forests: Bayesian Inference for Spatio-temporal Trends in Forest Composition Using the Fossil Pollen Proxy Record

August 19, 2008

Christopher J. Paciorek
Department of Biostatistics, Harvard School of Public Health
and
Jason S. McLachlan
Department of Biology, University of Notre Dame

Abstract

Ecologists use the relative abundance of fossil pollen in sediments to estimate how tree species abundances change over space and time. To predict historical forest composition and quantify the available information, we build a Bayesian hierarchical model of forest composition in central New England, USA, based on pollen in a network of ponds. The critical relationships between abundances of taxa in the pollen record and abundances as actual vegetation are estimated for the modern and colonial periods, for which both pollen and direct vegetation data are available, based on a latent multivariate spatial process representing forest composition. For time periods in the past with only pollen data, we use the estimated model parameters to constrain predictions about the latent spatio-temporal process conditional on the pollen data. We develop an innovative graphical assessment of feature significance to help to infer which spatial patterns are reliably estimated. The model allows us to estimate the spatial distribution and relative abundances of tree species over the last 2500 years, with an assessment of uncertainty, and to draw inference about how these patterns have changed over time. Cross-validation suggests that our feature significance approach can reliably indicate certain large-scale spatial features for many taxa, but that features on scales smaller than 50 km are difficult to distinguish, as are large-scale features for some taxa. We also use the model to quantitatively investigate ecological hypotheses, including covariate effects on taxa abundances and questions about pollen dispersal characteristics. The critical advantages of our modeling approach over current ecological analyses are the explicit spatio-temporal representation, quantification of abundance on the scale of trees rather than pollen, and uncertainty characterization.

keywords: Dirichlet-multinomial, Gaussian process, paleoecology, radial basis functions, smoothing, spatial statistics

1 Introduction

Scientific inference about forest composition in the past relies heavily on sediment records of fossil pollen taken from ponds and other depositional environments (Davis 1981; Delcourt and Delcourt 1987). Fossil pollen collected from multiple sites over time acts as a proxy for the abundance of different tree taxa (species or genera), telling us about spatio-temporal vegetation dynamics over thousands of years. Paleoecologists ask questions such as the following. How have relative population abundances and range boundaries changed over time? Do stable assemblages of species exist for long periods of time or are forest compositions constantly shifting? How have forest communities changed in response to past climate shifts and what can forest composition tell us about climate? Practical environmental questions relate to how human manipulation of forests compares to natural forest change.

However, inferring tree abundance on the landscape from pollen abundance in sediments is not straightforward, because the relationship between the relative abundance of trees near a pond and pollen in the sediment of that pond is not simple. Different tree species produce different amounts of pollen on average (Jackson 1990), and the representation of any individual tree in deposited pollen is a complex function of distance to the deposition basin, size of the deposition basin, landscape openness, forest structure, wind regime, and preservation in sediments (Prentice 1985; Jackson and Lyford 1999; Nielsen and Sugita 2005). Our understanding of the timing of sediment deposition depends on indirect and inexact measurements of sediment age (through radiometric dating and stratigraphic markers). The aggregate effect of these sources of uncertainty is pollen assemblages that are noisy reflections of the trees in the surrounding landscape.

Because of these uncertainties, most paleoecological studies do not attempt to make explicit inference about the distribution of trees based on fossil pollen data. Instead, they assume that robust changes in pollen abundances over space and time generally correspond to changes in vegetation at the scales described above, primarily using multivariate time series at one or more sites (Fuller et al. 1998). Efforts to explicitly correct for differential pollen production across taxa range from primarily statistical (Tauber 1965; Prentice et al. 1987) to more mechanistic approaches (Bunting and Middleton 2005). These studies highlight the difficulty of inferring a complicated spatial pattern of pollen source contributions across the landscape from pollen proportions in a single deposition site. The power to disentangle this spatial signal using a network of sites was explored by Webb (1974) and Sugita (1993, 1994, 2007a,b). Our work provides a statistical framework to estimate the signal and quantify the uncertainty in this process based on a spatial network of noisy data.

Building on recent work in Bayesian spatio-temporal statistics (e.g., Wikle et al. 2001; Banerjee et al. 2004; Fuentes and Raftery 2005; Royle and Wikle 2005; Gelfand et al. 2006; Haslett et al. 2006), we develop an approach for modelling forest composition based on vegetation data from two key time points and pollen data from sediment cores, using a multivariate latent spatio-temporal process representing the relative abundances of different taxa. The model allows inference across space and time, based on modeling the relationship between forest composition and pollen composition for locations and times at which both vegetation and pollen data are available. Assuming consistency in the relationships over time, the model then predicts vegetation in the past using proxy pollen

data. The statistical challenges are in computationally-efficient and sufficiently-resolved representation of the latent spatio-temporal surfaces, modelling spatially correlated compositional data, and carefully borrowing strength across space, time and taxa. We seek to allow the pollen data to provide as much information as possible, avoiding oversmoothing, while constraining the model sufficiently to achieve reasonable prediction that accounts for bias and noisiness. Finally, this high-dimensional model must be fit; MCMC in such situations is often time-consuming and prone to mixing difficulties (Knorr-Held and Rue 2002; Christensen et al. 2006; Paciorek 2007). There has been recent fruitful collaborative work between statisticians and ecologists in understanding patterns of species distributions (e.g., Hooten et al. 2003; Royle and Wikle 2005; Gelfand et al. 2006). Our work is in this tradition, but differs in its consideration of a multi-taxon spatial process and its use of proxy data to predict distributions over time, as well as through careful consideration of how to assess the significance of predicted spatial patterns.

Our analysis focuses on central New England in the northeastern United States over the past 2500 years. The network of pollen sites that we model is amongst the most dense sets of pollen data in existence and and has taken decades to produce. Our goals are both particular to this domain and quite general. In particular, we first want to understand the relationship between the pollen record in a pond and vegetation in the surrounding area. Second, we want to estimate and compare vegetation in our space-time domain in the colonial and modern eras. Third, our key application goal is to predict, and quantify uncertainty in, spatio-temporal patterns in tree abundances over the past 2500 years. More generally, we want to explore the ability of the pollen record to inform vegetation composition and dynamics spatially and temporally and create a modelling infrastructure useful in different areas and time periods.

Section 2 describes the pollen and vegetation data available from central New England. In Section 3 we build an estimation model to calibrate pollen to vegetation at times at which both types of data are available and then present a prediction model that uses parameter estimates from the estimation model to make predictions when only pollen data are available. We assess the model, considering the consistency and strength of the association between the proxy pollen composition and forest vegetation composition, as well as using cross-validation, and then use the model for prediction over the past 2500 years (Section 4). We also introduce innovative graphics that take advantage of the rich information in the posterior samples to assess a variety of contrasts of interest. The discussion in Section 5 highlights the contributions of the modeling approach to the ecological problem. Additional ecological analysis of model results is currently underway and will be presented in the ecological literature.

2 Data

2.1 Study area and study taxa

Our study area extends from $43^\circ,\ 21'$ N, $73^\circ,\ 30'$ W in the northwest to $41^\circ,\ 37'$ N, $71^\circ,\ 13'$ W in the southeast corner in south-central New England, USA, focusing on central and western Massachusetts, west of the Boston metropolitan area. In projected coordinates,

this defines a region, 192×192 km^2, which for computational reasons we divide into a 16 by 16 grid, with each grid cell 12 km on a side. All computations are done at the resolution of the grid cell.

We focus on 9 particular taxa (genus or species), including the most common taxa in the area: oak (*Quercus spp.*), pine (*Pinus spp.*), maple (*Acer spp.*), hemlock (*Tsuga canadensis*), and beech (*Fagus grandifolia*); as well as several additional taxa of particular interest, namely hickory (*Carya spp.*), birch (*Betula spp.*), spruce (*Picea spp.*), and chestnut (*Castanea dentata*). Other tree taxa, excluding taxa that are primarily shrubs and small trees, are grouped into a tenth category and included in the analysis as a tenth reference 'taxon'. Note that due to chestnut blight there have been essentially no adult pollen-producing chestnut in the study area in the last 100 years, so many of our figures omit chestnut.

2.2 Pollen data

Plant pollen from trees, shrubs and herbaceous plants falls on the surfaces of ponds, sinks to the bottom, and accumulates in sediment. Over time, these sediments are buried by layers from successive time periods, creating a sediment record of what fell into the pond. The scale of vegetation that corresponds most closely to the composition of pollen in the sediments of small ponds and depositional environments is generally vegetation within one to two km (Jackson 1990; Jackson and Lyford 1999; Nielsen and Sugita 2005), but more than half of the total pollen in those sediments may originate beyond that distance (Sugita 1994; Sugita et al. 1998; Sugita 2007b).

As described below, the period of colonial settlement and the modern period are times when vegetation and pollen data can be compared. While separated only by a few hundred years, these periods are likely to be as disparate in forest structure and composition as any two time periods considered, because of the drastic ecological effects of post-settlement land use (Foster et al. 1998; Fuller et al. 1998; Oswald et al. 2007). Colonial era surveys provide historical vegetation data, but the surveys occurred at different times in different parts of the study region, generally 1650-1700 in the eastern part of our region and in the Connecticut River valley and approximately 1700-1800 in the hill towns in the western part of our region. Therefore, we used the appearance of agricultural weed pollen to select pollen samples (~500 grains) at individual times from 23 ponds with archived sediment cores to best match the time at which the survey in the township encompassing each pond was completed (Fig. 1a). Because settlement occurred over a period of time, the colonial era data do not represent a fixed snapshot in time, but rather a reasonably consistent window within the settlement process, stretching over the years ca. 1635-1800. Because of the long lifespans of trees and the relatively quick settlement, we consider this treatment of the colonial data to be reasonable. For the modern era we use surface sediment samples to best represent current vegetation, taken from 38 ponds (Fig. 1b).

To make predictions back in time, we make use of the full archived sediment cores taken from the 23 ponds. The temporal coverage varies, with ponds having records of length varying between 1000 and 15000 years (Fig. 2), with 2500 years the full period of interest here. Each core is divided into intervals and approximately 500 grains from a sample of sediment in each interval are identified and counted. A subset of samples is dated using radiocarbon dating, with linear interpolation providing dates for all samples, resulting

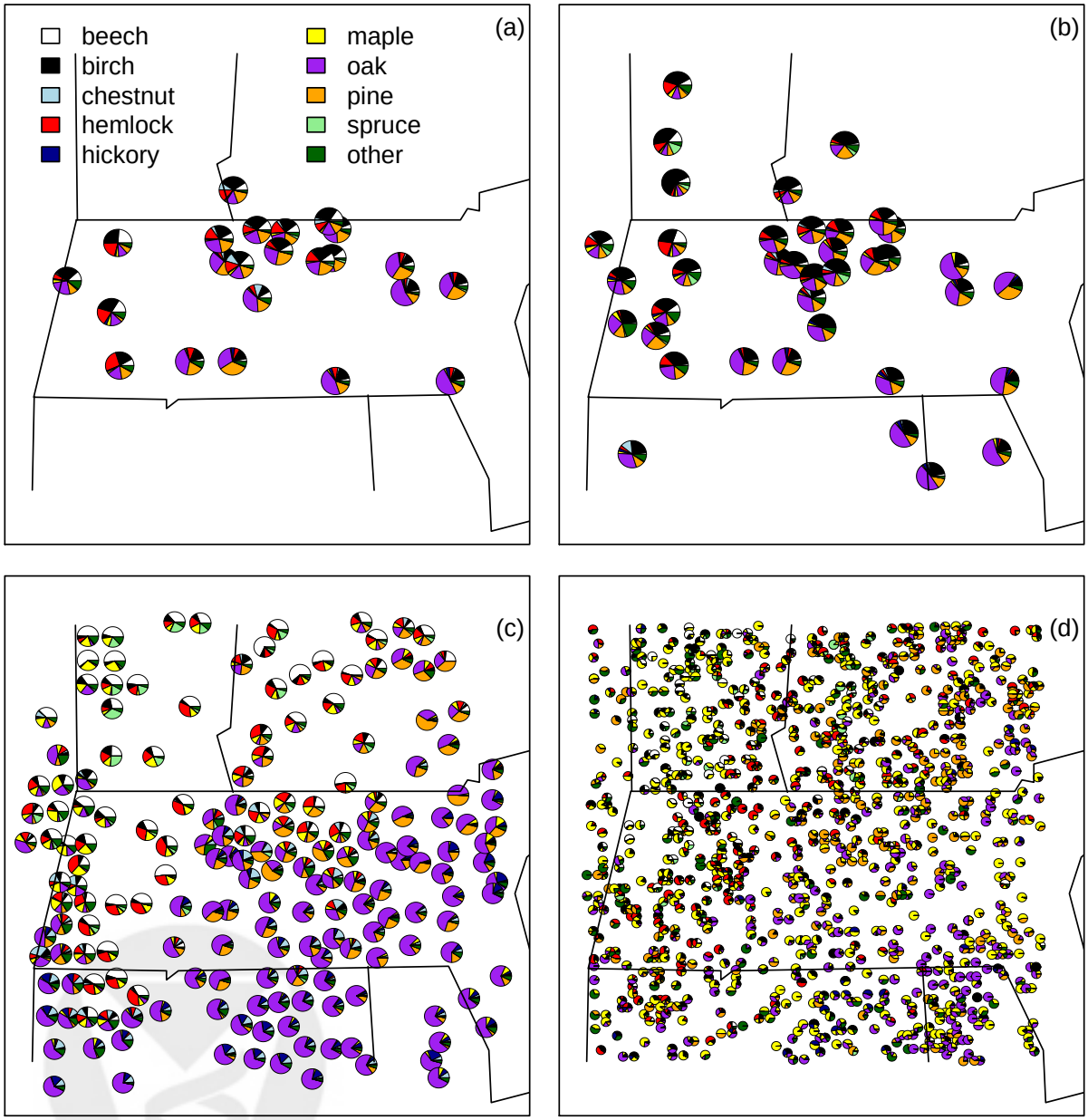

Figure 1. (a) Pollen composition by pond for the colonial era. (b) Pollen composition by pond for the modern era. (c) Witness tree vegetation composition for the colonial era (plotted at the centroids of colonial townships). (d) Forest service plot vegetation composition for the modern era.

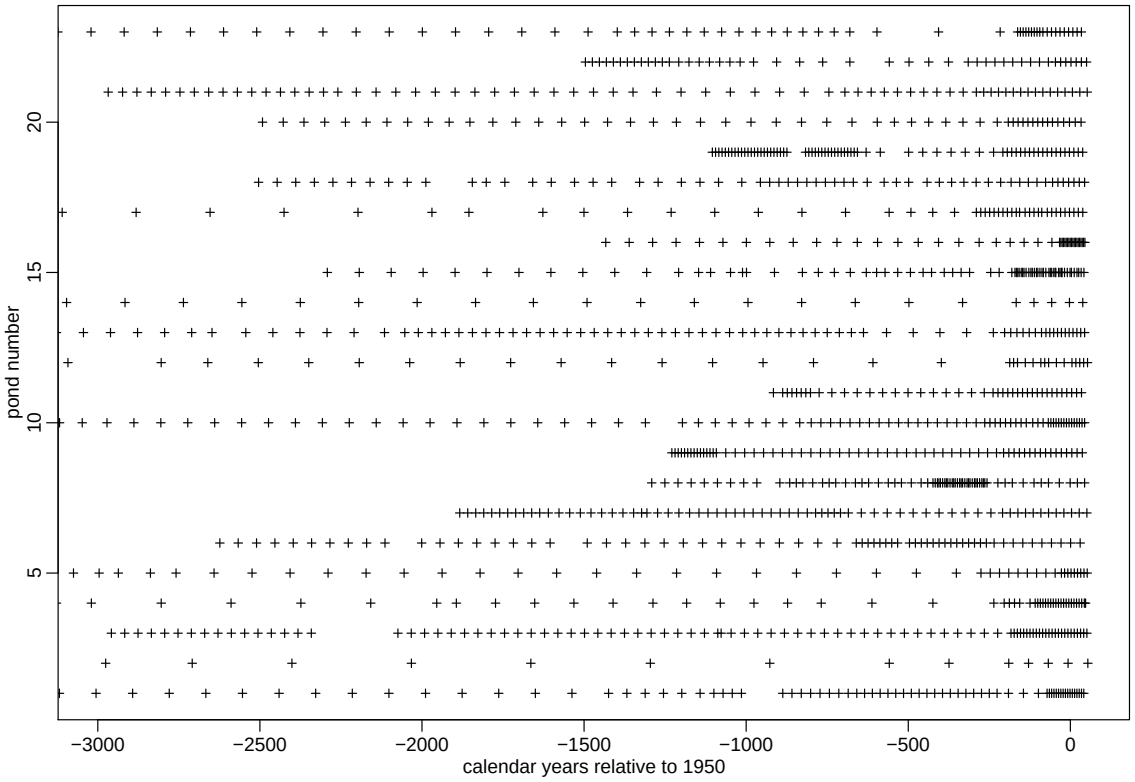

Figure 2. Sampling points over time for each of the 23 ponds, truncated at 3000 years before present.

in samples at different and irregular times for each pond. Some sediment mixing occurs in upper sediments, so any individual sample represents pollen deposited over a period of years, naturally smoothing the data. The long lifespan of trees also causes smoothness. Accordingly, in our spatio-temporal predictions, we aggregate all samples into intervals of 100 calendar years and base the prediction model on this set of discrete times. The first interval is centered on 1950 (defined to be the 'present' or year 0 in the paleoecological dating scheme) and the last is centered on 550 B.C. (denoted henceforth as year -2500).

2.3 Uncertainty in dating pollen samples

There are three primary sources of uncertainty in the dating of samples. First, dates earlier than ca. 300 years before present are based on measurements of ^{14}C while dates since the Industrial Revolution (the last ca. 150 years) are dated using Pb-210 chronologies developed for each core; for both these methods there is natural stochasticity in the counts of the isotopes for each sample that is dated. Second, sedimentation rates vary over time, introducing additional uncertainty for samples whose dates are from interpolation of radiocarbon-dated samples. The natural stochasticity can be considered independent between dates and between ponds, while the interpolation error might be considered to be roughly independent between ponds unless there were spatial structure in sedimentation

rates. The third source of error is the variability in radioactive decay over time, i.e., uncertainty in the radiocarbon model for calendar time, which introduces uncertainty in the relationship between radiocarbon years and calendar years. Note the uncertainty in the calibration of radiocarbon to calendar years applies equally to all of our samples and therefore only affects the calendar year labels we apply to our results. Essentially time is distorted to the extent the calibration curve is wrong, but it is distorted equally in all of the ponds. A full analysis, such as Blaauw and Christen (2005), could model the calendar dates in a way that accounts for all sources of uncertainty in a single pond. Their work makes use of published calibration curves estimated in a Bayesian way (e.g., IntCal04; Buck et al. 2006). A final source of uncertainty is uncertainty in the assignment of calendar age to the appearance of weed pollen in the cores, which combines variability in when weed pollen appears relative to settlement and uncertainty in when settlement actually occurred in a given area.

We do not account for the dating uncertainties here, in part because of our focus on changes at time scales of several hundred years or more, for which uncertainties in dating should have limited impact. However, we believe the following strategy, which extends Blaauw and Christen (2005) to the multiple pond spatial setting and includes uncertainty in stratigraphic markers, holds promise for the dating uncertainty until the period of settlement. First elicit a prior distribution on the age of the settlement horizon in each pond's core; we expect this to be a fairly simple unimodal distribution centered around the estimates we currently use, probably with the same level of uncertainty for all the ponds. The uncertainty might be correlated for nearby ponds based on the uncertainty in when settlement occurred but would be independent to the extent it reflects very fine-scale heterogeneity in when grass pollen builds up in a pond. Next, repeat the following steps many times. First, draw a sample calibration curve from the published calibration curve including uncertainty (e.g. IntCal04). Draw a calendar age from the settlement horizon prior distribution for each pond and convert to radiocarbon ages using the sample calibration curve. Using this radiocarbon age as an age with no uncertainty, use the model of Blaauw and Christen (2005) to derive the posterior in radiocarbon years for each pond separately, thereby accounting for the stochasticity in the isotope counts and interpolation uncertainty, which might be considered approximately independent between ponds. Finally convert all the radiocarbon ages for all ponds to calendar age using the sample calibration curve, thereby accounting for the shared uncertainty in calibration across ponds. Repeat these steps 50 times to get a posterior sample of 50 calendar ages for all pollen samples in all ponds and run the prediction model once for each sample of ages, as we have done to reflect uncertainty in the model parameters from the estimation runs. We plan to pursue this approach in our ongoing work.

Of course one could also build the dating model into the full model, but in line with our strategy of modeling in modules, we prefer to build a model for the dating uncertainty and then sample from the posterior of the dates and run multiple prediction runs, each with a different sample of dates. While the pollen data contain some information about the true dates, this is likely to be minimal given sediment mixing and tree lifespans.

2.4 Vegetation data

2.4.1 Colonial witness tree data

During settlement of central New England in the 17th and 18th centuries, colonial surveyors surveyed lots of size 0.5-65 ha for settlement, citing 'witness' trees as permanent markers of the lot corners within townships (approximately 6-mile square). Records of these witness trees have been recovered from town archives, and surveyor identifications have been mapped to modern taxonomic classification (Cogbill et al. 2002). These data are available aggregated to the township, with between 26 and 3149 trees per township for 183 townships with known boundaries in our study region, providing 87,114 trees in total (Fig. 1c).

2.4.2 Modern vegetation data

The U.S. Forest Service (USFS) Forest Inventory Analysis (FIA: www.fia.fs.fed.us) program samples vegetation using randomly-located plots on both public and private land, counting and identifying (to species) all trees in four 7.3 m radius subplots located 36.6 m apart. Our data consist of FIA tree counts of individuals greater than 10 cm diameter at breast height (1.3m) from 1990 for 1094 plots in the study area, with individuals aggregated into our ten taxa (Fig. 1d). Because of privacy concerns, USFS randomizes the plot locations to within 1.6 km of actual location. The plots contain between one and 115 trees per plot, with 29,938 trees in total. In the modeling described here, we use tree counts to simplify the error structure, allowing for an overdispersed multinomial error structure. However, basal (cross-sectional) area is likely to be more closely related to pollen production, with production increasing sharply with tree size, and should therefore be more closely related to pollen proportions in the sediment; in the discussion we further comment on this issue.

3 Model description

3.1 Notation

Let $p = 1, \ldots, P$ (for population) index the $P = 10$ tree taxa. The subscript i indexes the vegetation plots or townships. We work on a regular grid, with $s = 1, \ldots, S$ indexing the $S = 16^2$ spatial locations on the grid and $t = 1, \ldots, T$ indexing the T time points, discretized in 100 year intervals. To simplify the notation, we suppress the dependence on t when considering the modern and colonial periods. With this exception, where we omit subscripts, we indicate vectors, e.g., $\boldsymbol{v_i} = (v_{1,i}, \ldots, v_{P,i})$ is the count of trees of all taxa in the ith location.

3.2 Overview

Our modelling proceeds in two basic steps. First, in 'estimation runs', we use the modern and colonial data to estimate key parameters describing the pollen-vegetation relationships

and critical hyperparameters that constrain the model structure, borrowing strength across multiple ponds based on the spatial process structure. Second, in 'prediction' runs, we use only pollen data and the estimated key parameter values to make predictions in the past. The critical hyperparameters reflect the general structure of vegetation and parameterize spatial process variability, regression coefficient variability, and long-distance pollen dispersal. They serve to constrain the model to produce reasonable predictions with only a small number of ponds.

An alternative is to fit a coherent Bayesian model to all the pollen and vegetation data at all points in time. However, with a complicated model and multiple data sources, model misspecification and difficulty in model development and assessment are major concerns that would be exacerbated in a single integrated analysis. Our approach also allows us to carefully control what information is used to inform and constrain the inference at different points in time, for example, making inference about parameters related to the general structure of the vegetation based only on the vegetation data (Section 3.3.4). It also eases the computational burden. An additional advantage of the two-stage approach is the ability to more easily develop, diagnose, and improve the estimation model step by step before making use of the model to make predictions. See also Haslett et al. (2006) for a similar strategy.

We note that our multivariate non-normal outcome prevents conjugate updates of the latent process values and integration over these process values, greatly affecting MCMC mixing and limiting our ability to fit complicated structure in the model hierarchy, in contrast to much recent work with normal data that extends simple Bayesian spatial models to spatio-temporal, multivariate, nonseparable and other settings. This constraint, combined with sparse, noisy, and complicated data, necessitates careful attention to deciding upon the key aspects of reality to represent in the model structure.

3.3 Estimation model

3.3.1 Likelihood terms

Our likelihood terms are conditional on a latent multivariate spatial process, which provides the the composition vector for each grid cell, $\boldsymbol{r}(\boldsymbol{s}) = (r_1(s), \ldots, r_P(s))$, described in Section 3.3.2. Here we define the separate likelihoods for modern plot data, colonial witness tree data, and pollen data.

Vegetation For the vegetation, our basic strategy is to use a Dirichlet-multinomial ($\mathcal{DM}$) structure (also known as the compound multinomial distribution, a generalization of the beta-binomial) (Dey and Maiti 2002) to account for overdispersion in the vegetation data due to heterogeneity of vegetation within grid cells. First, consider the FIA plot data. We associate each plot, i, with the grid cell in which the plot falls, $s(i)$. The likelihood for the vector of tree counts, conditionally independent between plots, is $\boldsymbol{v_i} = \{v_{1,i}, \ldots, v_{P,i}\} \sim \mathcal{DM}(n_i, \alpha_{\text{FIA}}\boldsymbol{r}(\boldsymbol{s}(\boldsymbol{i})))$, where $n_i = \sum_p v_{p,i}$. The scalar Dirichlet precision parameter, α_{FIA}, is multiplied by each element of the composition vector for the grid cell in which the plot falls, $\boldsymbol{r}(\boldsymbol{s}(\boldsymbol{i}))$. For the witness trees, the structure is similar, except that the tree counts are aggregated into townships, which are generally larger than the grid cells and are misaligned

with respect to the grid. To account for this, we consider the count of trees in a township, i, to represent a weighted average of the trees in the grid cells that the township overlaps, $s \in O(i)$, where $O(i)$ is the set of overlapped grid cells. The weighting is based on the proportion of the township falling in each grid cell, $w_i(s)$. This gives us the likelihood for the ith township, $\boldsymbol{v_i} \sim \mathcal{DM}(n_i, \alpha_{\text{WT}}\bar{\boldsymbol{r}}(\boldsymbol{i})))$, where $\bar{\boldsymbol{r}}(\boldsymbol{i}) = (\bar{r}_1(i), \ldots, \bar{r}_P(i))$ and the proportion of the pth taxa in the ith township is $\bar{r}_p(i) = \sum_{s \in O(i)} w_i(s) r_p(s)$. In other words, the composition for the township is calculated as the integral over the gridded piecewise composition surface. Other approaches are possible, such as using the intersections of the grid cells and townships (Mugglin et al. 2000) in the discretization of the spatial domain, but seems unlikely to materially affect the results.

Pollen For the pollen, the likelihood must account for the fact that pollen production and dispersal vary by taxon, which causes the proxy pollen data to be biased for the local vegetation, even if one were to directly measure pollen falling to the ground. We again use the Dirichlet-multinomial form for the pollen count data for the modern and colonial eras, but we differentially scale the vegetation composition in the grid cell to account for the bias. The likelihood for the vector of pollen counts at location i, $\boldsymbol{c_i}$, is,

$$\boldsymbol{c_i} = \{c_{1,i}, \ldots, c_{P,i}\} \sim \mathcal{DM}(n_i, \boldsymbol{\phi} \bullet \boldsymbol{r}(\boldsymbol{s}(\boldsymbol{i}))), \tag{1}$$

where $\boldsymbol{\phi}$ is a vector of taxon-specific scaling factors that relate pollen to vegetation and $\boldsymbol{r}(\boldsymbol{s}(\boldsymbol{i}))$ is the vegetation composition of the grid cell in which the pond lies. Note that the multiplication is done element-wise (i.e., a Hadamard product). Because of chestnut blight, there are essentially no pollen-producing chestnut adults in the modern era, so we cannot estimate ϕ for chestnut in the modern era and assume this value is the same as for the 'other' category. Note that there is chestnut pollen in the modern sediment samples, because of sediment mixing at the sediment-water interface.

An added complication is that examination of the pollen data suggests a substantial fraction of pollen is derived from long-distance dispersal. Many ponds have taxa present despite little evidence in the vegetation data (for either the modern or colonial periods) or based on site visits by the authors that the taxa exist locally in sufficient quantity to explain the pollen abundance. The model assumes that $0 \leq \gamma \leq 1$ of the proportion of pollen produced in a cell remains in the cell and the remaining $1 - \gamma$ distributes in a distance-weighted fashion in a 15 by 15 grid of cells (some of which extend beyond our core grid) centered around each cell (see also Nielsen and Sugita 2005 for a similar decomposition of local and long-distance dispersal). The result is to replace $r_p(s(i))$ in (1) with

$$\gamma r_p(s(i)) + (1 - \gamma)\frac{1}{C} \sum_{s_k \neq s(i)} r_p(s_k) w(s(i), s_k), \tag{2}$$

where C is a normalization term calculated by summing $w(s(i), s_j)$ over cells s_j in the 15 by 15 grid surrounding the focal cell. The second term is a weighted average of the vegetation composition in the core grid cells other than $s(i)$, where weights,

$$w(s(i), s_k) = \exp\left(-\frac{d(s(i), s_k)^2}{\psi^2}\right),$$

are calculated based on the distance between the cell in which the pond resides and the other cells based on the grid cell centroids, $d(s(i), s_k)$, scaled by a dispersal distance parameter, ψ. The result is that the model attempts to distinguish the portion of the pollen data that is informative about the cell vegetation, ignoring pollen that reflects vegetation similar to the region as a whole, and essentially attempting a deconvolution.

The implied Dirichlet precision parameter for the pollen data depends on the scaling parameters and varies between ponds in different grid cells,

$$\alpha_{\text{pollen}}(i) = \sum_p \phi_p \left(\gamma r_p(s(i)) + (1-\gamma)\frac{1}{C} \sum_{s_k \neq s(i)} r_p(s_k) w(s(i), s_k) \right), \tag{3}$$

with somewhat lower values and therefore lower precision for ponds on the periphery of the domain because of the lack of modelled pollen input from cells outside the domain.

Ideally we would use an anisotropic, skewed dispersal kernel that reflects the effects of prevailing wind direction, but we were not able to find a reasonable skewed kernel parameterization. It would also be preferable to extend the domain to include vegetation at fairly large distances in all directions from the study ponds to limit boundary effects.

3.3.2 Spatially-correlated vegetation composition process

Using the spatial representation described below, which provides an approximate thin plate spline-based spatial process, $g_p(\cdot)$, for each taxon, we define the proportions of the ten taxa at a given location using the additive log-ratio transformation (Aitchison 1986, p. 113), where the proportion of taxon p at location s is

$$r_p(s) = \frac{\exp(g_p(s))}{\sum_{k=1}^{P} \exp(g_k(s))} \Rightarrow \sum_p r_p(s) = 1. \tag{4}$$

Note that the Aitchison (1986, p. 113) model has a one in the denominator in place of the contribution to the sum from the tenth, 'other', category, as well as replacing the numerator with one for $p = P$. For our MCMC implementation (Section 3.5), we specify $g_P(\cdot)$ in order to improve mixing. The result is that the processes are not fully identifiable, but the vegetation compositions are, because of the sum to one constraint (4). This approach allows us to use standard spatial models, yet create a multivariate framework for compositional data, and is very similar to the approach of Haslett et al. (2006).

Billheimer et al. (1997) take a similar approach to compositional data using a multivariate conditional autoregressive Markov random field model, while Tjelmeland and Lund (2003) take a Bayesian approach with a multivariate Gaussian process prior defined on the additive log ratio transformation. Pawlowsky and colleagues also focus on the additive log ratio transformation, investigating dependence structure amongst compositional components beyond that induced by the sum to one constraint and using kriging methods for modeling after transformation (Pawlowsky and Burger 1992; Pawlowsky-Glahn and Olea 2004). Note that unlike these approaches, for which composition proportions are the data, our data are in the form of counts, which requires the Dirichlet-multinomial structure (Section 3.3.1) in addition to the Gaussian structure on the transformed compositions.

Latent processes We take the $P = 10$ latent spatial processes to be independent spatial processes , $g_p(\cdot)$, defined at each grid cell location as $g_p(s)$, using a knot-based radial basis function approximation to a thin plate spline (Ruppert et al. 2003, Ch. 13). The value of the process at the 256 grid locations is

$$\boldsymbol{g_p} \;=\; \beta_{0,p}\mathbf{1} + \sum_k \boldsymbol{x_k}\beta_{k,p} + \boldsymbol{\Psi u_p}. \tag{5}$$

Here, $\boldsymbol{\Psi}$ is a reduced-rank basis matrix constructed using thin plate spline generalized covariance matrices on an equally-spaced 9 by 9 grid of knots. The 81 basis coefficients are taken to have prior distribution, $\boldsymbol{u_p} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$, with the single variance component controlling the amount of smoothing. The covariates are described below.

We recognize that vegetation is likely to be nonstationary, while the construction is stationary, but believe that by including key covariates in the mean structure, in particular elevation, we have accounted for a major source of nonstationarity. Our approach avoids the computational difficulties of nonstationary processes and recognizes the limitations on resolution caused by the sparseness of the pollen data. The inclusion of covariates also helps to justify our use of a single σ^2 common to all taxa (an approach that Haslett et al. (2006) also find to be sufficient) reflecting that taxon abundances tend to change in tandem spatially.

Also note that our model assumes prior dependence between taxa only to the extent induced by a sum to one constraint, reflecting our desire to avoid a dependence structure that because of data sparsity would need to be implausibly constant over space. For example, the ranges of two taxa may overlap in part of the domain, but only one taxon may be present in another part of the domain. We do not want to infer the presence of the missing taxon (it may in fact be beyond its range boundary) because of correlation inferred elsewhere. Critically, for every pond we have a large multinomial sample and direct information on each taxon from its count. Borrowing strength across taxa through a dependence structure introduces potential for bias from misspecification of the dependence structure, while the balanced sampling of data provides limited opportunity for variance reduction. Of course a posteriori, the taxa will be correlated insofar as they share similar covariate values and spatial process values. This approach reflects current ecological thinking that the distributions of tree species are driven by abiotic factors more than tight interspecific linkages (Davis 1981). There is real ecological dependence between taxa; individual taxa modify forest light levels and soil characteristics in a way that influences the abundances of other taxa (Pacala et al. 1996). However, these dependencies are likely to change over time and are in fact one of the outcomes of interest; we do not want to assume a correlation structure amongst the taxa that is constant over time.

Landscape covariates Vegetation abundance is strongly related to covariates such as elevation, soil type and climate. Covariates are represented in the spatial process representation (5), where $\boldsymbol{x_k}$ is a vector of values of the kth covariate at each grid cell, and $\beta_{k,p}$ is the coefficient for the pth taxon. To predict in the past, we are limited to covariates whose values are known at every time point, generally those that have not changed much over time. In particular, we use elevation (averaged over the grid cell) and latitude (after projection) in the current model, as these are readily available and are the covariates most likely

to influence vegetation at the spatial resolution of our grid. Note that latitude (at the cell centroid) is merely a linear spatial term. We include only a linear term in latitude and not in longitude because vegetation is likely to vary most substantially with climate differences that vary most strongly with latitude, and for the prediction runs, with 23 or fewer ponds, we wanted to estimate as few parameters as possible, leaving any variability by longitude to be accounted for in the radial basis portion of the spatial process. Both covariates are centered about their means, with elevation scaled to units of 1 km and latitude to 100 km. In the estimation runs, these covariates allow us to better match grid-level vegetation estimates with pollen abundance at ponds for which there is little nearby vegetation data in the estimation runs. In the prediction runs, they allow us to better predict vegetation at grid cells not near ponds. Other potential covariates include modern climate information and soil type, but for the moment we use only the two covariates above; in part because of the limited sample size.

3.3.3 Hyperparameter representation, prior distributions, and shrinkage

The goal for our prior distributions for the various parameters is to allow the data to play the primary role in estimating the parameters, while borrowing strength as necessary in contexts in which the data provide limited information, which is particularly relevant for the prediction runs, for which the small number of ponds provides limited information about spatial structure and covariate effects.

For the covariate effects, we use exchangeable prior structures to allow us to estimate hyperparameters in the estimation runs that can be used to constrain the relevant parameters in the prediction runs. We take $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$, the coefficients for elevation and latitude, respectively, $\boldsymbol{\beta_k} \sim \mathcal{N}(\mathbf{0}, s^2_{\beta_k}\boldsymbol{I}), k = 1, 2$. In the estimation runs, the coefficient for each taxon could be estimated individually with independent prior distributions with little difficulty based on the dense vegetation data, but the variance components allow us to stabilize the estimates of the coefficients in the prediction runs, while still allowing the coefficients to vary in time. Note that the coefficients are taken to have mean zero because (4) causes the mean to not be identifiable; only relative differences can be estimated.

The pollen scaling parameters, $\boldsymbol{\phi}$, are taken to be independent a priori with non-informative but proper priors, as are $\boldsymbol{\beta_0}$. The long-distance contribution parameter, γ, is taken to have a uniform distribution on $(0, 1)$. For the remaining parameters, $\{\sigma^2, \alpha_{\text{FIA}}, \alpha_{\text{WT}}, \psi, s^2_{\beta_1}, s^2_{\beta_2}\}$, we use independent, non-informative, but proper priors. In particular, for variance components, we have used uniform priors on the standard deviation scale to avoid the use of diffuse inverse gamma priors (Gelman 2006), which have sharp spikes in density at small values, and decay extremely rapidly to zero density at values smaller than the location of the spike. For all the parameters, we impose lower and upper limits on the parameter values to prevent the MCMC sampler from wandering in areas of the parameter space in which the data provide little information and ensure propriety. In all cases of non-informative priors, the posterior distributions were concentrated away from the limits, suggesting that the diffuseness of the prior is not of particular concern (see the considerations of Berger et al. 2001).

3.3.4 Model misspecification and model incoherence

With regard to model misspecification, we know that sediment pollen records are an error-prone and biased proxy for vegetation, while the modern plot data, and likely the colonial vegetation data to a lesser extent, are relatively error-free. Thus in doing the estimation runs, we would like to estimate the key parameters used to constrain predictions in such a way that our vegetation surface estimates are informed primarily by the vegetation data. In our joint estimation model for vegetation and pollen data, in cells with limited vegetation data, the vegetation estimates in a cell can overfit to the pollen data. To avoid this, in the estimation runs, our MCMC samples the parameters used to construct the latent vegetation process, $\boldsymbol{r}(\boldsymbol{s})$, are done conditional only on the vegetation data, 'cutting feedback' in a manner recently introduced into the BUGS software (Spiegelhalter et al. 2003) and discussed in detail in Rougier (2008). Yucel and Zaslavsky (2005) have also considered this issue in models with multiple data sources in which one data source directly informs a parameter, but a second, larger, set of data can also influence the inference more strongly than desired because of model misspecification. In our setting the pollen dataset acts as the 'larger' dataset within individual grid cells with ponds because of the large number of pollen grains compared to the tree counts. A sensitivity analysis suggested that the coherent model without cutting feedback overfits to some degree but not a substantial amount, with increased estimates of the precision in the pollen data and of the proportion of grid-cell pollen, γ.

In the same vein, but in the temporal domain, we wish to estimate the key prediction parameters from the time periods with vegetation data and use those parameter values in the prediction runs. We want to avoid the danger that extensive pollen data from older time periods would swamp the inference about certain parameters in some way, even though there is inherently no information in those time periods about the parameters. For example, inference about the spatial process parameter, σ^2, might be influenced by the prediction points, increasing the smoothness of the process because of the sparse pond data, even though the vegetation data provide much more information about level of spatial variability in vegetation. Our strategy of splitting the analysis into estimation and prediction runs allows us to make critical choices and estimates in the estimation runs and apply these choices to the prediction problem, albeit at the cost of a strictly coherent Bayesian approach.

3.4 Prediction model

After fitting the model in the estimation runs for the colonial and modern eras, we use fixed parameter values from those runs in the prediction runs, which have the same model form as the estimation runs, but with temporal autocorrelation introduced as described below. To account for uncertainty in the parameters in the estimation runs, we compute separate predictions conditional on samples from the posterior of the parameters from a given estimation run. In the prediction runs, only $\boldsymbol{\beta_{0,p}(t)}$, $\boldsymbol{\beta_{1,p}(t)}$, $\boldsymbol{\beta_{2,p}(t)}$, and $\boldsymbol{u_p(t)}$, which are the parts of the model that directly determine the vegetation composition at each time, and autocorrelation parameters for these time series, are estimated. This approach ensures that the vegetation predictions are primarily informed by the pollen proportions at the time of interest, but that structural information that is well-informed only with rich

vegetation data is based on the estimation runs.

The temporal structure gives us the ability to smooth over time to better estimate $r_p(s,t)$ and assess how the relationship between taxon abundances and covariates have changed over time (e.g., Williams et al. 2001) as inferred from the pollen data. For each of the temporally-varying terms, we include an overall mean that we integrate over for better MCMC mixing, giving us two temporal variance components. For example,

$$\boldsymbol{\beta_{0,p}} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2(\delta_0 \boldsymbol{J} + (1-\delta_0)\boldsymbol{R}(\rho_0)))$$

where $\delta_0 \in (0,1)$ is the proportion of variance for the long-term mean, $\boldsymbol{J}$ is a matrix of ones, σ_0^2 is the overall variance, and $\boldsymbol{R}(\rho_0)$ is the correlation matrix, a function of decay parameter ρ_0 and the relevant time lags. We use a Matérn correlation function with $\nu = 2$ but also consider the exponential (i.e., AR(1)) correlation function. The priors for $\boldsymbol{\beta_{1,p}}$ and $\boldsymbol{\beta_{2,p}}$ are analogous but with $s^2_{\beta_1}$ and $s^2_{\beta_2}$ in place of σ_0^2 and ρ_1 and ρ_2 in place of ρ_0. We choose σ_0^2 to be large, imposing no constraints on the overall mean, $\boldsymbol{\beta_0}$, while using the variance components for $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ from the estimation runs to stabilize their estimation. To provide for residual spatio-temporal structure, we specify an analogous temporal correlation structure for the basis coefficients

$$\boldsymbol{u_{k,p}} \sim \mathcal{N}(\mathbf{0}, \sigma^2(\delta \boldsymbol{J} + (1-\delta)\boldsymbol{R}(\rho))),$$

again independent between coefficients for different knots, $k = 1, \ldots, 81$, and where σ^2 is taken from the estimation runs. This constraint ensures that the amount of spatial heterogeneity is based on information from the rich vegetation data in the estimation runs. Nonseparability may come into play, particularly at times of range expansion and contraction, but would difficult to estimate based on the small number of sites, and would add even more complexity to the modeling. For the proportion of variance and the decay parameters we use uniform priors, where for the latter we impose upper and lower bounds based on the discrete time lag and length of the time period.

Because vegetation changed markedly upon European settlement (Fuller et al. 1998), introducing a likely nonstationarity in time, we run the prediction model separately for the pre-settlement (2500 to 300 years before present) and post-settlement periods (500 to 0 years before present). Note the inclusion of a buffer on either side of the settlement period for both runs to avoid boundary effects.

In our prediction runs, we make use of the posterior distributions for all parameters, except those mentioned above, from either the modern or colonial estimation run. To incorporate uncertainty in these parameters, we fit the prediction model using 50 draws from the joint posterior distribution of $\{\boldsymbol{\phi}, \gamma, \psi, \sigma^2, s^2_{\beta_1}, s^2_{\beta_2}\}$ from the chosen estimation run, running a separate MCMC for each draw, and combining the iterations from the 50 chains to estimate posterior quantities. In this way we incorporate parameter uncertainty into our predictions, but we do not update these distributions as the pollen data alone do not contain sufficient information to inform the parameters.

3.5 Implementation

We note that the lack of conjugacy and inability to integrate over the latent processes in our hierarchical multivariate space-time model with a non-normal likelihood seriously af-

fect mixing and require long run times, even with the simplifications in our model structure. We recognize that replication of complicated MCMC schemes is difficult and attempt here to describe in detail some of the modifications, relative to a straightforward blocked Metropolis-Hastings implementation, that we used to speed mixing.

3.5.1 MCMC sampling schemes

We first consider the sampling schemes used in the estimation runs. $\boldsymbol{\beta_0}$ is sampled as a block via simple Metropolis. The elements of $\boldsymbol{\beta_0}$ are not identifiable; all could be shifted by the same amount with no effect on the likelihood (4). Therefore in the MCMC for the estimation runs we center the values at zero at each iteration (Besag et al. 1995), which while not strictly appropriate because of the influence of the prior on $\boldsymbol{\beta_0}$, has no effect in our setting in which the prior is essentially flat. Similarly, in sampling $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$, we employ blocked Metropolis proposals. We also use a Metropolis proposal that shifts all the coefficients by the same amount. Since $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ are not identifiable in that a constant shift of all the elements in each has no effect on the likelihood, the proposal that shifts all the elements at once is constrained only by the prior, $\mathcal{N}(\mathbf{0}, s^2_{\beta_k}\boldsymbol{I})$, and allows the values to mix in this region that is unconstrained by the data. The spatial process values are also inherently not identifiable because we use all 10 taxa in (4) instead of holding out a taxon as a reference category, as is usually the case in the additive log-ratio transformation (Aitchison 1986). An initial assessment suggested that the this parameterization helps somewhat in mixing, albeit not substantially, compared to a model in which the extra non-identifiable component is excluded. Note that the process coefficients, $\boldsymbol{u_p}$, are constrained by their prior distributions, which helps with mixing. The coefficients are proposed via blocked Metropolis with separate blocks for each taxon. Note that the complicated relationship between the data and the latent processes, involving the Dirichelt-multinomial distribution and the additive log-ratio transformation of multiple processes makes it very difficult to use more targeted proposals than simple random walks. In sampling $\boldsymbol{\phi}$, we use two sampling approaches. First, we use a block Metropolis proposal. Second, to allow the overall magnitude to mix quickly, we propose to shift all the values by a constant amount (on the log scale) in a Metropolis proposal. This helps mixing with respect to the overall Dirichlet heterogeneity (3).

For some of the key hyperparameters that control population distributions ($s^2_{\beta_1}, s^2_{\beta_2}$) and process structure (σ^2), dependence between the hyperparameter and associated random effects can greatly slow mixing (Knorr-Held and Rue 2002; Rue and Held 2005; Paciorek 2007). Following Paciorek (2007), we use joint proposals that first propose the hyperparameter and then deterministically shift the associated random effects such that the random effects prior density remains the same. For example, after proposing $s^{2*}_{\beta_k}$, we jointly propose, $\boldsymbol{\beta^*_k} = \boldsymbol{\beta_k}\sqrt{s^{2*}_{\beta_k}/s^2_{\beta_k}}$. These joint proposals propagate changes in a hyperparameter to the level of the data, involving the log-likelihood in the acceptance decision. Acceptance is determined in a single decision based on the prior distribution of the hyperparameter and the log-likelihood because the random effects prior stays constant between the current values and proposed values, after accounting for the deterministic shift based on the Jacobian of the transformation. The legitimacy of this proposal follows from an argument used in justifying reversible-jump MCMC (Green 1995). For σ^2 the joint proposal is similar except

that the coefficients for all the taxa are scaled based on $\sqrt{\sigma^{2*}/\sigma^2}$ as above within the same joint proposal. For $s^2_{\beta_1}$ and $s^2_{\beta_2}$ we also propose to move them via simple Metropolis.

The remaining parameters are sampled by simple Metropolis sampling.

In sensitivity runs, we allow ψ and γ to vary between taxa, using an exchangeable prior distribution with mean m_ψ and variance s^2_ψ; for $\boldsymbol{\gamma}$ the prior is similar, but ensures that $\gamma_p \in (0, 1)$ by truncation. As above, to account for dependence between hyperparameters and random effects, we use joint proposals, which we detail here for $\boldsymbol{\psi}$; proposals for $\boldsymbol{\gamma}$ are analogous. For m_ψ, we propose m^*_ψ, and in the same joint proposal, propose $\boldsymbol{\psi}^* = \boldsymbol{\psi} - m_\psi \mathbf{1} + m^*_\psi \mathbf{1}$; this is similar to the joint proposals described for $\boldsymbol{\beta_1}$, and $\boldsymbol{\beta_2}$ above, except that in this case the mean of the vector parameter is a parameter in the model. The joint proposal involving s^2_ψ is analogous to that for $s^2_{\beta_k}$ above.

In the prediction runs, the time series $\boldsymbol{\beta_0}$ (and similarly for $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$) is sampled in a block with a correlated Metropolis proposal using the current temporal correlation induced by ρ_0 and δ_0. The coefficients for all taxa are sampled within the same block. Sampling of $\boldsymbol{u_p}$ is done in similar fashion but the coefficients for the different taxa are sampled in separate blocks. For $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ we also include a proposal in which the values for all 10 taxa are moved by the same amount, but with a correlated proposal using the current correlation. As before this helps with mixing with respect to the prior distribution but does not change the likelihood. Finally for $\{\rho, \rho_0, \rho_1, \rho_2\}$ and $\{\delta, \delta_0, \delta_1, \delta_2\}$, we use joint proposals of hyperparameters and process values as described above. For example, we jointly sample $\{\rho, \boldsymbol{u_p}, p = 1, \ldots P\}$as well as $\{\delta, \boldsymbol{u_p}, p = 1, \ldots, P\}$, with analogous sampling for the regression coefficient hyperparameters and their associated time series.

3.5.2 MCMC for the estimation runs

We first estimate the key prediction parameters in the estimation runs. The model is run from three different sets of initial values for 400,000 iterations each after an initial burn-in of 10,000, with every 40th iteration saved to economize on storage space and posterior computations. This gives us a sample of 30,000 values from the posterior. A sample of trace plots, with effective sample size estimates (Neal 1993, p. 105), is shown in Fig. 3 for key hyperparameters as well as the model log posterior (up to the normalizing constant) and log-likelihood terms. MCMC mixing is sufficiently fast to allow us to claim reliable inference, but some parameters do show slow mixing, and computation times are on the order of 36 hours for a full run of the 410,000 iterations in R with the Goto BLAS on a Linux computer with a 2.3 GHz processor. Note that key log-likelihood calculations are done in compiled C code called from within R.

3.5.3 MCMC for the prediction runs

For the prediction runs, using a single joint sample of the fixed parameters from the posterior distribution from the chosen estimation run (modern or colonial), we run the model for 60,000 iterations after a burn-in of 30,000. We average the posterior estimates over 50 such joint samples, giving a final sample of 37,500 values, after saving every 80th iteration. For prediction at a single time point, a sample of trace plots for the colonial era cross-validatory prediction model is shown in Fig. 4, with the sharp jumps occurring because of the changes

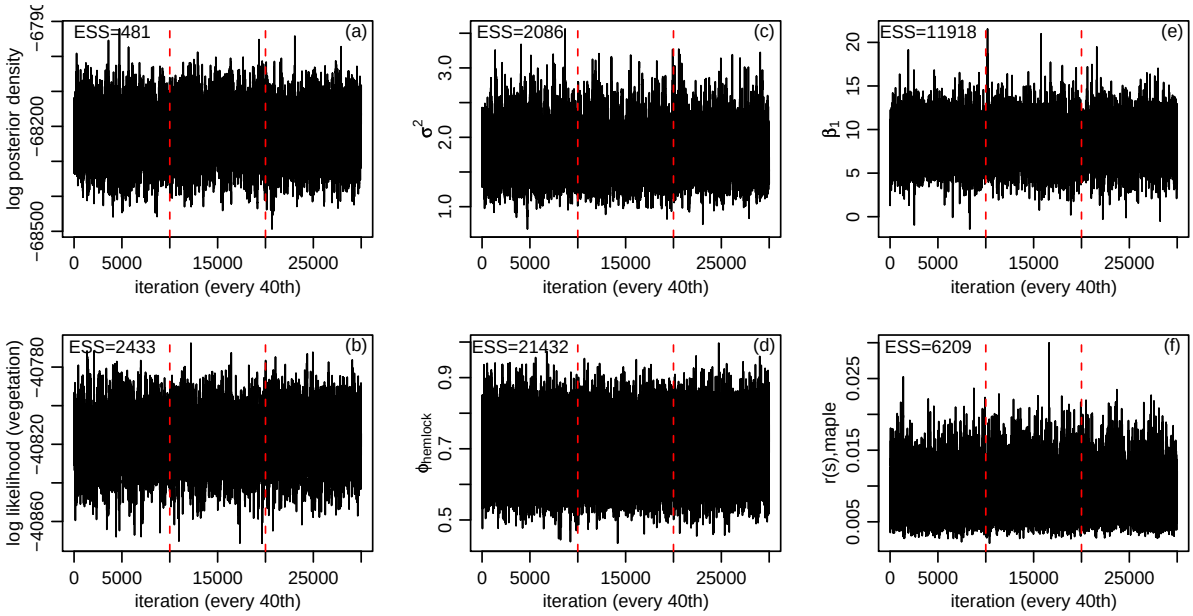

Figure 3. Trace plots of three combined chains (delineated by the red dashed lines) from the modern era estimation model for (a) log posterior density (up to the additive normalizing constant), (b) log likelihood for the FIA data, (c) σ^2, (d) ϕ for hemlock, (e) β_1 for spruce, and (f) proportion of maple in an arbitrary grid cell.

in the fixed parameter values. Mixing seems reasonable, with a large effective sample size after aggregating over all 50 runs. In contrast mixing for prediction runs over long time periods is much worse because of the higher-dimensional parameter space and temporal correlation (Fig. 5). Prediction runs take on the order of four days for 90,000 iterations for the 23 time points.

4 Model results and assessment

Results from the model come in several forms. In Section 4.1 we use the estimation runs to learn about the relationship between pollen and vegetation in the modern and colonial periods. We consider the ecological implications of parameter estimates and contrast results from the modern and colonial estimation runs to understand potential differences in vegetation structure. In Section 4.2, we assess the use of the model for prediction in a cross-validatory fashion. First, we focus on prediction of colonial era vegetation using parameter estimates from the modern estimation run and colonial pollen data. We compare the predictions to the vegetation as informed by the witness tree surveys based on point estimates, uncertainty estimates, and spatial pattern assessment. Second, we predict for the modern era based on the surface sediment pollen and colonial parameter estimates. Having argued that our model performs reasonably, in Section 4.3, we apply the prediction model to pollen data over the past 2500 years.

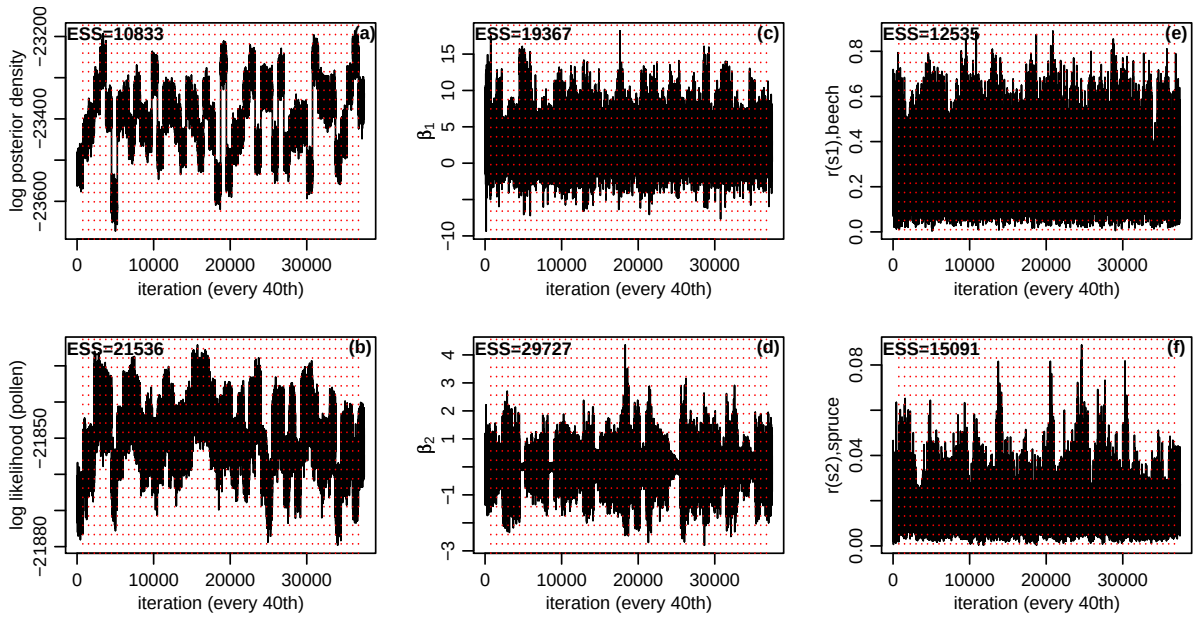

Figure 4. Trace plots of 50 combined chains (delineated by the red dashed lines, indicating changes in the fixed parameters between chains) from the prediction model for the colonial era for (a) log posterior density (up to the additive normalizing constant), (b) log likelihood for the pollen data, (c) β_1 for birch, (d) β_2 for hemlock, (e) proportion of beech for an arbitrary grid cell, and (f) proportion of spruce for a different arbitrary grid cell.

4.1 Estimation model results

4.1.1 Pollen as proxy for vegetation

Differential pollen production and dispersal The estimation runs allow us to characterize the relationship between pollen in sediments and local vegetation, thereby informing us about the ability of pollen to serve as proxy data for vegetation. Our model attempts to find the best fit between pollen and vegetation across a regional network of sites. As a residual diagnostic, we compare the pollen composition in each pond to the spatially-smoothed estimated vegetation composition of the encompassing grid cell from the model. While differences between pollen and vegetation composition may arise because the grid-scale vegetation is poorly estimated, most ponds fall in areas with nearby FIA plots or township data (see Figure 2.2), so we expect that most differences are due to long-distance pollen transport and local (within grid cell) vegetation heterogeneity.

For the colonial and modern eras, respectively, Figs. 6-7 plot relative pollen abundance in ponds versus model-smoothed grid cell relative vegetation abundance for each taxon (red crosses). Most taxa show increasing relationships. The lack of 1:1 relationship shows the importance of including ϕ to adjust for differential pollen production and dispersal. After scaling the smoothed vegetation by the estimated values of ϕ, we see the values (black squares) falling around a 1:1 line, albeit with some taxa, such as oak and hickory, showing more consistent relationships than others, such as spruce. The substantial remaining variability makes it difficult to precisely estimate ϕ.

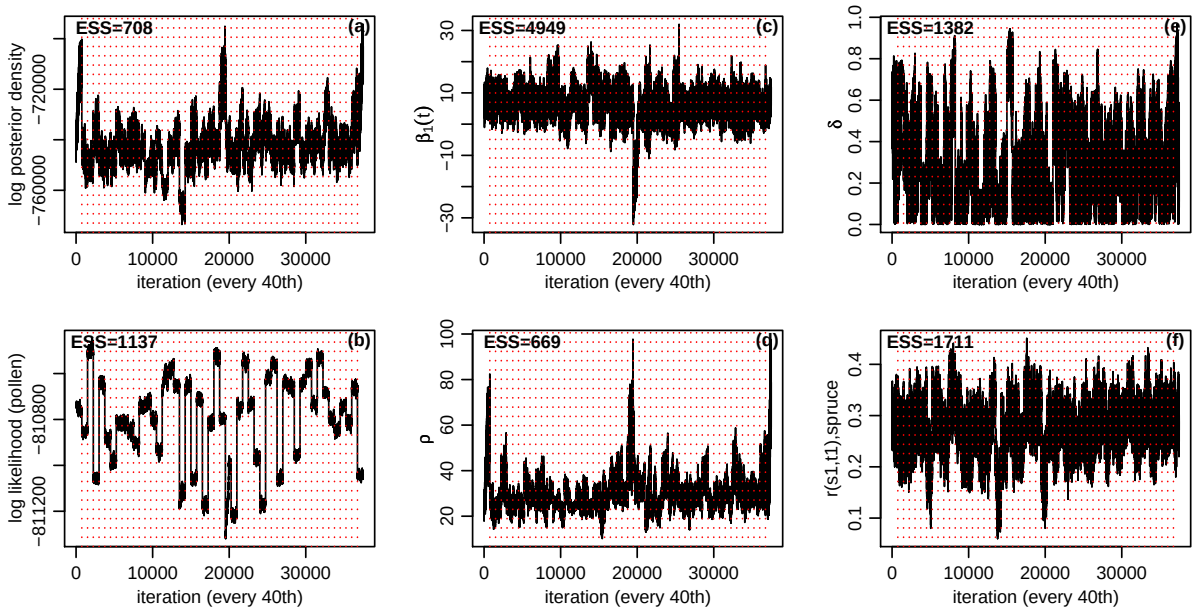

Figure 5. Trace plots of 50 combined chains (delineated by the red dashed lines, indicating changes in the fixed parameters between chains) from the prediction model for the period 2500 to 300 years before present for (a) log posterior density (up to the additive normalizing constant), (b) log likelihood for the pollen data, (c) $\beta_1(t = 8)$ for birch, (d) ρ, (e) δ, and (f) proportion of spruce for an arbitrary grid cell and time.

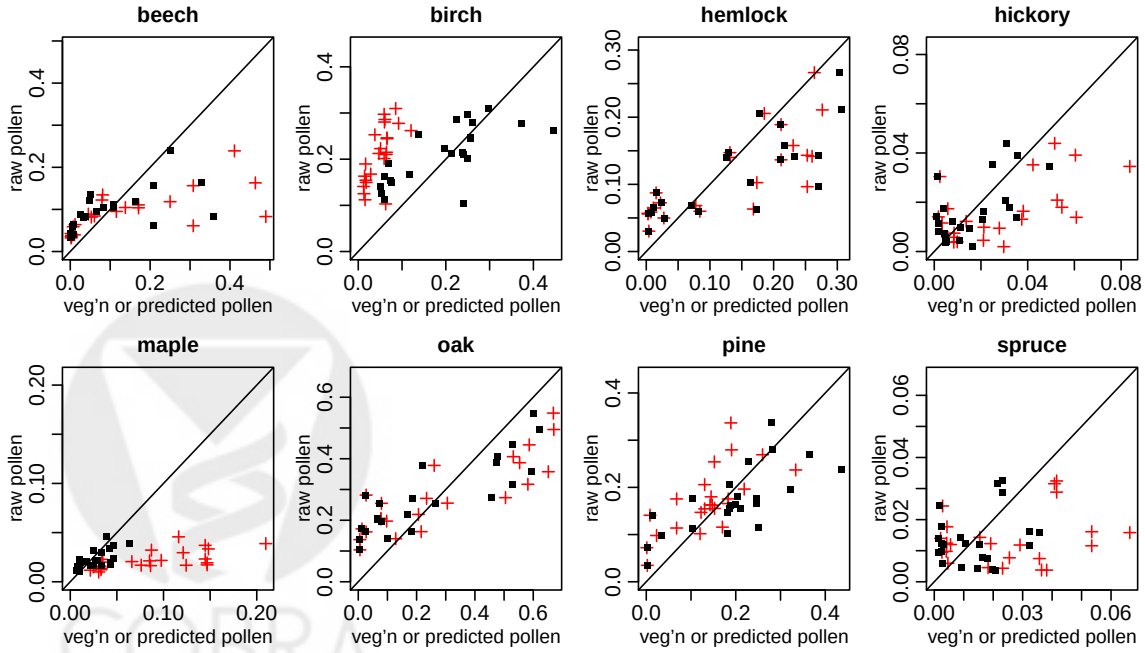

Figure 6. For the colonial era, scatterplots by taxa of pollen proportions in each pond against both the model-smoothed vegetation proportions in the grid cell of the pond (red crosses) or model-predicted pollen proportions based on scaling the smoothed vegetation in the cell by ϕ (black squares).

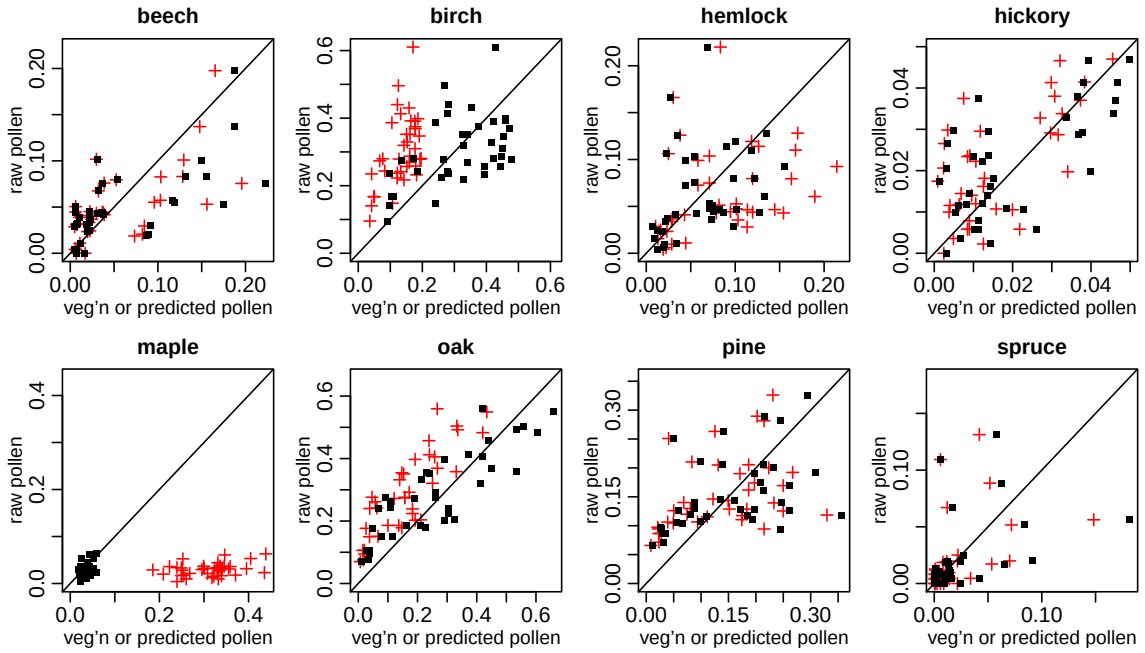

Figure 7. For the modern era, scatterplots by taxa of pollen proportions in each pond against both the model-smoothed vegetation proportions in the grid cell of the pond (red crosses) or model-predicted pollen proportions based on scaling the smoothed vegetation in the cell by ϕ (black squares).

Based on plots by pond rather than by taxon (not shown), most ponds show an increasing relationship between relative abundance of each taxon in the pollen and in pollen as predicted from vegetation, scaling by ϕ, although some ponds show sharp differences, particularly for some of the more abundant taxa. Fortunately, in almost all cases, taxa with low abundance in the vegetation can be distinguished from taxa with high abundance in the vegetation based on the pollen. Further exploration has not indicated any relationships with covariates or spatial patterns that might explain which ponds have more noisy relationships between the model-scaled pollen proportions and smoothed vegetation in the cell of the pond. Nor are the ponds with the noisy relationships consistent between the modern and colonial eras. This makes more sophisticated error modeling difficult.

The estimation runs also allow us to investigate differential taxon-specific pollen production and dispersal. For both the modern and colonial parameter estimates, large uncertainties prevent us from readily distinguishing among most taxa based on ϕ, but the two taxa whose estimates are clearly different than the others are maple, with low production/dispersal, and birch, with high production/dispersal (Fig. 8). These results agree with previous finer-scale analyses of the relationship between trees and pollen assemblages in the eastern U.S. (e.g., Jackson 1990), in which birch, oak, and pine are estimated to have well-dispersed pollen and maple poorly-dispersed pollen. Note that the estimates have been scaled so that the mean across taxa is one (this scaling is done by MCMC iteration), because it is the relative magnitudes of the parameters that are relevant in the likelihood (1). The large uncertainties in ϕ will make it difficult to compare abundances across taxa

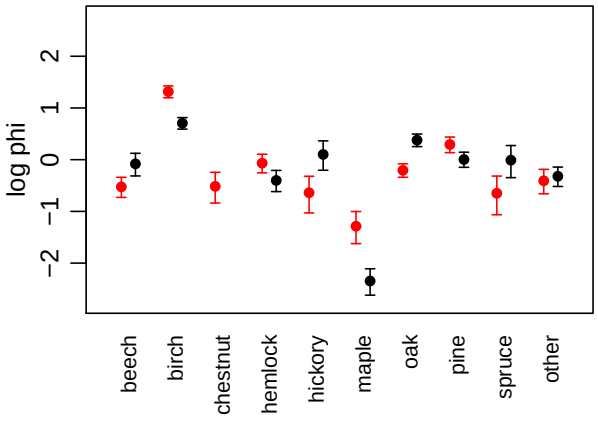

Figure 8. 95% credible intervals for $\log \phi$ *for the colonial (red) and modern (black) estimation runs. Note that chestnut is not shown for the modern period because of the absence of mature chestnut.*

in the predictions, although we should be able to distinguish common from rare taxa and to compare abundances spatially within taxa (as the latter inference does not rely on ϕ). For ecological reasons, such as different landscape openness and different assemblages of taxa, particularly over long time periods, the values of ϕ may vary over time. While we can only estimate ϕ for two time periods close in time, the colonial and modern periods represent stark changes in conditions; also the estimated values of ϕ represent averages across different conditions and assemblages in the different ponds, so our inference may be reasonably robust. Comparing between the two periods, the general patterns are consistent. However, note that as we discuss in the cross-validation assessment (Section 4.2), changes in ϕ over time can substantially affect the inferred overall abundance of taxa.

Long-distance dispersal To create diagnostic plots similar to Figs. 6-7, but that account for long-distance dispersal, we create the weighted average of vegetation from the focal grid cell and that from other cells and then scale by ϕ to represent the estimated pollen contribution to the pond; plotting this against the actual pollen composition gives us a residual diagnostic (black squares in Figs. 9-10). The relationship between pollen and predicted pollen based on the mixture approach appears to be closer than when comparing with pollen predicted solely based on grid cell contributions, suggesting that the model with long-distance transport fits better than a model that attributes all pollen to grid cell vegetation. Using DIC for model comparison also indicates that the simpler model fits substantially worse than allowing γ to be estimated in the estimation runs, with a ΔDIC of 307 (427) for the modern (colonial era). Not surprisingly, the estimated precision for the pollen data in the likelihood is smaller when $\gamma \equiv 1$, with values of $\bar{\alpha}_{\text{pollen}}$ of 35 compared to 60 for the modern run and 27 compared to 97 for the colonial run, as the additional heterogeneity is accounted for in the Dirichlet heterogeneity parameter. Further assessment using cross-validation supported the use of the mixture model, with the model without long-distance dispersal producing predicted vegetation surfaces with much less distinct

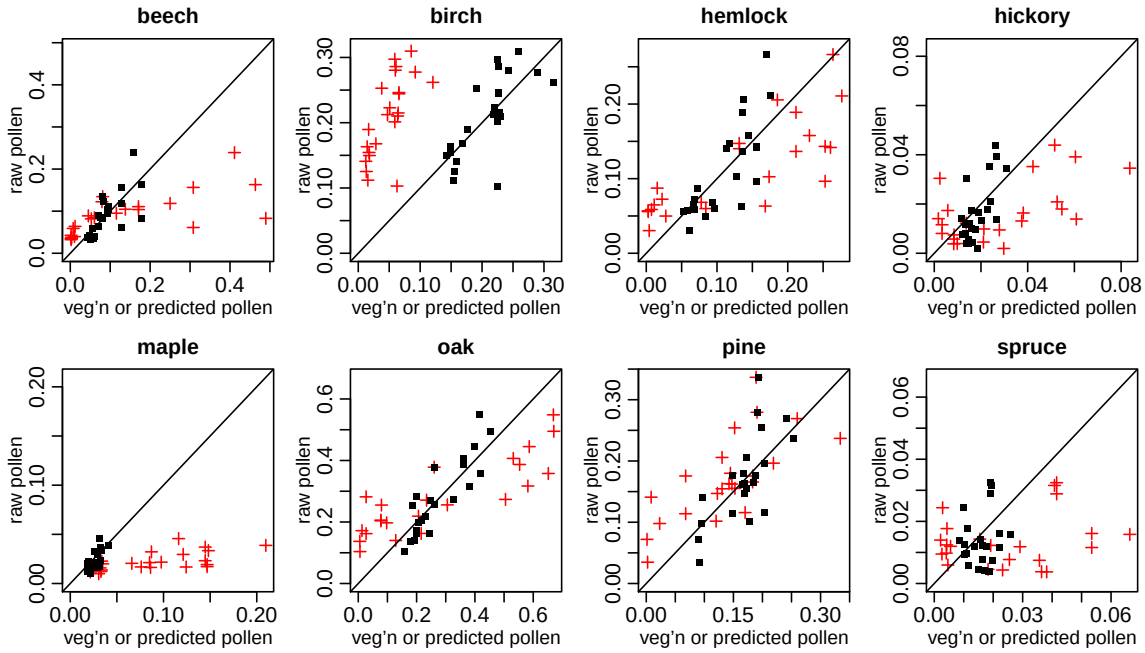

Figure 9. For the colonial era, scatterplots by taxon of pollen proportions in each pond versus model smoothed vegetation in the grid cell of the pond (red crosses) and model-predicted pollen from the mixture of local and long-distance dispersal, based on adjusting by ϕ, ψ, and γ (black squares) .

spatial patterns (not shown) and less posterior confidence about feature significance, as we would expect with the smaller estimated Dirichlet heterogeneity parameter for the pollen data. In future ecological analyses we will consider different approaches for pollen source contributions in more detail, as this is an issue of critical paleoecological importance and others have attempted to infer relative contributions by various methods (e.g., Jackson and Lyford 1999; Nielsen and Sugita 2005).

For the pollen data, γ represents the proportion of pollen data consistent with vegetation estimated in the encompassing grid cell, with $1-\gamma$ the proportion based on weighting the composition in the other grid cells in the domain. In both the colonial and modern eras, γ is about one-half: 0.48 for the modern era (with a 95% credible interval of 0.30, 0.61) and 0.50 (0.41, 0.59) for the colonial era, indicating that much of the pollen in the ponds is not consistent with the grid-cell-estimated vegetation. The pollen could be associated with long-distance transport, reflecting the vegetation in other grid cells, or with local sub-grid-scale vegetation that happens to be more similar to the region-wide vegetation than the model-estimated vegetation in the grid cell of the pond. While local variability and lack of identifiability in the model surely contribute to some extent, site visits by the authors suggest that many of the ponds visited had few nearby trees of the type indicated by the anomalous pollen, suggesting that much of the pollen may be due to long-distance transport. Our results are consistent with previous paleoecological work (Jackson and Lyford 1999; Davis 2000; Nielsen and Sugita 2005), which suggests that mixing of pollen sources makes distinguishing local from regional sources. For taxa that are at high abundance in

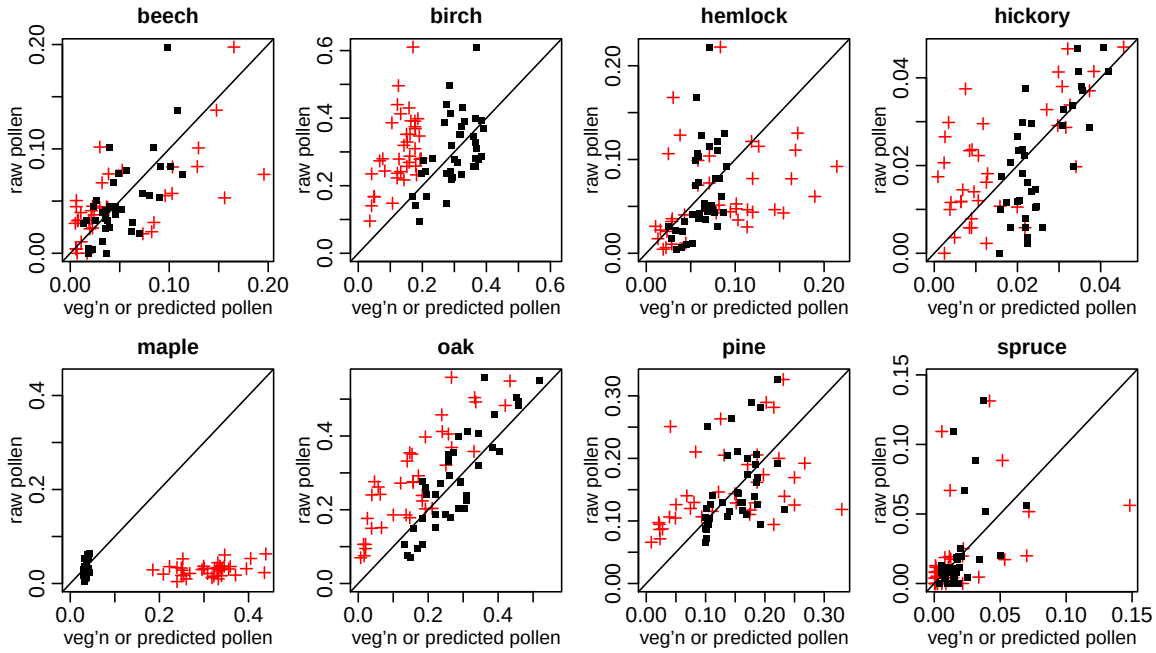

Figure 10. For the modern era, scatterplots by taxon of pollen proportions in each pond versus model smoothed vegetation in the grid cell of the pond (red crosses) and model-predicted pollen from the mixture of local and long-distance dispersal, based on adjusting by ϕ, ψ and γ (black squares).

most locations, such as maple in the modern era, it is particularly difficult to distinguish pollen from the grid cell compared to long-distance transport. Additional vegetation data from field surveys near ponds could help estimate local vegetation, thereby distinguishing long-distance from local pollen and improving our estimation of ϕ, ψ, and γ. A strength of the model is that it synthesizes already-existing data, but it could readily incorporate local vegetation data.

Ecologists expect the contribution of local pollen dispersal, γ, and the distance of dispersal, ψ, may differ by taxa (Jackson 1990), but a model with $\boldsymbol{\gamma}$ and $\boldsymbol{\psi}$ varying by taxa, both parameterized by exchangeable priors, showed little ability to distinguish differences between taxa (Fig. 11), albeit with a small improvement in DIC (6.7 for the modern era and 9.0 for the colonial). These parameters are difficult to estimate as the model involves a deconvolution of the deposited pollen, so all taxa show high levels of posterior uncertainty.

4.1.2 Error structure

In the model, we specify a Dirichlet-multinomial distribution for the vegetation and pollen data, with the mean for the distribution of the pollen data based on a mixture of grid cell and long-distance components. For the modern era, the posterior mean estimate of the Dirichlet precision parameter for the FIA data is 3.41 (3.24, 3.59) while for the colonial era for the witness tree data it is 39.1 (34.6,44.2). As expected much more precision is indicated in the colonial era, because the witness tree data are aggregated to the township level, smoothing

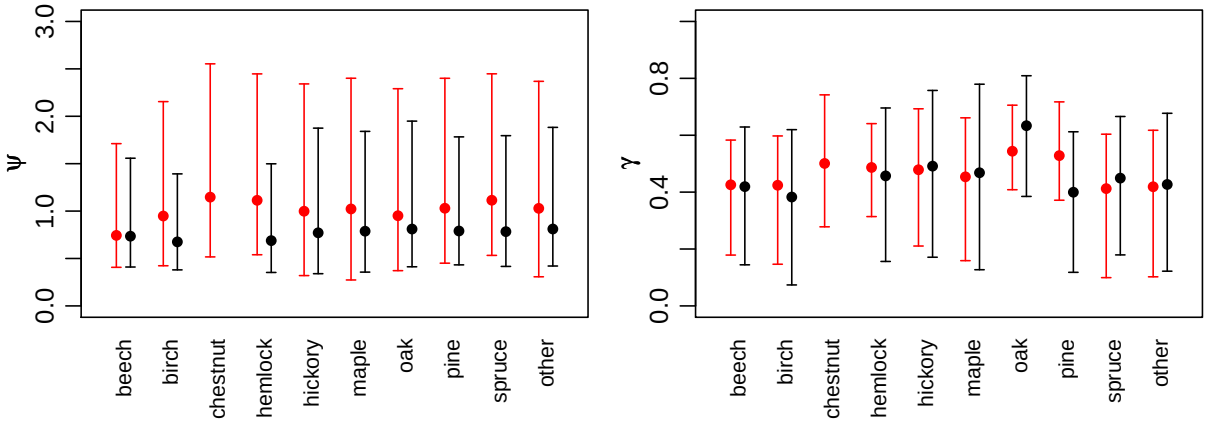

Figure 11. (a) 95% credible intervals for ψ for the colonial (red) and modern (black) estimation runs. (b) 95% credible intervals for γ for the colonial (red) and modern (black) estimation runs. Note that chestnut is not shown for the modern period because of the absence of mature chestnut.

over local heterogeneity, compared to scattered FIA plots with relatively few trees. Of course, we are not sure how representative the witness trees are of the actual vegetation in those areas, though the comparable estimates of abundance based on witness tree surveys and predictions from pollen (Fig. 15) suggest the surveys are reasonably representative as does the design of the original surveys and subsequent data processing (Cogbill et al. 2002).

For the pollen, the Dirichlet precision varies slightly between ponds because of the weighted summation introduced by the use of ϕ and the mixture model (3). The lack of pollen contribution from ponds outside the domain lowers the precision for ponds on the periphery. Averaging the pondwise posterior means, the precision for the modern surface sample data is 60.2 with a standard deviation of these pondwise means of 8.3. For the colonial data, the average posterior mean across ponds is 96.6 with a standard deviation of 6.6. The higher precision for the pollen data in the colonial era may reflect better estimation of the grid-scale vegetation with the large witness surveys, for which approximately three times as many trees were measured than in the FIA dataset, and possibly additional local vegetation heterogeneity in the modern era due to land use change

Note that the precision parameters for pollen and vegetation are not directly comparable to each other because of the very different sources of the data and different biological processes producing the compositions.

4.1.3 Spatial smoothing of composition data

By running the model for the modern and colonial eras, we can smooth the available vegetation data and provide estimates of colonial and modern vegetation in a visually appealing fashion, with associated uncertainty. In accounting for the count data structure, this simple application of the model has advantages over non-statistical smoothing and graphical display, allowing us to consider the ecological differences since European settlement in light

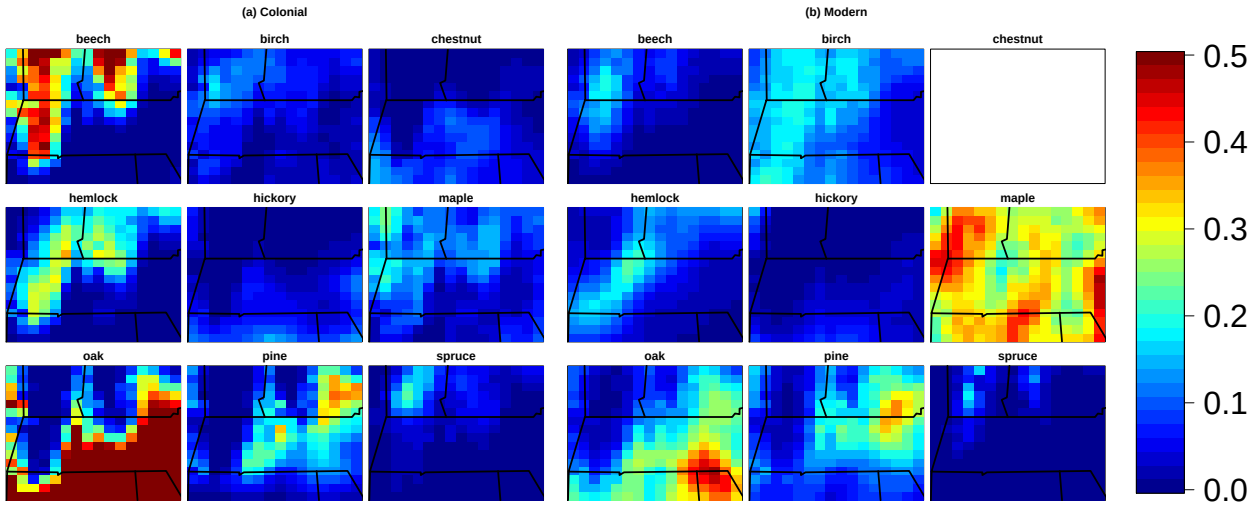

Figure 12. Posterior mean vegetation estimates from the (a) colonial and (b) modern estimation runs. Note that some beech and oak proportions are truncated to 0.5 in the colonial era.

of the estimated uncertainty. In Fig. 12, we see the smoothed composition estimates. Spatial gradients in vegetation appear to have become less distinct with European settlement, consistent with the ecological literature (Foster et al. 1998; Oswald et al. 2007). Fig. 13 shows uncertainty estimates for each taxon, suggesting that with the rich vegetation data of the FIA surveys we have reasonably precise estimates. However note that this is done at the grid level, and there is certainly a large amount of within-cell heterogeneity that causes individual stands of trees to have compositions that differ drastically from the composition estimate in a cell. The standard deviations are larger for more common taxa, but this reflects only that we can be quite certain in absolute terms that less common taxa are uncommon; the coefficient of variation (not shown) indicates that relative uncertainty is greater for the less common taxa and, for a given taxon, in locations in which the taxon is less common.

Our model relies on key parameters to translate between pollen data and vegetation predictions in the prediction runs. In particular, the variance component for the basis coefficients of the spatial process representation influences the amount of smoothing, which not only influences point predictions by determining the degree of local averaging, but perhaps more importantly determines the degree of uncertainty, with uncertainty increasing rapidly with increasing distance from ponds when the smoothing parameters specify more unsmooth spatial processes. As expected because of changes in vegetation post-settlement and sparser data in the modern surveys, the estimated heterogeneity is less in the modern era with an estimate of σ of 1.8 (1.2, 2.5) compared to 5.9 (4.1, 8.2) in the colonial era. This difference and the difference in the estimates of ϕ highlight the importance of choosing between the estimation parameters estimated for the modern and colonial periods when predicting in the past.

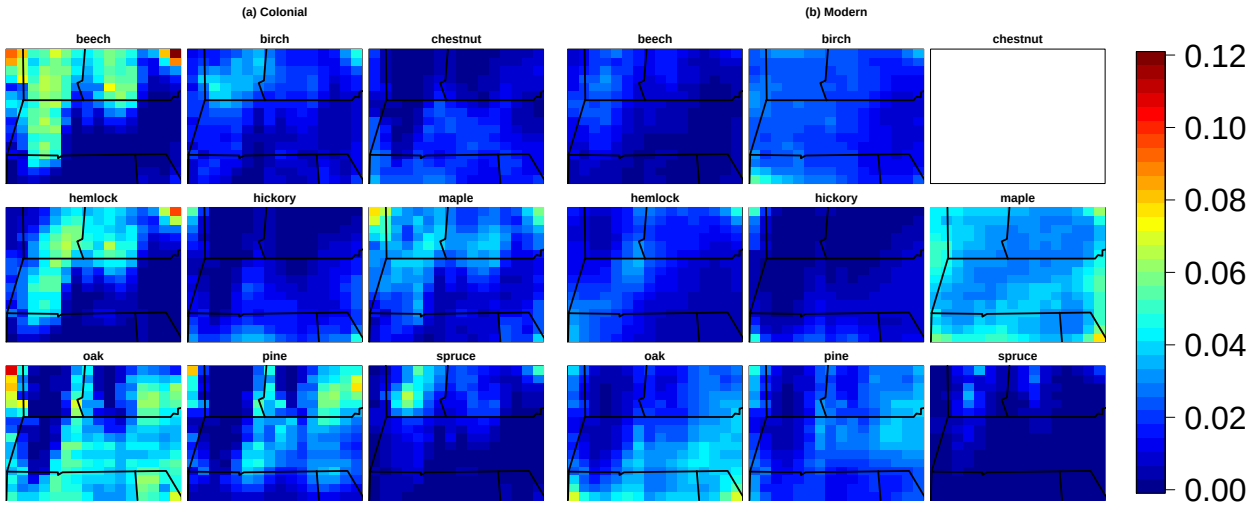

Figure 13. Posterior standard deviations of vegetation estimates from the estimation runs, using vegetation data, for (a) colonial and (b) modern eras.

4.1.4 Covariate effects

As shown in Fig. 12, spatial gradients in vegetation appear to have become less distinct with European settlement. We can also assess the possible change in associations between taxa and covariates affecting vegetation composition by comparing the elevation and latitude (as a proxy for climate, particularly temperature) effects in the colonial and modern models. Fig. 14 shows the estimated coefficients and uncertainty, suggesting that the covariate effects were generally more pronounced in the colonial era, as expected because of the influence of European colonists on vegetation through major land use changes, albeit with high uncertainty in both cases.

4.2 Cross-validation

Our estimation runs allow us to compare reasonable model specifications, but the true test of the model is its ability to predict and provide good uncertainty estimations for vegetation when only pollen data are available. Our next assessment uses cross-validation, first using modern parameter estimates to predict in the colonial period and then using colonial parameter estimates to predict in the modern period. Note that while only several hundred years apart, the modern and colonial eras are separated by vast ecological changes induced by European settlement, as great as any differences expected over the past 2500 years (Fuller et al. 1998; Oswald et al. 2007), so this provides an important check on the model. The results below suggest that our model is performing as well as may reasonably be expected, able to resolve many spatial patterns and temporal changes at coarse scale but missing the fine-scale details of vegetation and some coarse patterns.

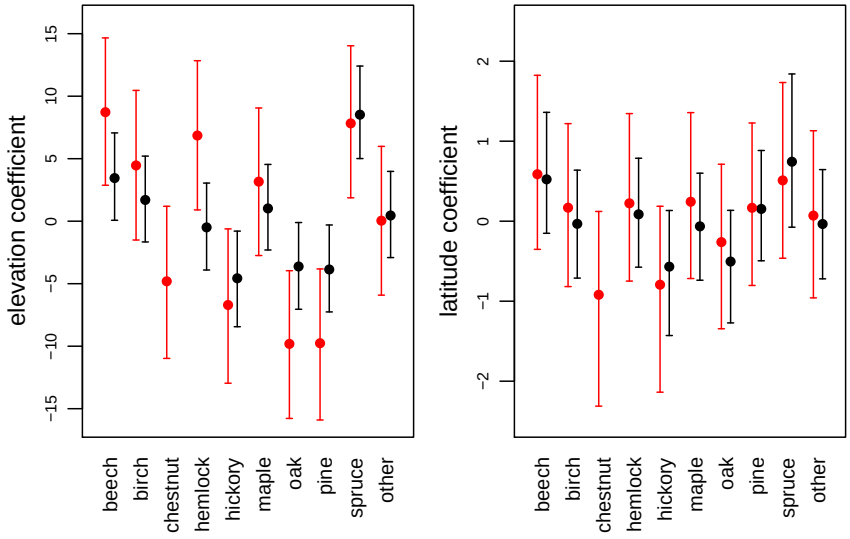

Figure 14. Regression coefficients for (a) elevation and (b) latitude with the colonial estimates in red and the modern estimates in black. Note that chestnut is not shown for the modern period.

4.2.1 Feature Significance

Uncertainty assessment is a major concern for a complicated model with an ambitious prediction goal. Pointwise standard deviations of prediction for each taxon provide some information about how certain we can be about our point predictions of vegetation composition, including which taxa are most reliably predicted and at which spatial locations we can be most certain. However, this does not give us a complete picture concerning our certainty about spatial features of the predictive surfaces. Such feature significance can take several forms, including inference about significant gradients (Banerjee et al. 2003), inference about significant extrema (Chaudhuri and Marron 1999), and rankings of location based on abundance (Laird and Louis 1989). Our primary interest lies in determining which spatial areas can be reliably determined to have higher or lower abundances of a given taxon than other areas, although detection of gradients and extrema may also be of interest. We also want to compare abundances of individual taxa across time and between taxa at individual times and spatial locations.

To make assessments about relative abundances of a single taxon across locations at fixed time, we make use of the posterior distributions of contrasts between different locations. Ideally, we would make joint statements about contrasts between different areas, aggregating over multiple grid cells. For example, we would like to be able to state, with 90% probability, that a group of specified grid cells has more oak than another group of grid cells. This approach might sacrifice too much power, so we may want to phrase our statements in terms of False Discovery Rate (FDR)-type criteria (Benjamini and Hochberg 1995), such as: with 95% probability, 90% of the pairwise comparisons of abundances between two groups of cells show higher abundance in the first group. In principle such statements can be made based on the full posterior distribution obtained from MCMC. Two

difficulties stand in the way. First, finding areas of significant difference requires searching a large dimensional space, particularly if one wants to make FDR-type statements. Greedy searches might provide one approach, but there would be no guarantee that the areas obtained were optimal in terms of providing the largest areas of difference with highest posterior probability. Second, results would focus attention on certain areas and provide little information about what can be said about differences between other areas.

We take a graphical approach focused on pointwise comparisons, hoping to provide approximate inference about all the locations at the expense of some loss of information about joint properties. In general, the use of exchangeable prior distributions with the resulting shrinkage justifies not adjusting for multiplicity (Berry and Hochberg 1999, Carlin and Louis 2000, p. 339, Gelman et al. 2008). We do not adjust for multiplicity because our spatial model has this flavor of exchangeability through the spatial process structure for the vegetation processes that smooths abundances towards each other, potentially giving flat surfaces ($\sigma^2 \approx 0$) if the data suggest little spatial variability. The prediction model also smooths in time.

Our approach is to conduct pairwise tests for differential abundance for a given taxon and plot the results in an informative way, demonstrated in the third column of plots in Fig. 15. For each pair of grid cells, we compute the posterior probability that the abundance of the taxon in one grid cell is higher than the abundance in the other grid cell. We sequentially consider each grid cell as the focal cell, making a subplot in which we color the other grid cells for which pairwise differences between the focal cell and the other cell have at least 90% posterior probability of lying on one side of zero. The colors indicate the sign of the difference and the posterior probability of lying on that side of zero. Finally we make a mosaic of the subplots, with the subplot placed on a map in the position of the focal grid cell and an 'x' marking the relative position of the focal location within the subplot. By tiling the subplots into a full plot, we present a color map of pointwise, pairwise probabilities of differential abundance. Viewing the mosaic of subplots as a single plot, areas of substantial probability of differential abundance from other areas show themselves as deep colors, while individual subplots can still be examined to assess differences between a given focal location and all other locations. Note that to preserve the scaling of the full plot, each subplot needs to be square, even if the full region is not square, inducing some spatial distortion within the subplots. In Fig. 15 we see that for beech, the northwestern and north-central areas indicated in dark red show high probability of higher abundance than the south-central and eastern areas. In contrast, for hickory, the evidence is less strong, with moderate probability of a small area (in blue) in the northwest having lower abundance than most of the rest of the region.

The posterior probabilities are calculated as an expectation, over the 50 sampled values of the estimation model parameters, of the posterior probability of one cell's abundance exceeding the other cell,

$$\int P(r_p(s_i) > r_p(s_j)|\zeta_{\text{est}}, Z_{\text{pred}})\pi(\zeta_{\text{est}}|Z_{\text{est}})d\zeta_{\text{est}} \approx \sum_{k=1}^{50} P(r_p(s_i) > r_p(s_j)|\zeta_{\text{est},k}, Z_{\text{pred}}),$$

where Z_{est} and Z_{pred} are the data from the estimation and prediction runs and ζ_{est} is the set of estimation parameters and where the probability on the right-hand side is computed as the average over the posterior samples from the prediction model run with parameters $\zeta_{\text{est},k}$.

Similar reasoning with regard to multiple testing, based on the spatio-temporal smoothing in the model, applies to comparisons over time for a fixed taxon, demonstrated in Section 4.2.3. We can take an analogous approach to compare abundance between two taxa for a single time period, finding posterior probabilities of differential abundance in each grid cell, although the difficulty in estimating ϕ makes it harder to detect differences between taxa.

4.2.2 Assessment of Predicted Surfaces and Uncertainty Characterization

In Fig. 15 for the colonial period, we compare our best estimate of vegetation, from the colonial estimation run, with predictions based on pollen from the colonial period and parameter estimates from the modern period. We also show feature significance plots and plots of posterior standard deviations of prediction to assess uncertainty. Fig. 16 does the same for the modern period, using modern pollen and colonial parameters. Based on comparisons of the vegetation-predicted surfaces with the pollen-predicted surfaces, interpreted in light of the feature significance plots, it appears the model is doing a reasonably good job of predicting spatial patterns. For the colonial period, the features are quite similar in the prediction and estimation runs, and patterns detected in the feature significance plots are seen in the vegetation-predicted surfaces when considered at a fairly coarse resolution, suggesting minimal type one error, with few non-existent patterns detected. For the modern predictions, the results are not as good, particularly for hemlock. The model fails to capture some large-scale patterns and is overly confident about the patterns it does estimate. The poorer results in predicting modern vegetation may more strongly reflect difficulties in predicting modern vegetation than problems with the colonial parameter estimates per se. Land-use change post-colonization makes spatial patterns in modern vegetation less distinct and less strongly associated with the covariates than in the colonial era (Foster et al. 1998; Fuller et al. 1998), making prediction more difficult. The larger precision of the pollen data as a proxy for vegetation in the colonial estimation runs than in the modern estimation runs causes overconfidence in the modern predictions. Given that vegetation structure in the past is very likely to be more similar to the colonial vegetation than the modern vegetation, the success of the model in predicting colonial patterns gives us confidence in the ability of the model to make predictions.

In terms of absolute abundance, the model generally indicates the taxa with high and low abundance reasonably (maple is an exception) but often incorrectly predicts overall relative abundance of a taxon. This relates to whether the estimated values of ϕ are appropriate for the time period. In particular, maple is overpredicted in the colonial era and underpredicted in the modern, while the reverse is true for oak and beech. This occurs because the estimated values of ϕ for the two taxa are different for the two eras (Fig. 8); use of the parameter estimates from estimation runs for the same era as the prediction runs improves prediction of the overall level, indicating the sensitivity of predictions to this key parameter. Given that we expect vegetation before settlement to be more similar to the colonial vegetation than the modern vegetation, this suggests we should focus on the colonial parameter estimates for prediction before settlement.

Posterior uncertainty varies widely between taxa and across space. Uncertainty is the greatest far from ponds, as should be the case. Maple is particularly uncertain, because the

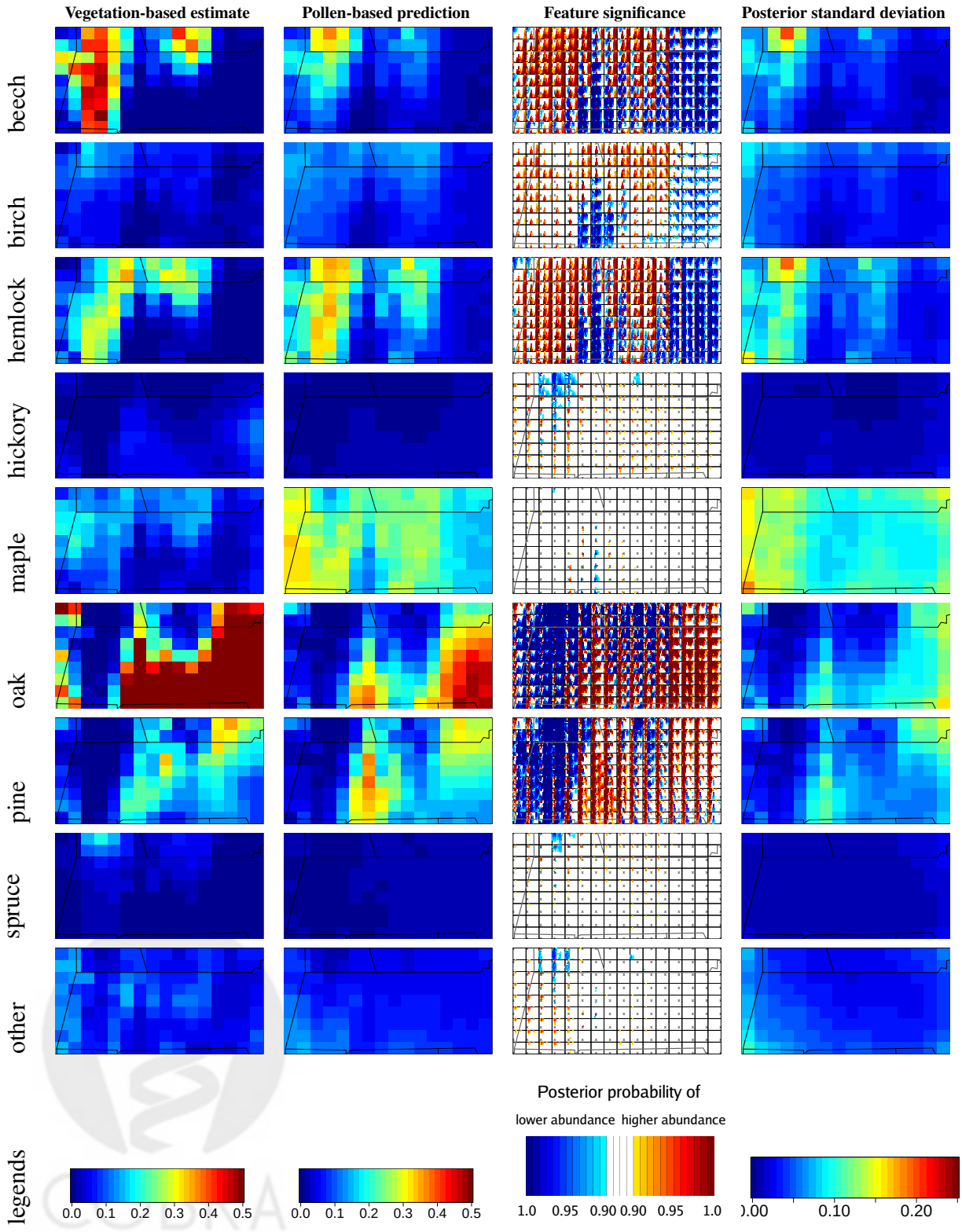

Figure 15. For the colonial period, vegetation estimated in the colonial estimation run (first column), vegetation predicted in the colonial prediction run based on colonial pollen and modern parameter estimates (second column), feature significance for the prediction run (third column) and posterior prediction standard deviations (fourth column). Note that chestnut is not shown as we cannot obtain parameter estimates for the modern era with no mature chestnut. In the first column, some cell abundances are truncated to 0.5.

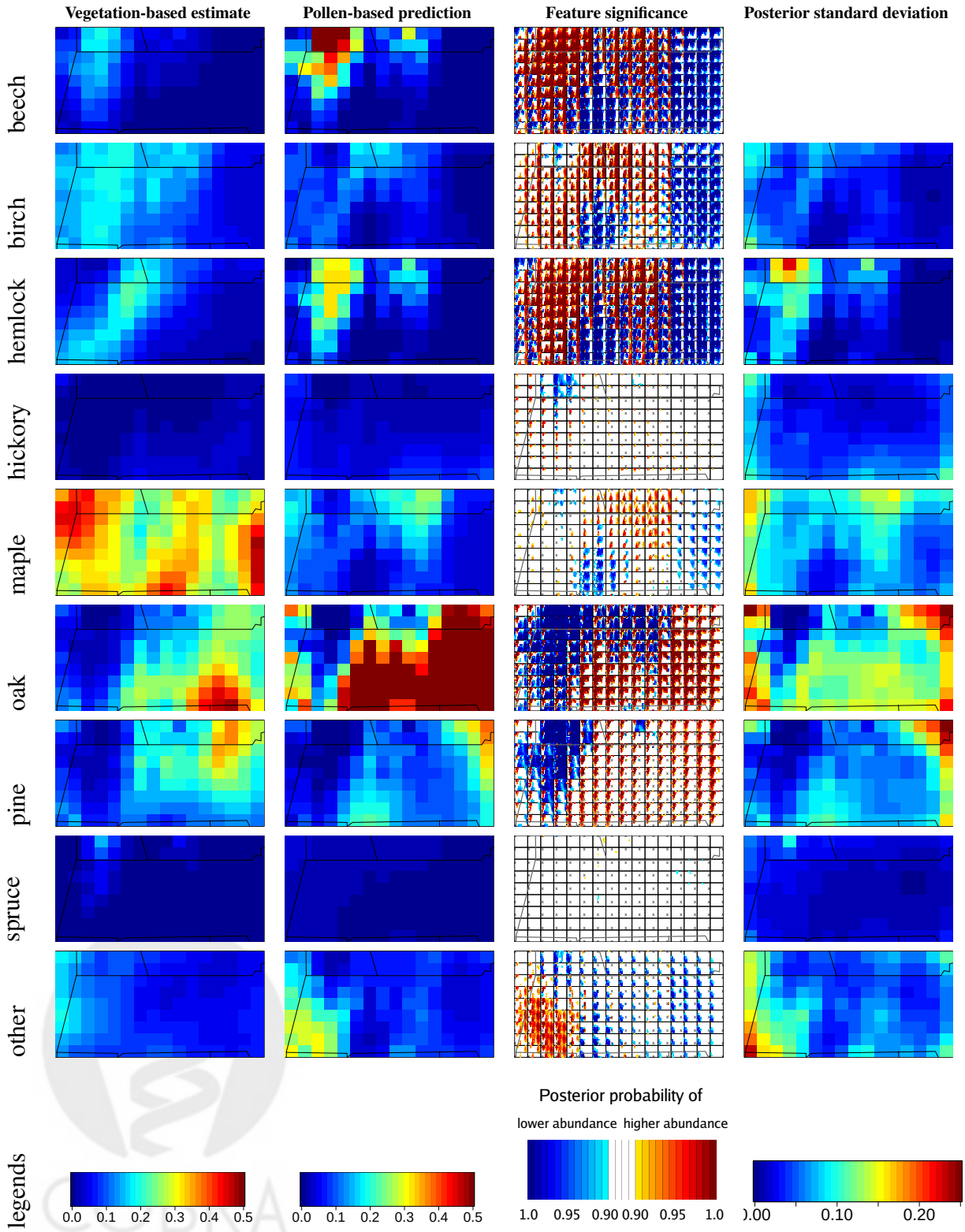

Figure 16. For the modern period, vegetation estimated in the modern estimation run (first column), vegetation predicted in the modern prediction run based on modern pollen and colonial parameter estimates (second column), feature significance for the prediction run (third column) and posterior prediction standard deviations (fourth column). Note that chestnut is not shown because there are no mature chestnut in the modern era. In the second column some cell abundances are truncated to 0.5 and in the last column some standard deviations to 0.25.

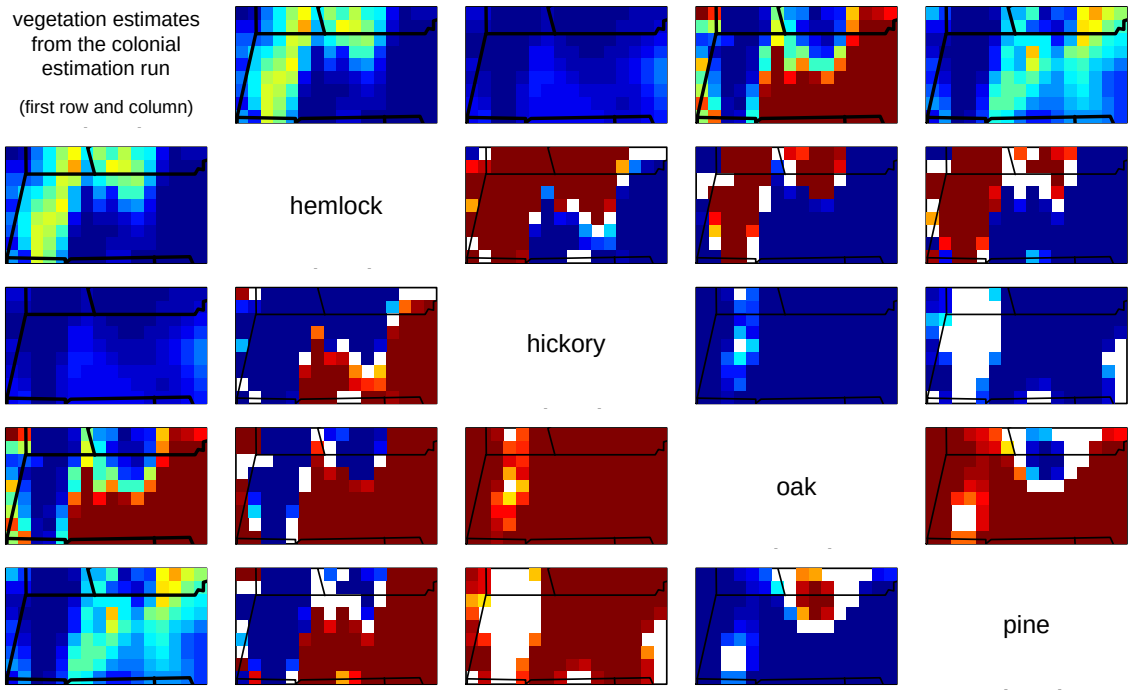

Figure 17. Posterior probabilities of differences in abundance between selected taxa in the colonial era based on the cross-validation run in which the modern parameter estimates were used with the colonial pollen data. The first row and first column plot the estimated vegetation composition for the selected taxa based on the colonial estimaton run. Each cell within the table of figures indicates the posterior probability (with a threshold of 90%) that the taxon distinguished by the row had higher (red) or lower (blue) abundance than the taxon distinguished by the column. For example hemlock appears to be more abundant than hickory over the north-central and western areas (indicated by the red) but less abundance in the south-central and eastern portions. See Fig. 15 for legend.

low pollen production/dispersal of maple causes inference about maple to rely on a small number of pollen grains in each pond, creating a large signal to noise ratio. In contrast, the predictions for rare taxa are relatively certain. Note that considered relative to the small magnitude of the proportions (e.g., using the coefficient of variation), the relative uncertainty in the rare taxa may be large, which would affect our ability to make statements about range boundaries for rare taxa. We recognize that we can say little about relative differences in rarity.

We can also compare abundance between taxa at individual grid cells. In Fig. 17 we show some example comparisons between taxa for the colonial era. These suggest differences between taxa can often be determined.

4.2.3 Assessment of temporal contrasts

To assess temporal changes across time, we can contrast abundance estimates for each taxon between any pair of time points on a pointwise cell by cell basis, again without post

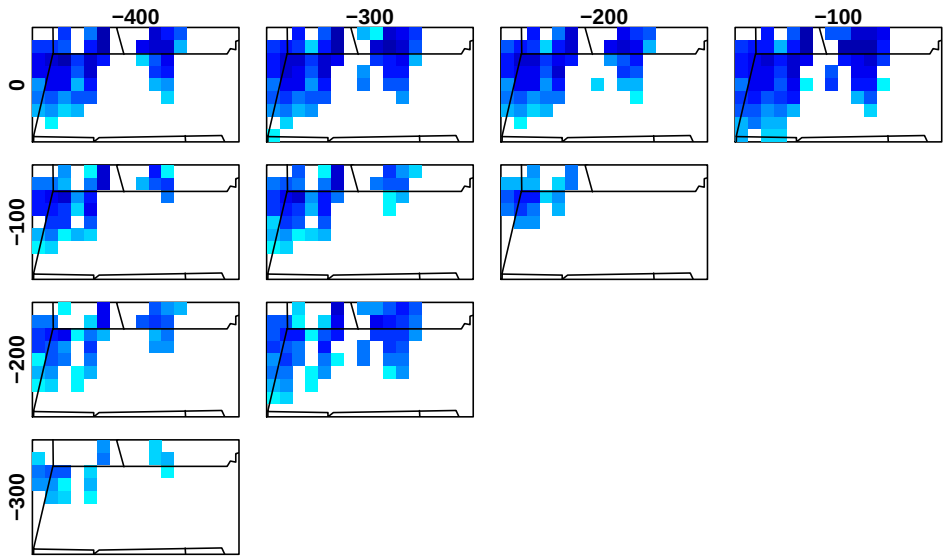

Figure 18. Posterior probabilities of differences in recent beech abundance between pairs of time points 0 to 400 years before present (1950) based on modern parameter estimates. Each cell indicates the posterior probability (with a threshold of 90%) that the cell has lower (blue) or higher (red - no examples here) abundance in the later time period than the earlier time period. See Fig. 15, third column for color legend.

hoc correction for multiple testing because of the temporal smoothing done by the model. The comparison takes the simple form, at each grid cell, of computing the posterior probability for the chosen taxon that the abundance is greater in one period than a second period. If this posterior probability exceeds a threshold (we use 90% in our plots) for either period, we plot the posterior probability as the color shade in the cell. As an example, Fig. 18 shows distinct changes in beech over time when comparing the present (i.e., 1950), denoted as year 0, with the years 100 to 400 years before present. The areas of predicted robust decrease in beech match estimated declines based on the colonial and modern vegetation data seen in Fig. 12 (but note that this is not full cross-validation given our use of the modern estimation run parameters). Assessment of this and other such contrast plots in light of the modern and colonial vegetation data suggests the model can detect changes over time with reasonable specificity and some sensitivity to real changes in composition.

4.3 Operational prediction model results

Here we describe initial results from the prediction runs over the past 2500 years. Many other uses of the full posterior distribution are possible. For display purposes, we use predictions based on colonial parameter estimates; more detailed ecological analysis will assess robustness with respect to the estimation run used. We also considered running the prediction model at time intervals of 50 years and with an exponential temporal correlation function. For most aspects of the predictions these changes had little effect, but there was some sensitivity in the temporal contrasts.

The parameter estimates for the temporal variance components indicate that changes over time occur smoothly, particularly for the regression coefficients, but also for the residual spatio-temporal structure.

4.3.1 Surface predictions, feature significance, and uncertainties

Figs. 19 and 20 shows surface predictions, feature significance, and posterior standard deviations for oak and beech, respectively, at 500-year intervals between 500 and 2500 years before present. These plots allow inference about potential spatial patterns and shifts over time. For these taxa, the spatial patterns in the posterior means seem to be robust, with little apparent change over time for oak, but potential declines in abundance of beech, which could be assessed more confidently with temporal contrast plots.

In Fig. 21, we show temporal contrast plots for the pre-settlement period using colonial parameter estimates for oak, which suggest that there was a trend toward increased oak abundance in higher-elevation areas in central Massachusetts in the period 1000 to 500 years before present and a decreasing trend in roughly the same area 2500 to 2000 years before present. There are no detectable trend in areas furthest to the east that have high oak abundance. One concern is that with only 23 ponds, spatial predictions may be sensitive to the loss of ponds as the number of ponds with data available drops over time (Fig. 2a). Sensitivity analysis could be done using predictions from runs that rely only on the fixed set of ponds available for the entire time window of interest.

4.3.2 Location-specific time trends

A common presentation of pollen data is in the form of a pollen diagram showing changes in pollen composition in a pond over time. An analogous presentation using model output is to estimate vegetation composition and associated uncertainty in a given grid cell, demonstrating the ability of the model to estimate vegetation based on pollen with uncertainty estimates. In Fig. 22 we compare pollen composition to model-estimated vegetation, including a decomposition into the average across time and temporal deviation from that average that allows assessment of contrasts across time. Maple, chestnut, spruce, and hickory pollen are all represented at low abundance throughout the period of interest. After accounting for taxon-specific biases in pollen representation and borrowing strength both spatially and based on environmental covariates, the model provides some evidence that birch, chestnut, maple, oak, and spruce increased over time pre-settlement, while the more common beech and hemlock do not show a robust trend. In general the model-estimated trends in the grid cell match those from the pollen in the single pond, but spatial smoothing in the model can cause differences between raw pollen and estimated vegetation, such as seen for maple.

4.3.3 Covariate effect trends

In Fig. 23 we plot the covariate effects at all times, with uncertainty, showing how our estimates of the relationships between vegetation abundance and elevation and latitude have varied over time. The estimates have a large amount of uncertainty, but when one considers

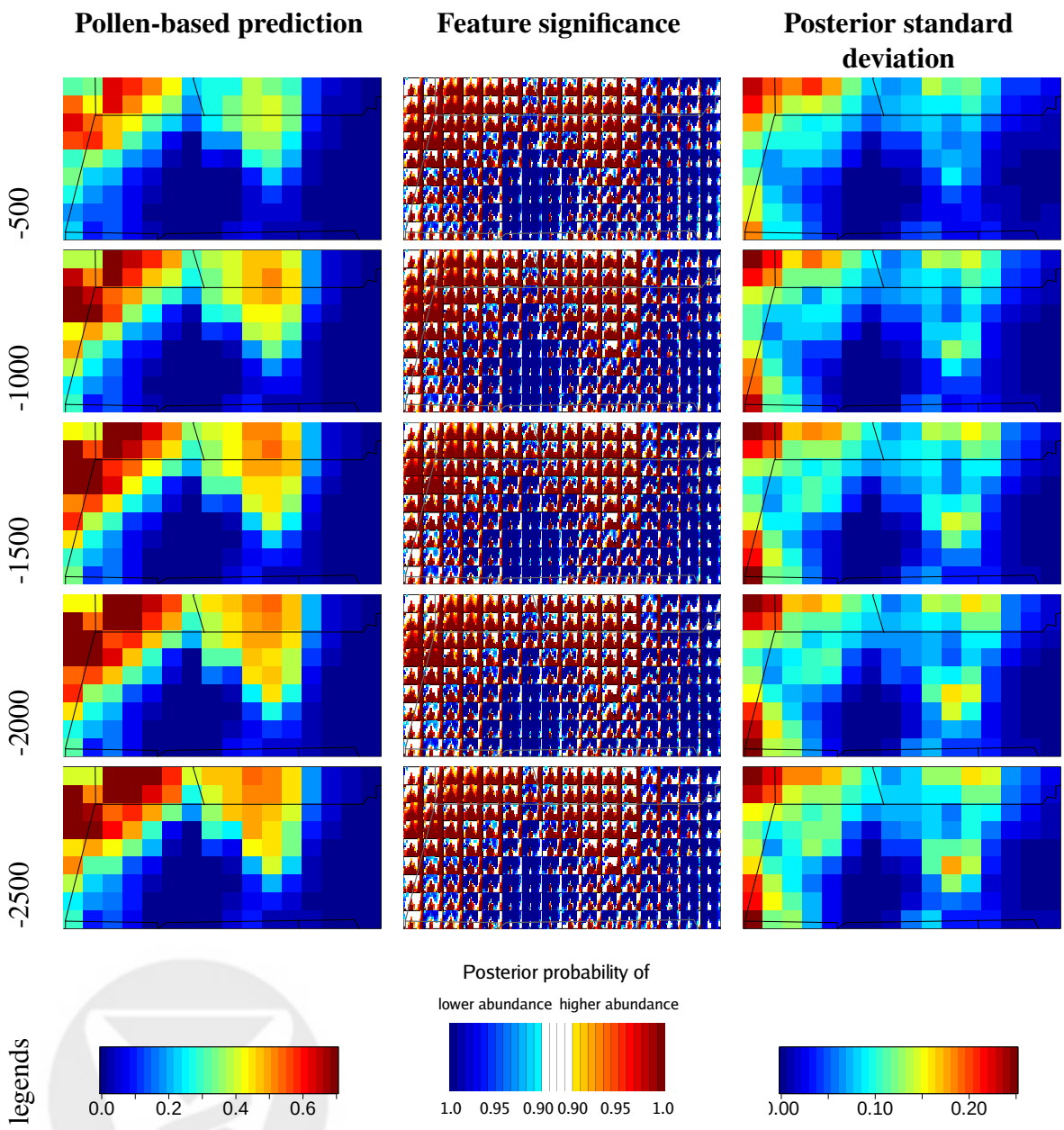

Figure 19. For beech, vegetation prediction information, based on colonial parameter estimates, at 500 year intervals, starting 500 years before present (top row) and ending 2500 years before present (second from bottom row). Plots are the posterior mean vegetation abundance (first column), feature significance (second column) and posterior prediction standard deviations (third column). In the first and third columns the values in some cells are truncated.

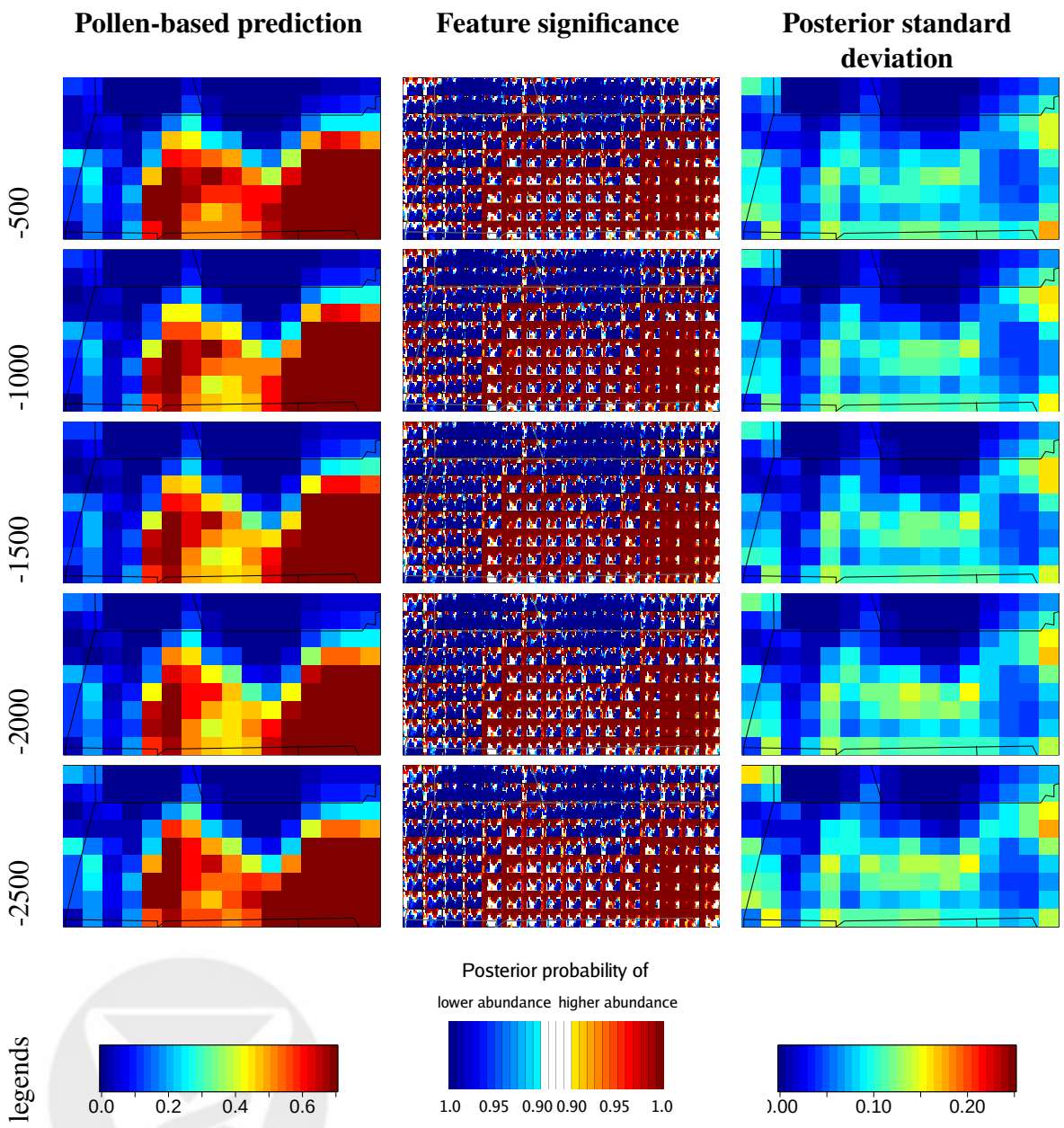

Figure 20. For oak, vegetation prediction information, based on colonial parameter estimates, at 500 year intervals, starting 500 years before present (top row) and ending 2500 years before present (second from bottom row). Plots are the posterior mean vegetation abundance (first column), feature significance (second column) and posterior prediction standard deviations (third column). In the first column the values in some cells are truncated.

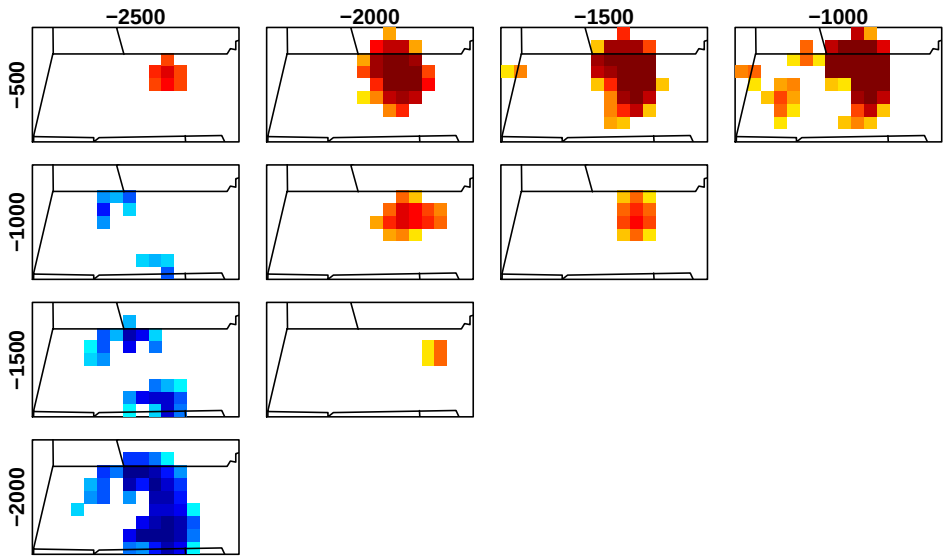

Figure 21. Posterior probabilities of differences in pre-settlement oak abundance between pairs of time points 500 to 2500 years before present (1950) based on colonial parameter estimates. Each cell indicates the posterior probability (with a threshold of 90%) that the cell has lower (blue) or higher (red) abundance in the later time period than the earlier time period. See Fig. 15, third column for color legend.

uncertainty in the coefficient estimates relative to the average of the coefficient over time, uncertainty decreases and there appears to be some limited evidence for changes over time, for example, for beech with respect to elevation. The wiggliness in the confidence bands in some plots is caused by a small number of prediction runs in which the estimated value of the correlation decay parameter is small. Note that the asymmetries are caused primarily by the averaging across relatively symmetric individual posteriors from the 50 prediction runs. As mentioned above, over time the loss of ponds with substantial leverage could contribute to any temporal changes seen here.

5 Discussion

Almost 100 years ago, von Post (1917) described the problem of interpreting forest composition from fossil pollen assemblages. Long-distance pollen dispersal, differential pollen production, and a generally high level of process noise have continued to be major obstacles for the interpretation of paleoecological data ever since. Analyses of pollen data have identified important trends in the data (Berglund 1991; Davis et al. 1998; Fuller et al. 1998; Soepboer et al. 2007), but they have not quantified these trends in an inferential framework that explicitly accounts for the various sources of uncertainty and the natural spatial context of the data. Although theory and models about pollen production, dispersal, and accumulation have continuously evolved (Webb 1974; Jackson 1990; Davis 2000; Haslett et al. 2006; Sugita 2007a,b), most paleoecological literature simply presents raw pollen percentages and asks the reader to understand that these are rough and unquantified

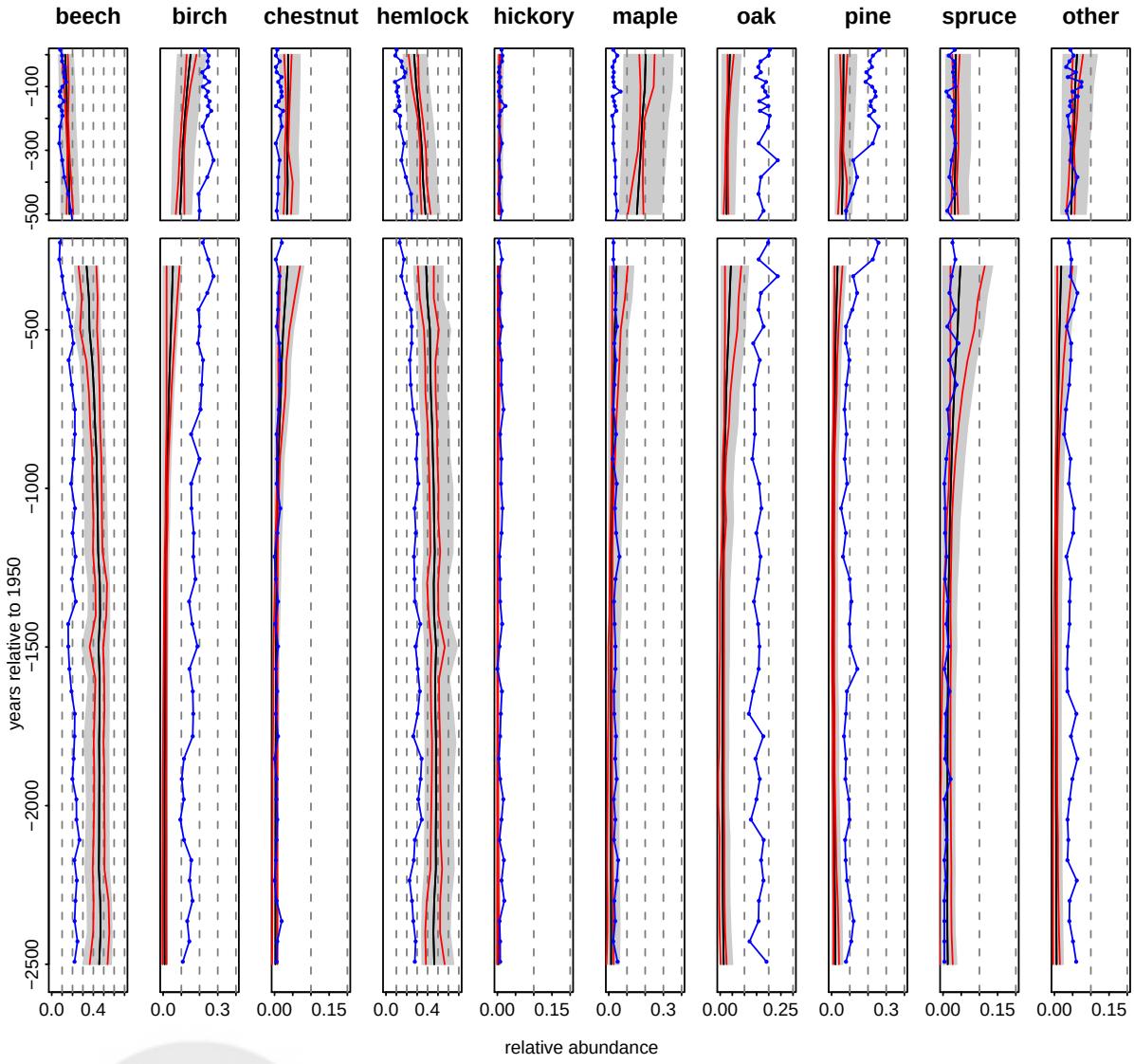

Figure 22. Vegetation diagrams for the grid cell encompassing pond 20 (Snake Pond) with the recent period based on modern parameter estimates (top; 0 to 500 years before present) and the pre-settlement period based on colonial parameter estimates (bottom, 300 to 2500 years before present). Black lines represents the posterior mean and gray shading the 95% credible intervals for vegetation abundance, $r_p(s,t)$, with blue lines showing corresponding pollen proportions from Snake Pond. Red lines represent 95% pointwise credible intervals for the deviations over time in the vegetation ($r_p(s,t) - \bar{r}_p(s)$), which are plotted as offsets relative to the posterior mean of $\bar{r}_p(s)$. Plotting the credible interval for the deviation in this way removes the effect of uncertainty in ϕ_p, which affects all times in the same way, and avoids the overly conservative contrasts of abundance across time indicated by the gray shading.

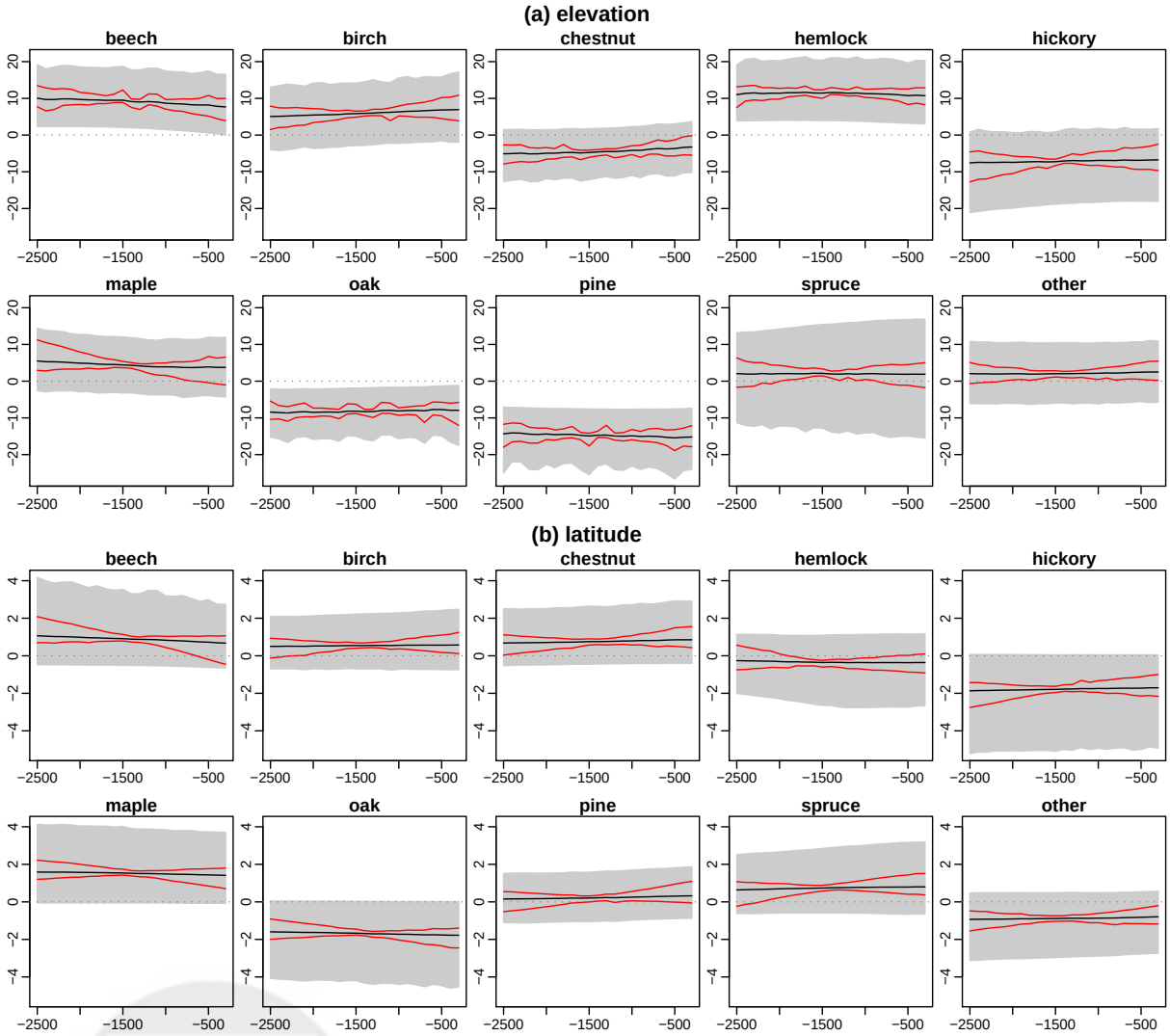

Figure 23. Time plots of posterior mean coefficient estimates (black lines) with pointwise 95% credible intervals (grey shading) for elevation (a) and latitude (b) regression coefficients. The x-axes denote years relative to 1950. Red lines represent 95% pointwise credible intervals for deviations over time in the coefficients ($\beta_{k,p}(t) - \bar{\beta}_{k,p}$), which are plotted as offsets relative to the posterior mean of $\bar{\beta}_{k,p}$. Plotting the credible interval for the deviation around the mean in this way removes the effect of uncertainty in the overall magnitude of the coefficient, which affects all times in the same way, and avoids the overly conservative contrasts of abundance across time indicated by the gray shading

approximations of the forest composition, which is the real variable of interest.

Our work tackles this problem, building a statistical framework for inferring historical forest composition based on proxy pollen sediment data. We present a multivariate spatio-temporal model for compositions. We build the model in stages, allowing easier model assessment and avoiding highly complicated space-time formulations that may be difficult to understand and assess. Under a set of simple assumptions about the relationship between trees and pollen through space and time, our model adds a quantitative estimate of uncertainty to inference about changing vegetation, providing the first spatially-explicit statistical analysis of paleoecological data, and borrowing strength across multiple ponds and across time in a coherent way. Innovative graphical assessments of feature significance based on the full posterior distribution suggest that the pollen data can reliably indicate certain large-scale spatial features for some taxa, but that features on scales smaller than ~50 km are not possible to distinguish, nor are large-scale features for some taxa, such as those with low pollen production/dispersal relative to other taxa. The model does not resolve the substantial problems involved in using pollen data to estimate forest composition, but does suggest which inferences are more reliable and what additional data would be most helpful.

Specific results from the model demonstrate the advantages of the spatially-explicit modeling approach that calibrates pollen to vegetation. For example, Fig. 22 estimates the extent to which a classic pollen diagram misrepresents changing forest composition. Most paleoecological studies (e.g., Fuller et al. 1998) would show only the blue pollen proportions and interpret forest change by acknowledging that the representation of certain tree taxa is likely to be biased in pollen data. Our analysis quantifies this in a coherent probabilistic framework. The recent decrease in beech trees suggested in Fig. 22 is depicted in a regional context in Fig. 21. Previous studies (Fuller et al. 1998; Oswald et al. 2007) have identified this regional decrease, but were unable to describe this trend in a continuous spatial setting. Graphical representations of output from our model allow a resolution of spatial analysis previous unavailable to paleoecologists. More importantly, confidence in the strength of the inferences is articulated. A similar set of maps for maple (not shown) shows very little significant trend, due to the large uncertainty about maple abundance. Given the amount of noise in the pollen representation and the relative sparseness of fossil pollen datasets, it is important that paleoecologists are able to confidently detect patterns emerging above the noise in their data.

The results also suggest particular avenues for future data collection that would help refine our estimates and predictions. Our assessment of the model suggests that two critical areas for data collection are 1.) to improve estimation of ϕ with vegetation data collected close to each pond, and 2.) to improve spatial prediction and estimation of covariates with additional ponds. The first area for data collection is feasible. The mismatch between pollen composition and estimated vegetation composition in the cell in which each pond resides is a critical area for improvement. At this point it is difficult to distinguish long-distance dispersal contributions to the pollen data from contributions from local vegetation that is more similar to the vegetation in the larger spatial region than the grid cell. The result is to make it difficult to distinguish long-distance dispersal from local within-cell heterogeneity. Several types of data could shed light on this and allow us to better estimate the long-distance contribution. First, field surveys near ponds could help estimate local vegetation. Also, additional vegetation data from established surveys, in particular

Massachusetts state parkland data, may help to better estimate grid cell vegetation. Getting many additional cores would be difficult: the current dataset represents decades of effort by paleoecologists and is one of the most extensive sets of such data available, but it may be possible to add a small number of additional ponds, selected based on examining the model output to see which ponds would add the most information. In addition, sampling of ponds that are likely to be anomalous in their pollen in such a way that the pollen is from less common taxa (e.g., from ponds at high elevation relative to their grid cells) would better allow us to distinguish the relative contributions of local and long-distance pollen, thereby distinguishing the contribution of local heterogeneity and long-distance transport to the pollen-vegetation mismatch. This is because the anomalous pollen in these ponds would generally not be more similar to the larger spatial region than the grid cell, preventing the model from attributing the pollen to long-distance transport. Additional covariates may help to better predict vegetation. For example, soil type and information about local topographic patterns may be useful. Covariates defined at the actual locations of the ponds and FIA plot data may help us to explain local heterogeneity and distinguish local effects from long distance pollen transport. In particular, the difference between pond elevation and the average elevation of the grid cell as a whole may help explain part of the local heterogeneity in pollen, accounting for some of the anomalies we have identified. However, our ability to estimate covariate effects from the pollen data is limited by the number of ponds with data.

In the modern era, we have used counts of trees larger than 10 cm DBH from the FIA surveys as our vegetation data, but basal area (total cross-sectional area of trees) is available. Because pollen production scales with tree size, basal area is likely to be more closely associated with pollen production than tree counts. While exploratory plots suggest there may not be a major difference between the two types of data, it would be worthwhile to create a likelihood function for basal area in place of the likelihood based on FIA counts. However, note that our estimates of ϕ do account for different average tree sizes between taxa as this effect is reflected in the pollen. The main advantage of accounting for basal area would be if the relative sizes of trees of the different taxa vary spatially in a way that cannot be accounted for by space-invariant values of ϕ. The modeling difficulty is that the basal area distribution is strongly zero-inflated. This occurs because of local vegetation heterogeneity, the discrete nature of trees and the fact that small trees are not included in the surveys (although there are separate plots for smaller trees that could be included in the analysis). This requires a distribution that mixes a mass at zero with a continuous distribution truncated at the smallest included basal area of a single tree. Such a distribution is non-standard and is likely to be more highly-parameterized than our current Dirichlet-multinomial distribution. In addition, the multinomial structure of the count data naturally more heavily weights plots with more trees; a distribution for the continuous basal area does not necessarily have this effect. Also note that the use of basal area would make it more difficult to compare between the colonial and modern estimation runs as the witness tree data are available only as counts.

We plan to apply the model in other domains. Michigan is an area of particular interest because it covers an important vegetation gradient, and we believe the model could be directly applied to that dataset. Second, we are interested in application of the model to eastern North America post-glaciation. While the density of ponds is less than in our current

application, the larger spatial differentiation in vegetation allows us to consider vegetation dynamics on a larger scale and to consider post-glaciation dynamics, tree migration into areas vacated by the glaciers, and species range boundaries over time. However, analyses presented here are based on recent millennia of forest change. Over this time period, forest structure and composition are not likely to have deviated far beyond our reference vegetation data from modern and early colonial times (Oswald et al. 2007). Extending such analyses farther back in time would eventually strain this relationship (Jackson and Williams 2004), and the estimates of key parameters from our estimation runs may no longer be reasonable. In such contexts, it will be important to either ensure that the relationship between pollen and vegetation remains consistent over time or to somehow account for differences, potentially using parameter information from other datasets reflecting a wider array of climate conditions or based on informative prior distributions. Finally, fossil pollen spectra are increasingly used to complement population genetic data to understand past population shifts (McLachlan et al. 2005; Magri et al. 2006). Current efforts at melding these complementary data sources are hampered by the difficulty of inferring tree population size from networks of fossil pollen data. Our model can provide estimates of population size over time including uncertainty estimates, which should help make integrating paleoecological and genetic data more straightforward.

The long-term and broad-scale nature of modern environmental problems ensures that networks of paleoecological sites will continue to provide important benchmarks for environmental change (Botkin et al. 2007). Our model provides the framework for testing ecological theory through the incorporation of covariates and through its ability to distinguish important spatial trends from noise. The model was designed with few biological assumptions, but it could be modified to incorporate such constraints, as well as additional data such as more finely-specified pollen dispersal data, environmental covariates, or spatial genetic information, as these data become available. We anticipate that our work, along with parallel efforts by others to interpret paleoecological data in better articulated statistical terms (Haslett et al. 2006; Sugita 2007a,b), will allow this longstanding data source to be better integrated into modern environmental analysis.

References

Aitchison, J. (1986), *The Statistical Analysis of Compositional Data*: Chapman & Hall Ltd.

Banerjee, S., Carlin, B., and Gelfand, A. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton, Florida: Chapman & Hall.

Banerjee, S., Gelfand, A. E., and Sirmans, C. F. (2003), "Directional Rates of Change Under Spatial Process Models," *Journal of the American Statistical Association*, 98(464), 946–954.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society, Series B*, 57, 289–300.

Berger, J. O., De Oliveira, V., and Sansó, B. (2001), "Objective Bayesian Analysis of Spatially Correlated Data," *Journal of the American Statistical Association*, 96(456), 1361–1374.

Berglund, B. (1991), *The Cultural Landscape During 6000 Years in Southern Sweden: The Ystad Project*: Blackwell Publishing.

Berry, D. A. and Hochberg, Y. (1999), "Bayesian Perspectives on Multiple Comparisons," *Journal of Statistical Planning and Inference*, 82, 215–227.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995), "Bayesian Computation and Stochastic Systems (Disc: P41-66)," *Statistical Science*, 10, 3–41.

Billheimer, D., Cardoso, T., Freeman, E., Guttorp, P., Ko, H.-W., and Silkey, M. (1997), "Natural Variability of Benthic Species Composition in the Delaware Bay," *Environmental and Ecological Statistics*, 4, 95–115.

Blaauw, M. and Christen, J. A. (2005), "Radiocarbon Peat Chronologies and Environmental Change," *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 54(4), 805–816.

Botkin, D., Saxe, H., Araujo, M., and et al. (2007), "Forecasting the Effects of Global Warming on Biodiversity," *Bioscience*, 57, 227–236.

Buck, C., Gomez Portugal Aguilar, D., Litton, C., and O'Hagan, A. (2006), "Bayesian Non-Parametric Estimation of the Calibration Curve for Radiocarbon Dating," *Bayesian Analysis*, 1, 265–288.

Bunting, M. and Middleton, D. (2005), "Modelling Pollen Dispersal and Deposition Using HUMPOL Software, Including Simulating Windroses and Irregular Lakes," *Review of Palaeobotany and Palynology*, 134, 185–196.

Carlin, B. P. and Louis, T. A. (2000), *Bayes and Empirical Bayes Methods for Data Analysis*: Chapman & Hall Ltd.

Chaudhuri, P. and Marron, J. (1999), "SiZer for Exploration of Structures in Curves," *Journal of the American Statistical Association*, 94, 807–823.

Christensen, O., Roberts, G., and Sköld, M. (2006), "Robust Markov Chain Monte Carlo Methods for Spatial Generalized Linear Mixed Models," *Journal of Computational and Graphical Statistics*, 15, 1–17.

Cogbill, C., Burk, J., and Motzkin, G. (2002), "The Forests of Presettlement New England, USA: Spatial and Compositional Patterns Based on Town Proprietor Surveys," *Journal of Biogeography*, 29, 1279–1304.

Davis, M. (1981), "Quaternary History and the Stability of Forest Communities," in *Forest Succession: Concepts and Application*, eds. D. West, H. Shugart, and D. Botkin, Springer-Verlag, pp. 132–153.

——— (2000), "Palynology After Y2K / Understanding the Source Area of Pollen in Sediments," *Annual Review of Earth and Planetary Sciences*, 28, 1–18.

Davis, M., Calcote, R., Sugita, S., and Takahara, H. (1998), "Patchy Invasion and the Origin of a Hemlock-Hardwoods Forest Mosaic of Pollen in Sediments," *Ecology*, 79, 2641–2659.

Delcourt, P. and Delcourt, H. (1987), *Long Term Forest Dynamics of the Temperate Zone: A Case Study of Late-Quaternary Forests in Eastern North America*: Springer-Verlag.

Dey, D. and Maiti, T. (2002), "Dirichlet Multinomial Distribution," in *Encylopedia of Environmetrics*, eds. A. El-Shaarawi and W. Piegorsch, Stuttgart: Fischer, pp. 522–523.

Foster, D., Motzkin, G., and Slater, B. (1998), "Land-Use History As Long-Term Broad-Scale Disturbance: Regional Forest Dynamics in Central New England," *Ecosystems*, 1, 96–119.

Fuentes, M. and Raftery, A. (2005), "Model Evaluation and Spatial Interpolation by Bayesian Combination of Observations with Outputs from Numerical Models," *Biometrics*, 61(1), 36–45.

Fuller, J., Foster, D., McLachlan, J., and Drake, N. (1998), "Impact of Human Activity on Regional Forest Composition and Dynamics in Central New England," *Ecosystems*, 1, 76–95.

Gelfand, A., Silander, J., Wu, S., Latimer, A., Lewis, P., Rebelo, A., and Holder, M. (2006), "Explaining Species Distribution Patterns Through Hierarchical Modeling," *Bayesian Analysis*, 1, 41–92.

Gelman, A. (2006), "Prior Distributions for Variance Parameters in Hierarchical Models (comment on Article by Browne and Draper)," *Bayesian Analysis*, 1(3), 515–534.

Gelman, A., Hill, J., and Yajima, M. (2008), "Why We (Usually) Don't Have to Worry About Multiple Comparisons," Technical report, Department of Statistics, Columbia University.

Green, P. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

Haslett, J., Whiley, M., Bhattacharya, S., and et al. (2006), "Bayesian Palaeoclimate Reconstruction," *Journal of the Royal Statistical Society, Series A*, 169, 395–438.

Hooten, M., Larsen, D., and Wikle, C. (2003), "Predicting the Spatial Distribution of Ground Flora on Large Domains Using a Hierarchical Bayesian Model," *Landscape Ecology*, 18, 487–502.

Jackson, S. (1990), "Pollen Source Area and Representation in Small Lakes of the Northeastern United States," *Review of Palaeobotany and Palynology*, 63, 53–76.

Jackson, S. and Lyford, M. (1999), "Pollen Dispersal Models in Quaternary Plant Ecology: Assumptions, Parameters, and Prescriptions," *The Botanical Review*, 65, 39–75.

Jackson, S. and Williams, J. (2004), "Modern Analogs in Quaternary Paleoecology: Here Today, Gone Yesterday, Gone Tomorrow?" *Annual Review of Earth and Planetary Sciences*, 32, 495–537.

Knorr-Held, L. and Rue, H. (2002), "On Block Updating in Markov Random Field Models for Disease Mapping," *Scandinavian Journal of Statistics*, 29(4), 597–614.

Laird, N. M. and Louis, T. A. (1989), "Empirical Bayes Ranking Methods," *Journal of Educational Statistics*, 14, 29–46.

Magri, D., Vendramin, G., Comps, B., and et al. (2006), "A New Scenario for the Quaternary History of European Beech Populations: Palaeobotanical Evidence and Genetic Consequences," *New Phytologist*, 171, 199–221.

McLachlan, J., Clark, J., and Manos, P. (2005), "Molecular Indicators of Tree Migration Capacity Under Rapid Climate Change," *Ecology*, 86, 2088–2098.

Mugglin, A. S., Carlin, B. P., and Gelfand, A. E. (2000), "Fully Model-based Approaches for Spatially Misaligned Data," *Journal of the American Statistical Association*, 95(451), 877–887.

Neal, R. (1993), "Probabilistic Inference Using Markov Chain Monte Carlo Methods," Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.

Nielsen, A. and Sugita, S. (2005), "Estimating Relevant Source Area of Pollen for Small Danish Lakes Around AD 1800," *The Holocene*, 15, 1006–1020.

Oswald, W., Faison, E., Foster, D., Doughty, E., Hall, B., and Hansen, B. (2007), "Post-Glacial Changes in Spatial Patterns of Vegetation Across Southern New England," *Journal of Biogeography*, 34, 900–913.

Pacala, S., Canham, C., Saponara, J., Silander, J., Kobe, R., and Ribbens, E. (1996), "Forest Models Defined by Field Measurements: Estimation, Error Analysis and Dynamics," *Ecological Monographs*, 66, 1–43.

Paciorek, C. (2007), "Bayesian Smoothing with Gaussian Processes Using Fourier Basis Functions in the SpectralGP Package," *Journal of Statistical Software*, 19, 2.

Pawlowsky, V. and Burger, H. (1992), "Spatial Structure Analysis of Regionalized Compositions," *Mathematical Geology*, 24, 675–691.

Pawlowsky-Glahn, V. and Olea, R. (2004), *Geostatistical Analysis of Compositional Data*, Oxford: Oxford University Press.

Prentice, I. (1985), "Pollen Representation, Source Area, and Basin Size: Toward a Unified Theory of Pollen Analysis," *Quaternary Research*, 23, 76–86.

Prentice, I., Berglund, B., and Olsson, T. (1987), "Quantitative Forest Composition Sensing Characteristics of Pollen Samples from Swedish Lakes," *Boreas*, 16, 43–54.

Rougier, J. (2008), "Comment on Article by Sansó et al.," *Bayesian Analysis*, 3, 45–56.

Royle, J. A. and Wikle, C. K. (2005), "Efficient Statistical Mapping of Avian Count Data," *Environmental and Ecological Statistics*, 12(2), 225–243.

Rue, H. and Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, Boca Raton: Chapman & Hall.

Ruppert, D., Wand, M., and Carroll, R. (2003), *Semiparametric Regression*, Cambridge, U.K.: Cambridge University Press.

Soepboer, W., Sugita, S., Lotter, A., Van Leeuwen, J., and Van der Knaap, W. (2007), "Pollen Productivity Estimates for Quantitative Reconstruction of Vegetation Cover on the Swiss Plateau," *The Holocene*, 17, 65–77.

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2003), "WinBUGS User Manual, Version 1.4," Technical report, MRC Biostatistics Unit.

Sugita, S. (1993), "A Model of Pollen Source Area for an Entire Lake Surface," *Quaternary Research*, 39, 239–244.

——— (1994), "Pollen Representation of Vegetation in Quaternary Sediments: Theory and Method in Patchy Vegetation," *Journal of Ecology*, 82, 881–897.

——— (2007a), "Theory of Quantitative Reconstruction of Vegetation I: Pollen from Large Sites REVEALS Regional Vegetation Composition," *The Holocene*, 17, 229–241.

——— (2007b), "Theory of Quantitative Reconstruction of Vegetation II: All You Need Is LOVE," *The Holocene*, 17, 243–257.

Sugita, S., Gaillard, M.-J., and Bronstrom, A. (1998), "Landscape Openness and Pollen Records: A Simulation Approach," *The Holocene*, 9, 409–421.

Tauber, H. (1965), "Differential Pollen Dispersion and the Interpretation of Pollen Diagrams," *Danm. Geol. Unders. II Raekke Ser.*, 89, 1–69.

Tjelmeland, H. and Lund, K. (2003), "Bayesian Modelling of Spatial Compositional Data," *Journal of Applied Statistics*, 30, 87–100.

von Post, L. (1917), "Om Skogstradpollen i Sydsvenska Tormfmosselagerfolker," *Geologiska Foreningens i Stockholm Forhandlingar*, 38, 384–390.

Webb, T. (1974), "Corresponding Distributions of Modern Pollen and Vegetation in Lower Michigan," *Ecology*, 55, 17–18.

Wikle, C. K., Milliff, R. F., Nychka, D., and Berliner, L. M. (2001), "Spatiotemporal Hierarchical Bayesian Modeling Tropical Ocean Surface Winds," *Journal of the American Statistical Association*, 96(454), 382–397.

Williams, J., Shuman, B., and Webb, T. (2001), "Dissimilarity Analyses of Late-Quaternary Vegetation and Climate in Eastern North America," *Ecology*, 82, 3346–3362.

Yucel, R. and Zaslavsky, A. (2005), "Imputation of Binary Treatment Variables with Measurement Error in Administrative Data," *Journal of the American Statistical Association*, 100, 1123–1132.