Post-glacial vegetation dynamics: Bayesian inference for spatio-temporal trends in forest composition using the fossil pollen record

Chris Paciorek
Department of Biostatistics
Harvard School of Public Health

Jason McLachlan
Center for Population Biology
UC-Davis

October 2006

www.biostat.harvard.edu/~paciorek

LyX - FoilTEX - pdfLATEX

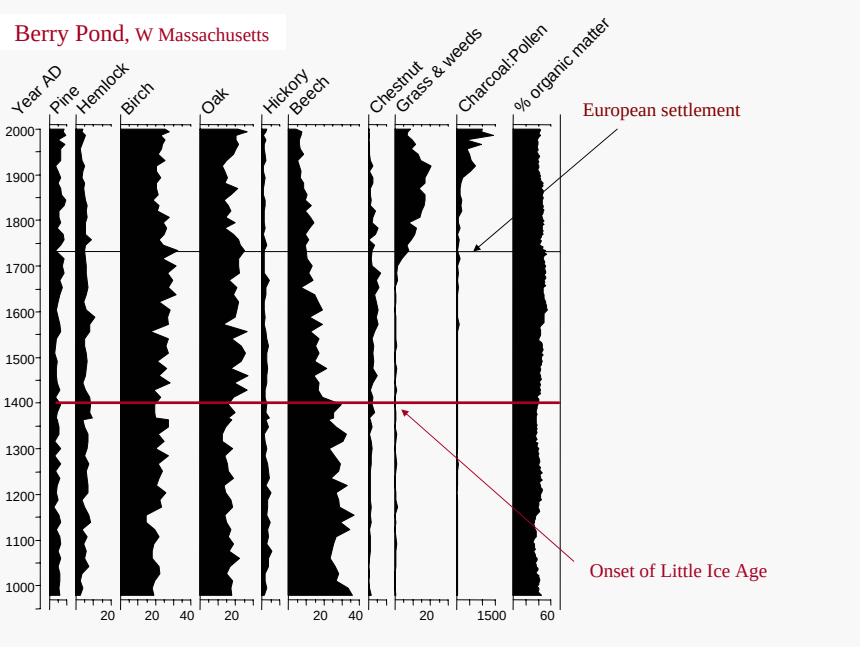
cores are taken from the sediment of ponds

a sediment core

courtesy of David Foster, Harvard Forest

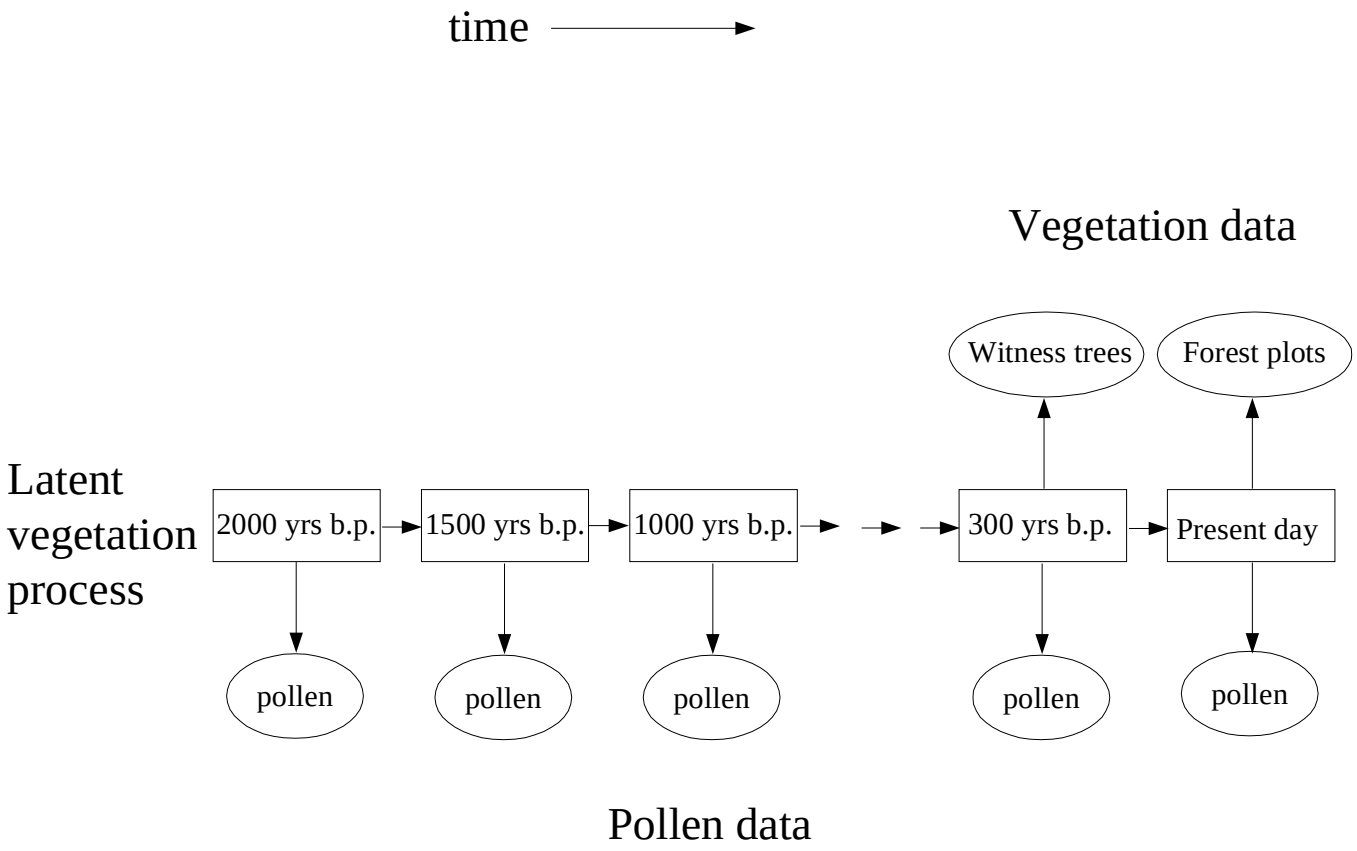
Scientific Setting

- Tree pollen accumulates in lake sediments over time; vertical cores sample the sediment
- Pollen identified to genera helps estimate tree composition over time; radiocarbon dating estimates times
- Tree composition is useful for understanding vegetation dynamics, tree migration, and climate
  - Particular interest in post-glacial vegetation structure and migration into ice-vacated areas
- The pollen record is biased and noisy
- Current analysis methods: time series plots of individual pond records



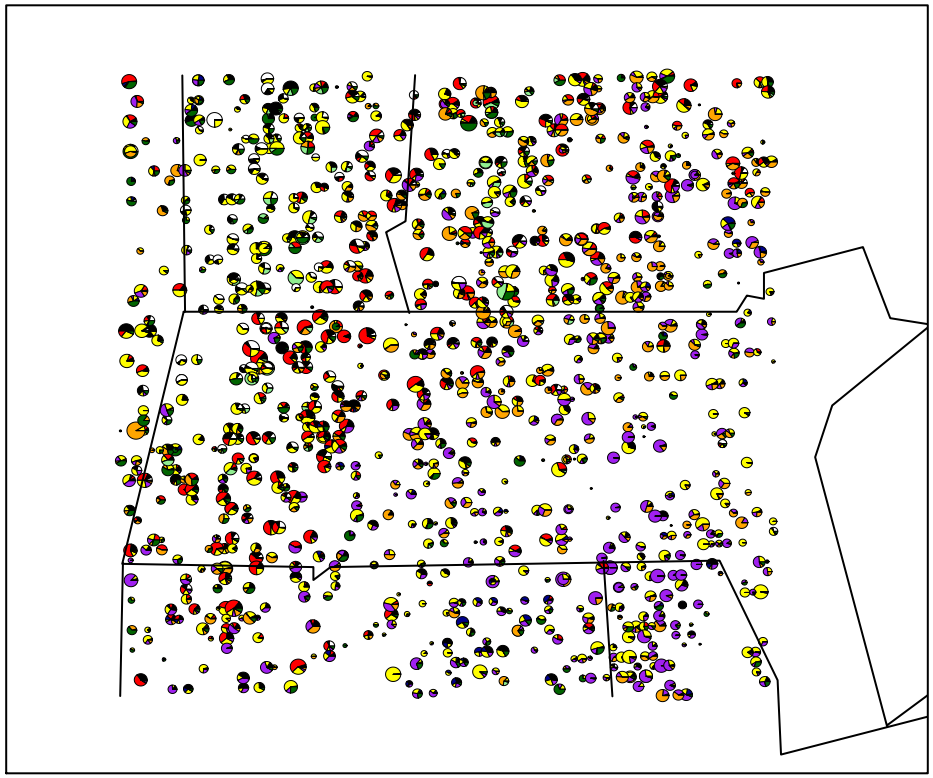
courtesy of David Foster, Harvard Forest

Basic problem structure



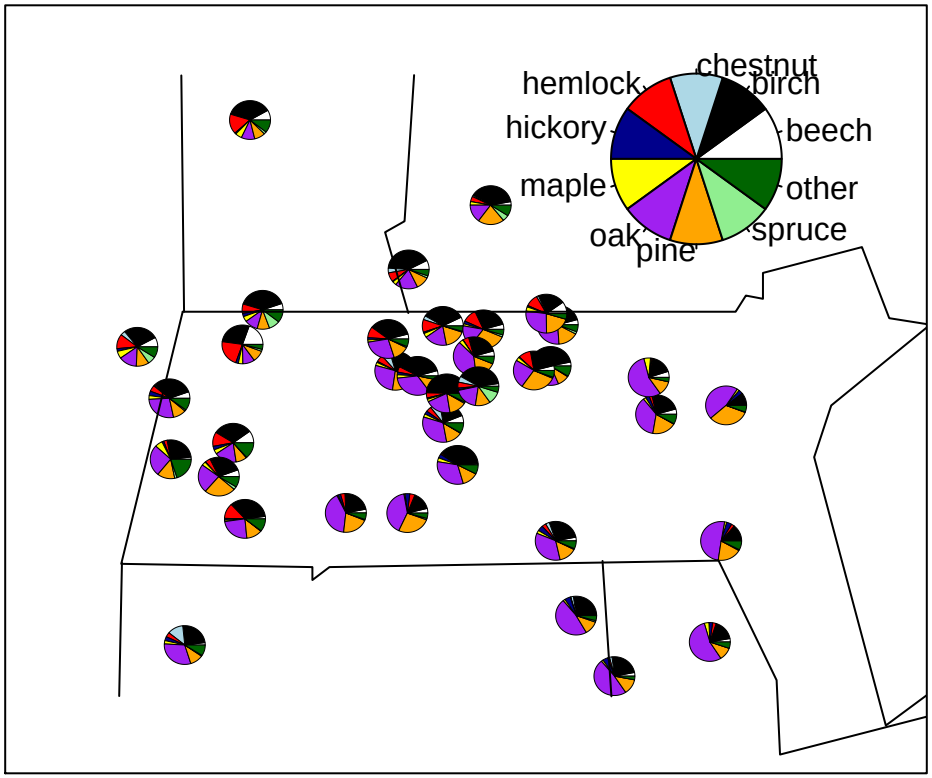
Central New England modern data

USFS vegetation plot composition



1161 plots, 1-115 trees per plot

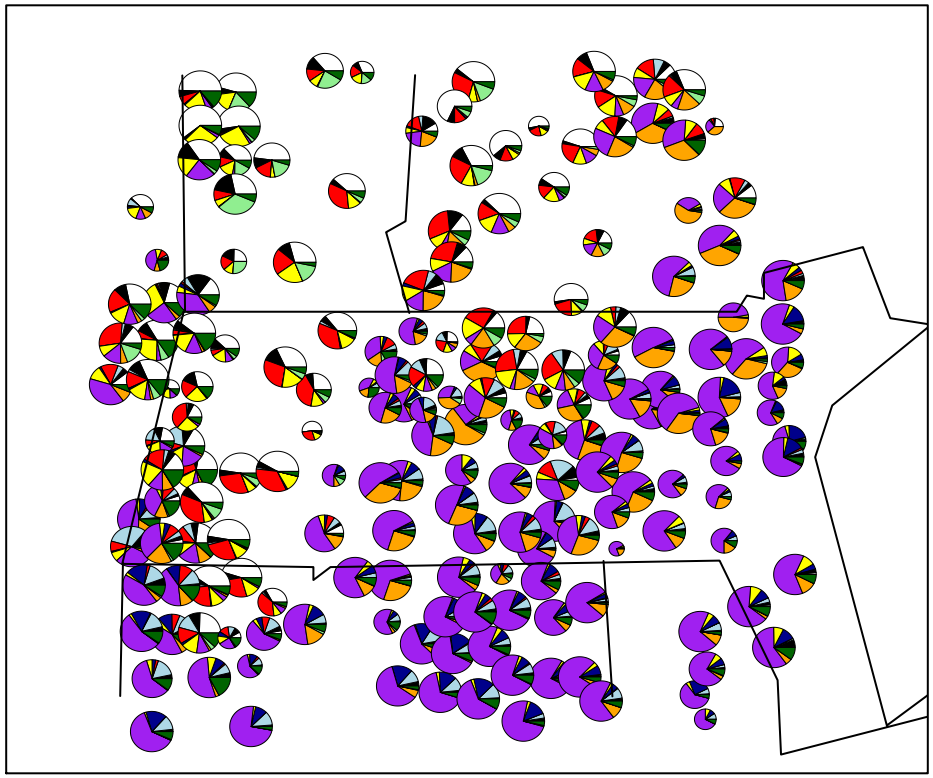
Pollen composition



38 ponds, 113-582 grains per pond

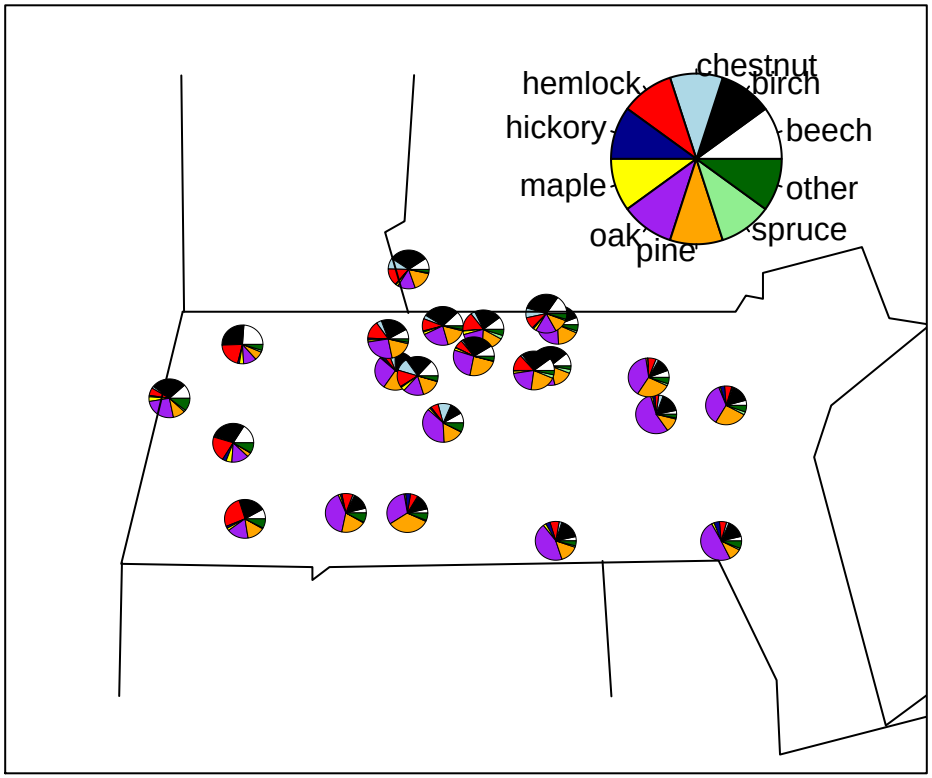
Central New England colonial data

Township witness tree composition



183 towns, 26-3149 trees per town

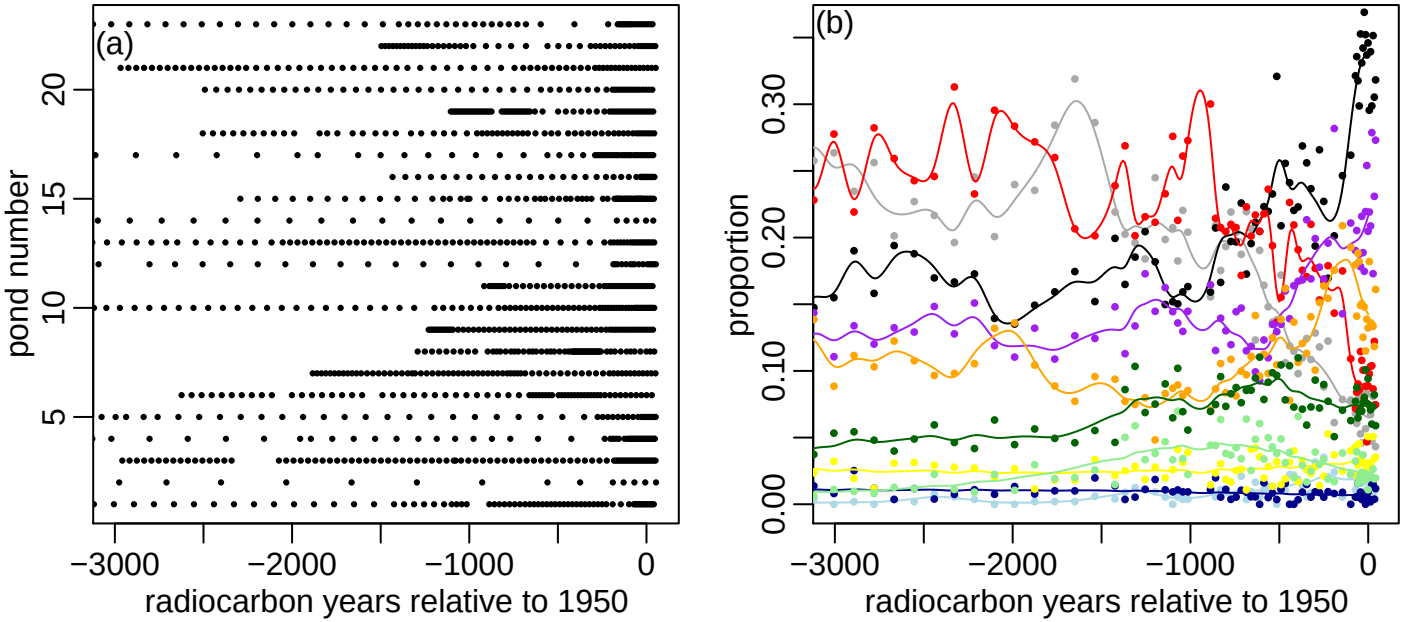
Pollen composition



23 ponds, 439-621 grains per pond

R help: “Pie charts are a very bad way of displaying information.”

Pollen data availability and smoothing



time series of pollen samples

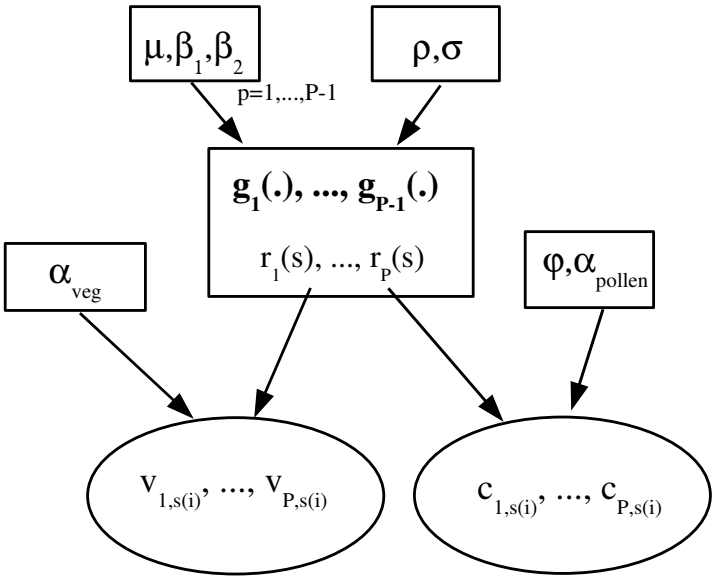
smoothed proportions in one pond

Goals

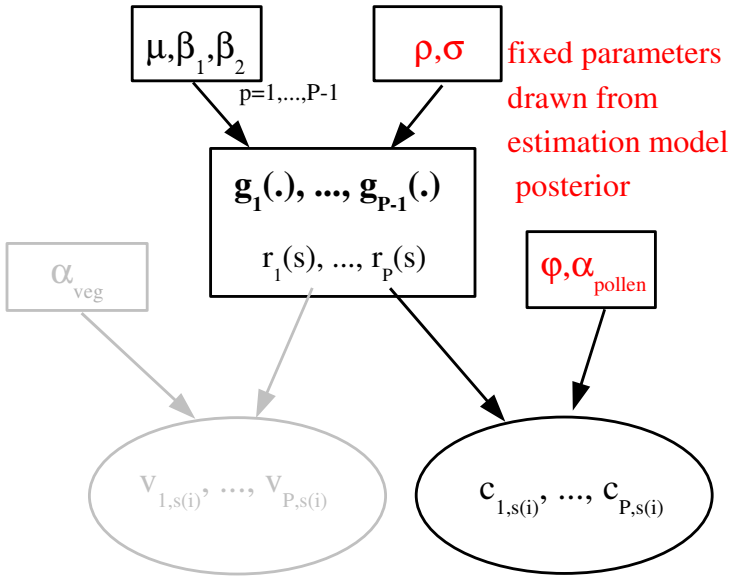
- Understand the relationship between pollen and vegetation based on modern and colonial data.
  - At what resolution are ponds a good proxy for vegetation?
  - How noisy is the relationship between the pollen record and vegetation?
- Estimate and compare vegetation in the colonial and modern eras.
  - Assess reliability of witness tree records.
- **Predict spatial patterns in tree abundances over the past 3000 years.**
  - Provide uncertainty estimates to allow inference about spatio-temporal patterns.
  - Assess changes in taxa relationships with covariates
  - Use the predictions to understand vegetation dynamics: changing abundance and ranges of tree taxa over time.
- Use the model as a research framework
  - Assess ecological hypotheses about population growth
  - Integrate genetic data to understand migration patterns

Basic models

Estimation model (veg and pollen)



Prediction model (pollen only)



Model (1): Latent spatial processes

For fixed time, $P = 10$ latent Gaussian spatial processes:

$$g_p(\cdot) \sim \text{GP}(\mu_p \mathbf{1} + \beta_1 \text{elevation}(\cdot) + \beta_2 \text{latitude}(\cdot), \sigma^2 R(\rho, \nu))$$

Proportion of taxa p at location s, $r_p(s)$, via additive log-ratio transformation (Aitchison 1985):

$$r_p(s) = \frac{\exp(g_p(s))}{\sum_{k=1}^{10} \exp(g_k(s))}; \qquad \sum_p r_p(s) = 1$$

Processes efficiently represented on a 16 by 16 grid:

$$\boldsymbol{g}_p \quad = \quad \mu_p \mathbf{1} + \beta_1 \text{elevation} + \beta_2 \text{latitude} + \sigma \Psi \boldsymbol{u}_p; \quad \boldsymbol{u}_p \sim \text{N}(\mathbf{0}, V(\rho, \nu))$$

Ψ is the Fourier basis matrix

$\text{V}(\rho, \nu)$ is a diagonal variance matrix based on the spectral density of the Matern (ρ, ν) correlation function

One ρ and one σ^2 common to all taxa seem sufficient when covariates included

Model (2): Vegetation likelihood

- Modern plot data (tree counts), $i = 1, \ldots, 1161$:
  - $\boldsymbol{v}_i \sim \text{Dir-multi}(n_i^{(v)}, \alpha_{\text{veg}} \boldsymbol{r}(\boldsymbol{s}(\boldsymbol{i})))$
  - α_{veg} is extra-multinomial heterogeneity, giving a Dirichlet mixture of multinomials
    $\boldsymbol{r}(\boldsymbol{s}(\boldsymbol{i}))$ is composition vector $(r_1(s(i)), \ldots, r_{10}(s(i)))$
- Colonial surveys (witness tree counts in townships), $i = 1, \ldots, 183$:
  - $\boldsymbol{v}_i \sim \text{Dir-multi}(n_i^{(v)}, \alpha_{\text{veg}} \overline{\boldsymbol{r}(\boldsymbol{s}(\boldsymbol{i}))})$
  - $\overline{\boldsymbol{r}(\boldsymbol{s}(\boldsymbol{i}))}$ is the weighted composition based on town-gridbox overlap

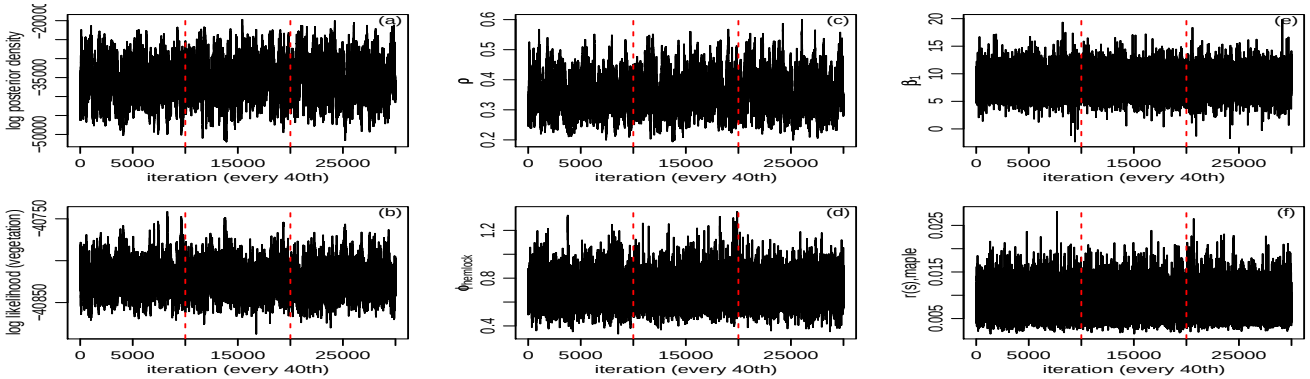
Model (3): Pollen data likelihood

- estimation model:
  - pollen grain counts from 23 ponds for colonial and 38 ponds for modern
  - $\boldsymbol{c}_i \sim \text{Dir-Multi}(n_i^{(c)}, \boldsymbol{\phi} \cdot \widetilde{\boldsymbol{r}(\boldsymbol{s}(\boldsymbol{i}))})$
  - ϕ are taxa-specific pollen-to-vegetation scaling factors
    - account for pollen production and dispersal variability between taxa
  - $\widetilde{\boldsymbol{r}(\boldsymbol{s}(\boldsymbol{i})}$ is weighted average of grid cell vegetation and distance-weighted vegetation in other cells

- prediction model
  - pollen grain counts from 23, $i = 1, \ldots, 23$, ponds at haphazard times
  - smooth counts over time using gam() at each pond to get predicted composition at fixed times, $\hat{\boldsymbol{c}}_i$
  - $\hat{\boldsymbol{c}}_i \sim \text{Dir}(n_i^{(c)}, \boldsymbol{\phi} \cdot \widetilde{\boldsymbol{r}(\boldsymbol{s}(\boldsymbol{i}))})$

MCMC performance

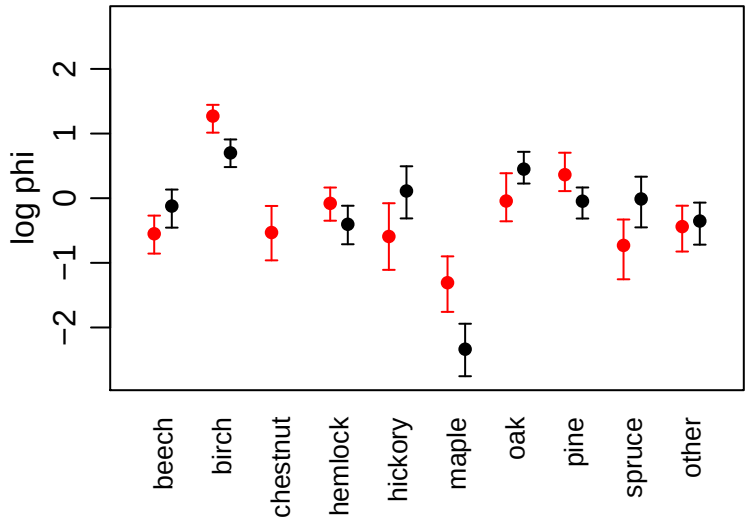
- MCMC mixing is rather slow but convergence seems reasonable; modern estimation run shown here for 3 chains, 400k iterations each:



- Are there better sampling schemes than Metropolis-Hastings for the spatial processes?
  - Given the relative nature of the processes, good proposals are hard
- Current implementation of Fourier approach seems to outperform thin-plate spline
- Modification allowing Gibbs sampling of coefficients via introduction of additional variance component provides little improvement (see model in Paciorek, in prep; Wikle 2002)
- Prediction runs propagate hyperparameter $\{\phi, \alpha_{\text{pollen}}, \rho, \sigma\}$ uncertainty using multiple runs

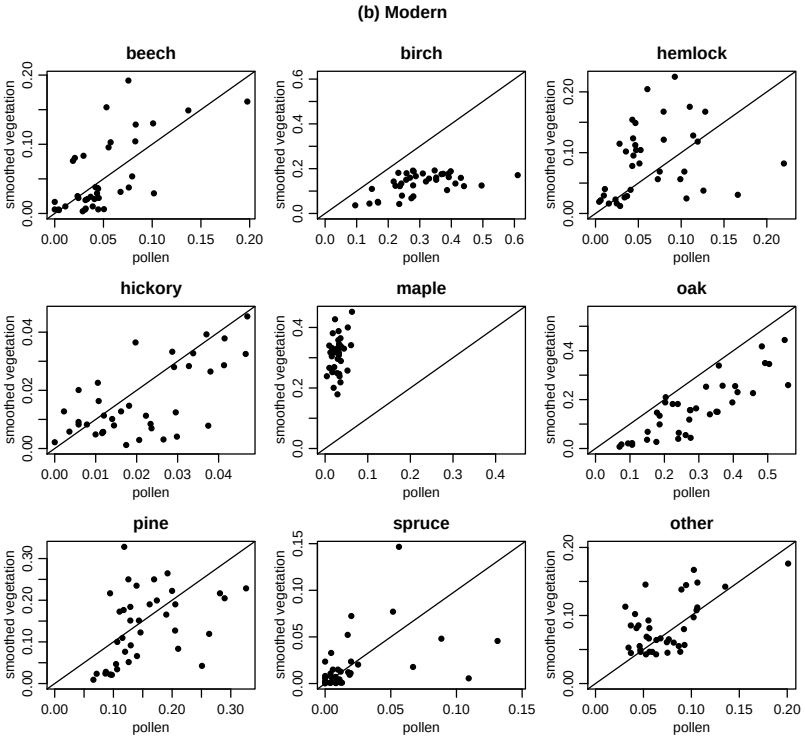
Estimation model inference

- colonial and modern parameter estimates reasonably comparable
- 25-30% of pollen attributed to grid cell vegetation (rather low, but scientifically provocative)
- Pollen scaling parameters (ϕ) relatively uncertain:

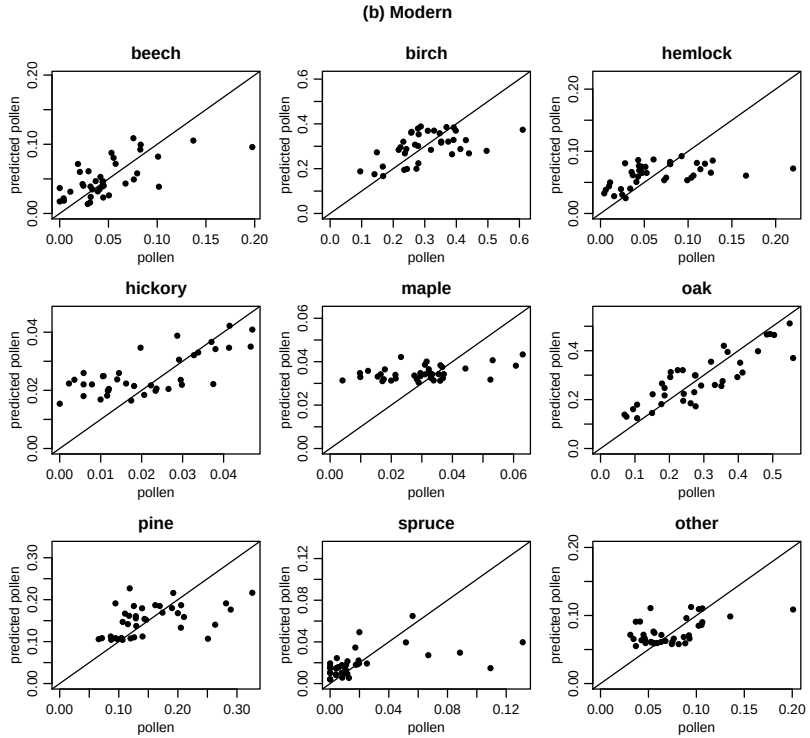


Pollen as a vegetation proxy

Unadjusted pollen-vegetation relationship



Adjusted pollen-vegetation relationship

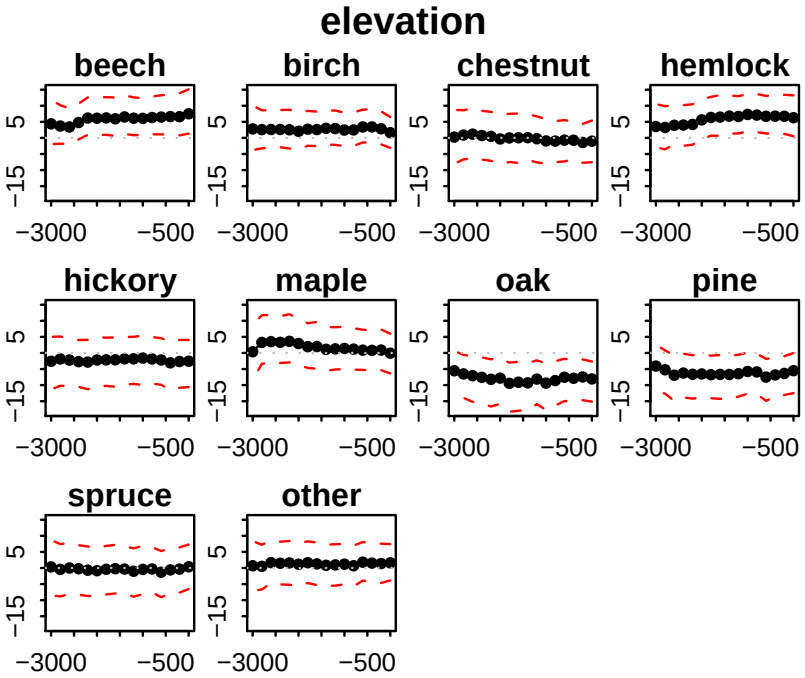


ϕ parameters scale pollen to vegetation. After adjustment by ϕ and for long-distance transport, most ponds show reasonable, albeit noisy, relationships between pollen and vegetation estimated in the grid box.

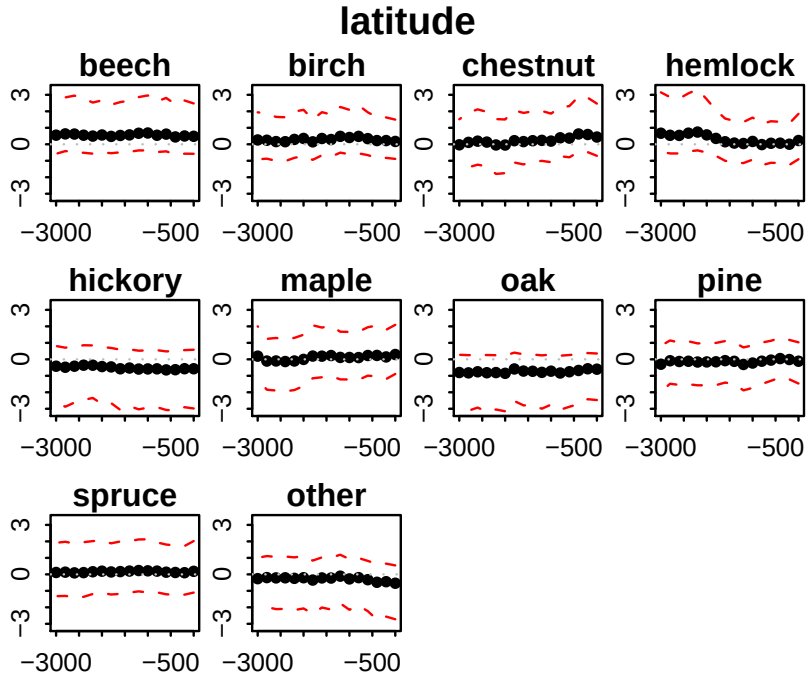
Cross-validation check: Colonial predictions

Covariate effects through time

Elevation

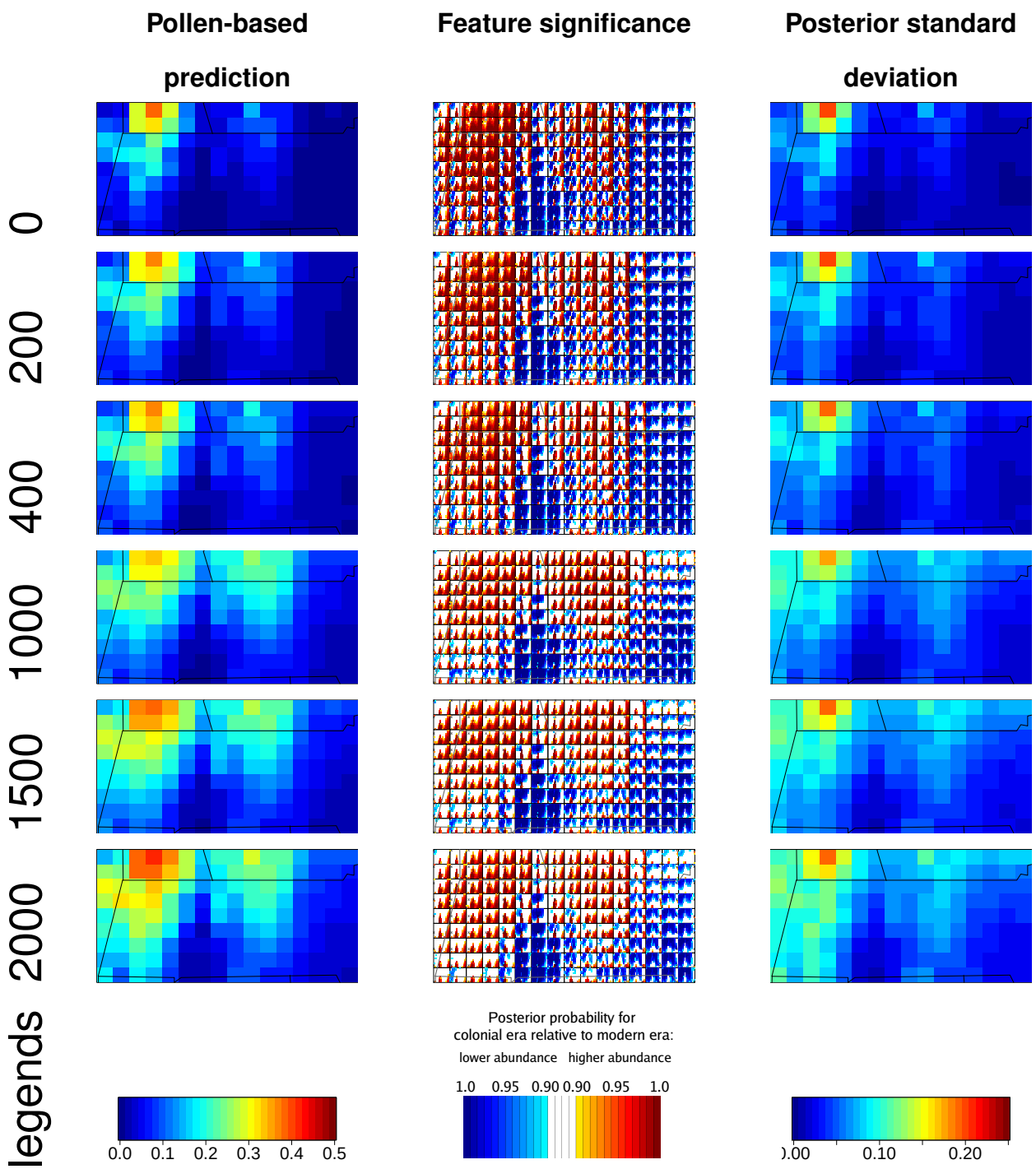


Latitude



Covariate effects appear consistent through time (albeit with high uncertainty).

Prediction in time for beech



For beech, vegetation prediction information for six distinct times between 0 years before 1950 (top row) and 2000 years before 1950 (bottom row). Plots are of relative vegetation abundance predicted based on smoothed pollen for the time point and modern parameter estimates (first column), feature significance for that prediction run (second column) and posterior prediction standard deviations (third column). Patterns are fairly similar over time. Can we detect changes over time?

Prediction in time for oak

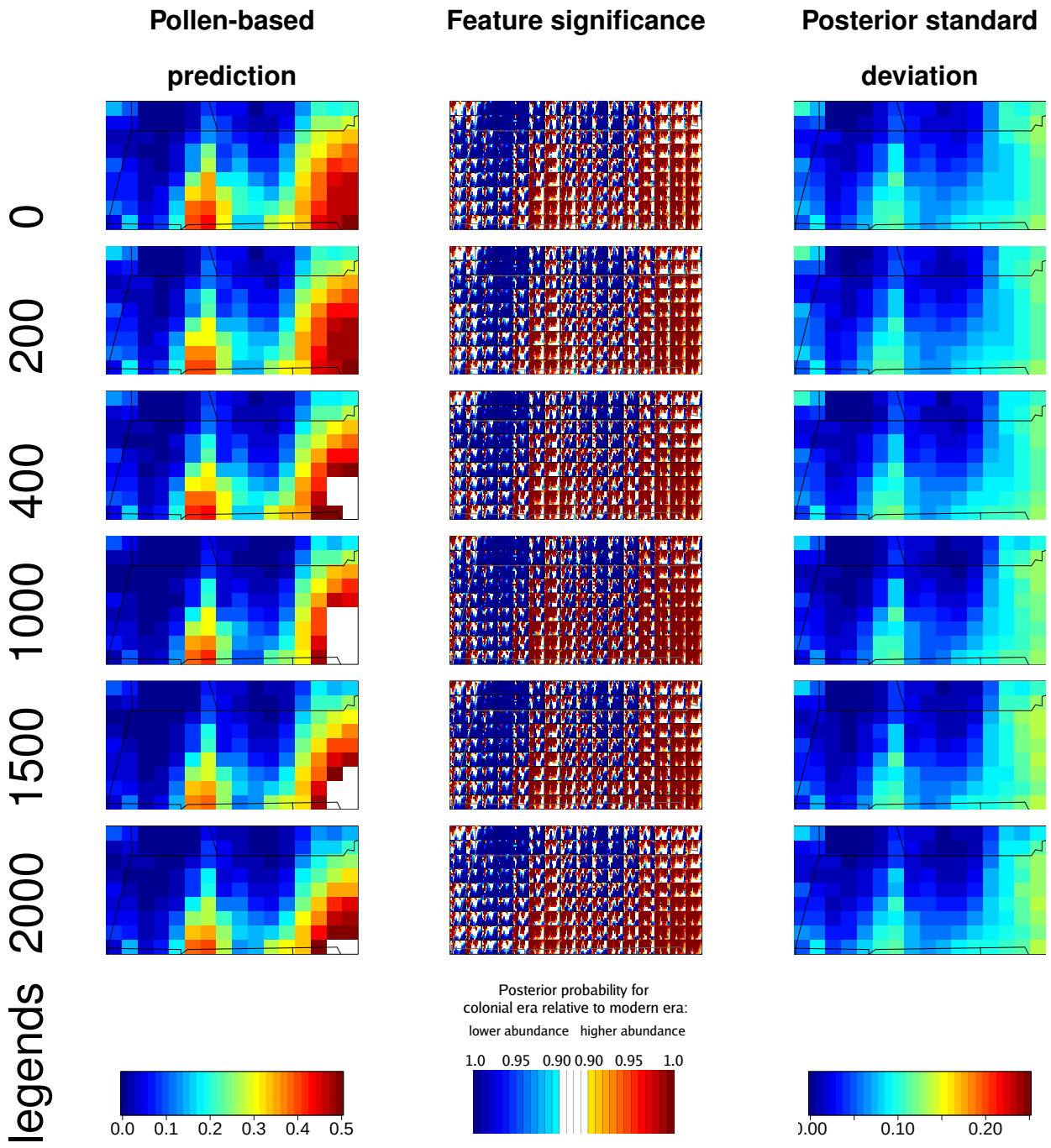
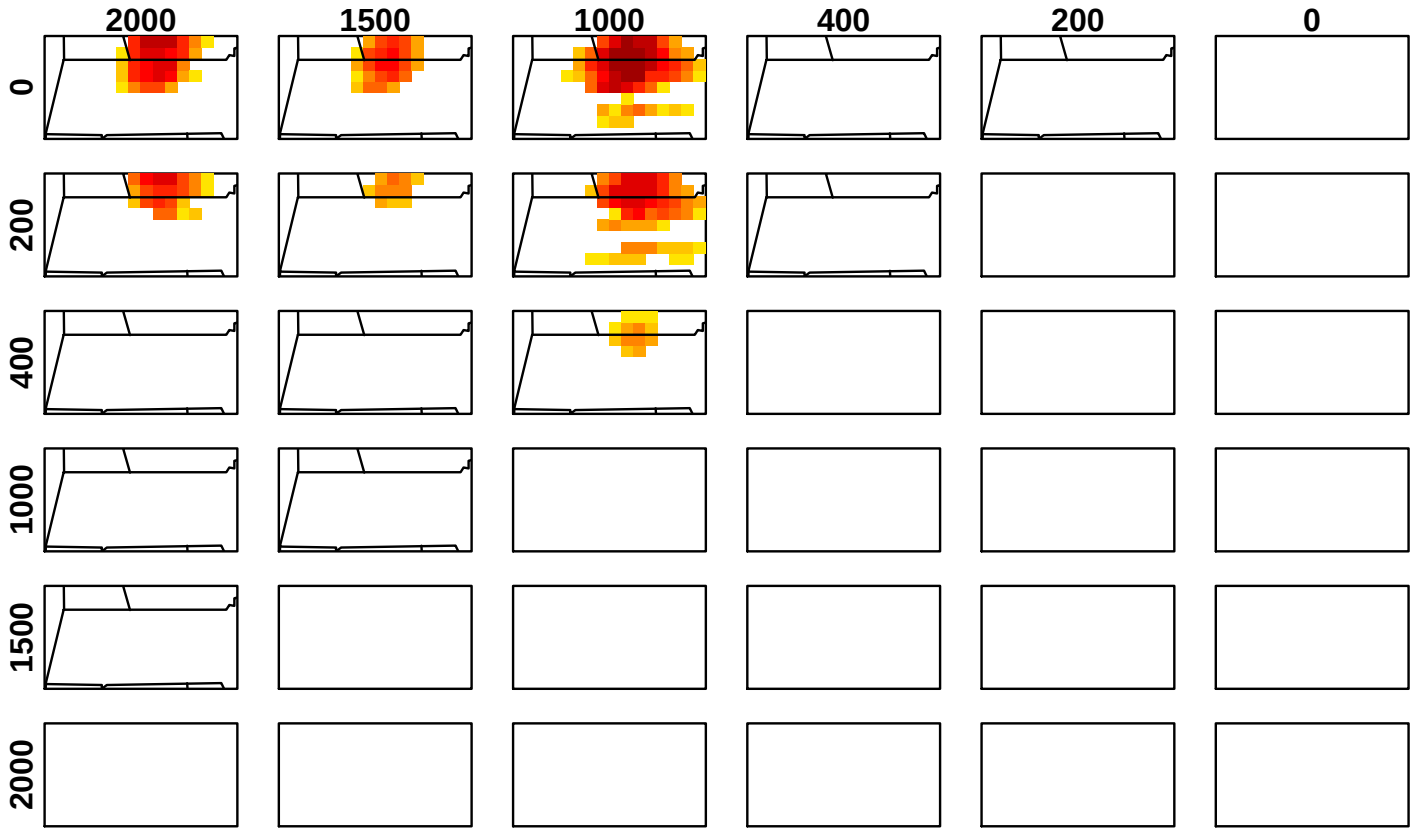


Figure 1: For oak, vegetation prediction information for six distinct times between 0 years before 1950 (top row) and 2000 years before 1950 (bottom row). Plots are vegetation predicted based on smoothed pollen for the time point and modern parameter estimates (first column), feature significance for that prediction run (second column) and posterior prediction standard deviations (third column). Patterns are fairly similar over time. Can we detect changes over time?

Significance of changes over time for oak



Changes over time grid cell by grid cell compared between pairs of time points based on years before present. Here red indicates that the later time period has more of that taxa in the grid cell, with posterior probability based on shading (with a threshold of 90%).

Future Work

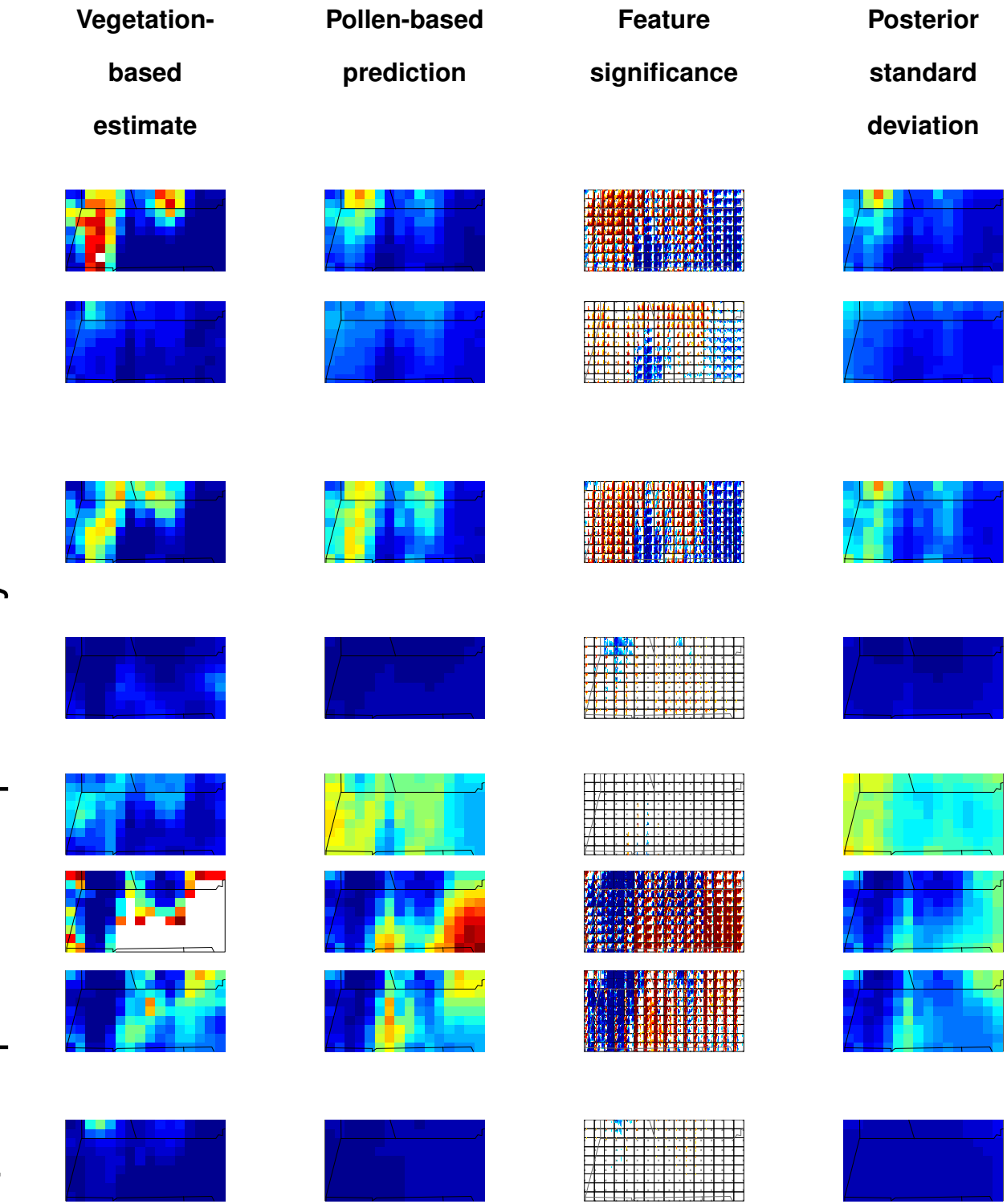
- Possible sampling of additional ponds to distinguish local pollen from long-distance transport
- Consideration of a full space-time model to lessen concerns about uncertainty bounds on comparisons across time
- Incorporation of ecological models for changes over time to understand population dynamics
- Analysis of pollen data from Michigan
- Expansion to the northeastern United States + southeastern Canada post-glaciation
  - better resolve spatial heterogeneity
  - assess tree migration
- Use of vegetation composition estimates as input/constraints to a model of genetic change over time

Fourier representation

Computationally efficient basis function construction
(Wikle 2002, Royle and Wikle 2005, Paciorek and Ryan 2005)

- $\boldsymbol{g}^{\#} = \Psi \boldsymbol{u}$
  - Piecewise constant gridded surface on k by k grid
  - additional observations are computationally 'free' for fixed grid
- Ψ is the Fourier (spectral) basis and $\Psi \boldsymbol{u}$ is the inverse FFT
  - $O\left((k^2)\log(k^2)\right)$ computations, $k = 32$
  - fast calculation of surface given coefficients
- $\Psi \boldsymbol{u}$ is approximately a Gaussian process (GP) when...
  - $\boldsymbol{u} \sim N(0, \text{diag}(\pi_\theta(\boldsymbol{\omega}; \rho, \nu)))$ for Fourier frequencies, $\boldsymbol{\omega}$
  - spectral density, $\pi_\theta(\cdot; \rho, \nu)$, of GP covariance function defines V($\boldsymbol{u}$)
- a priori independent coefficients
  - fast computation of prior density
  - improved mixing (sometimes)

tmp


Vegetation-based estimate
Pollen-based prediction
Feature significance
Posterior standard deviation
spruce pine oak maple hickory hemlock birch beech