# 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

June 2006

www.biostat.harvard.edu/~paciorek

LyX - FoilTEX - pdfLATEX

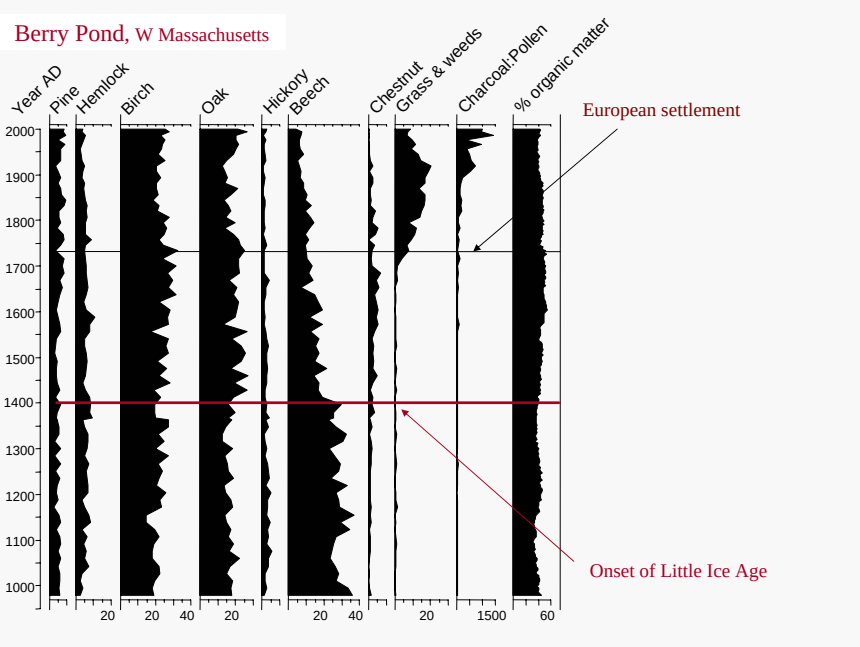
cores are taken from the sediment of ponds

a sediment core

courtesy of David Foster, Harvard Forest

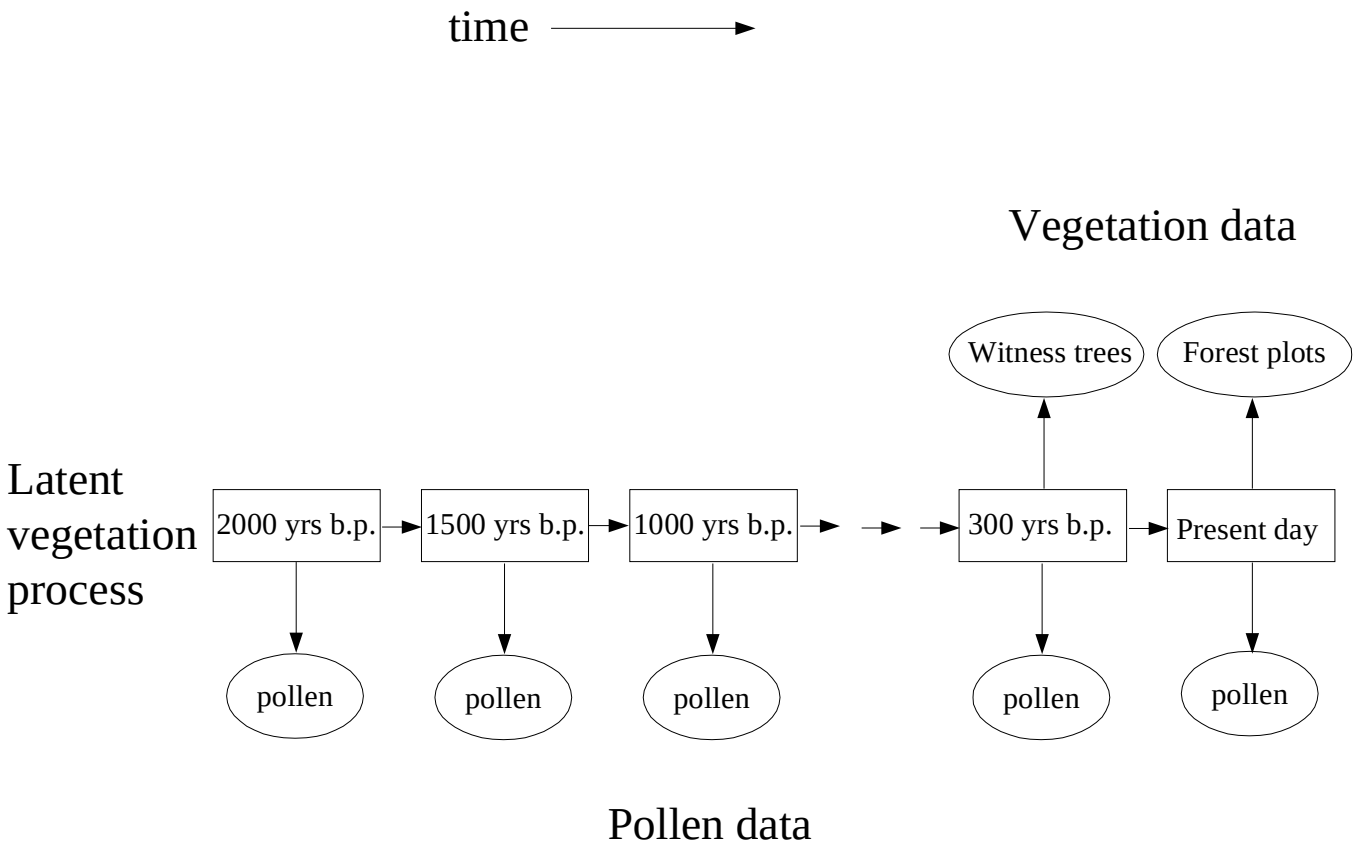
# Scientific Setting

- 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



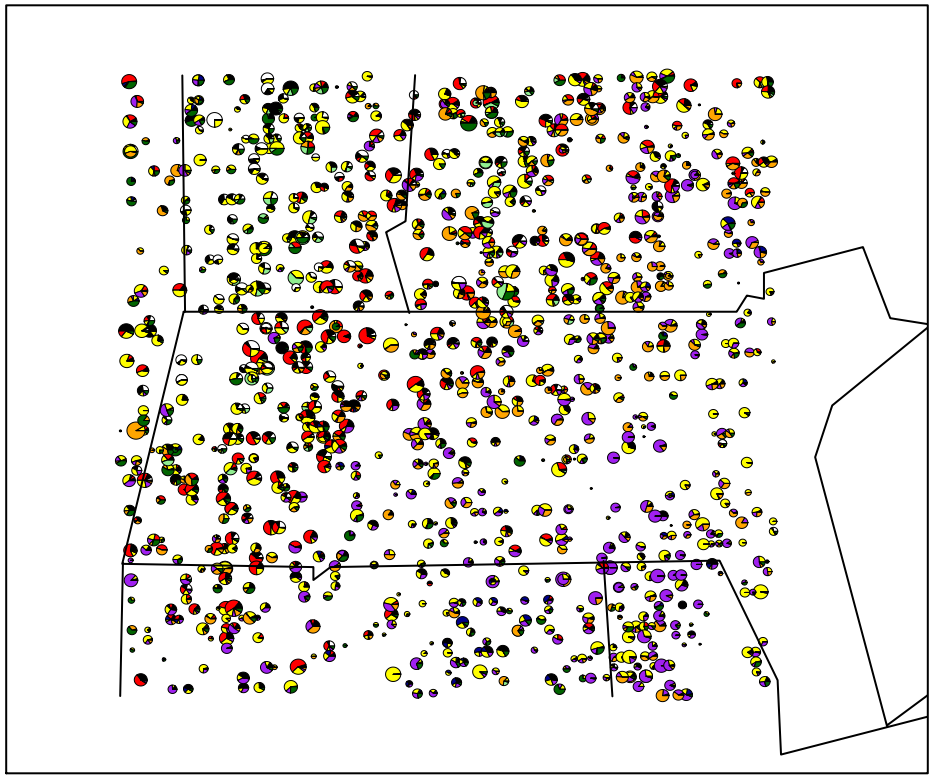

courtesy of David Foster, Harvard Forest

# Basic problem structure



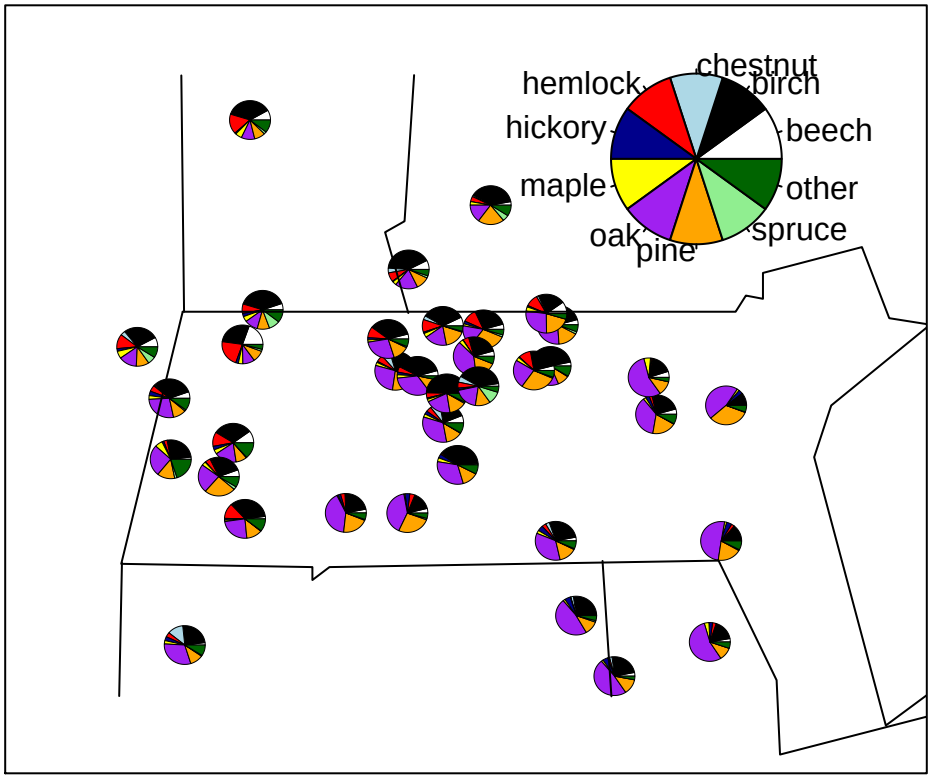
# Central New England modern data

**USFS vegetation plot composition**



1161 plots, 1-115 trees per plot

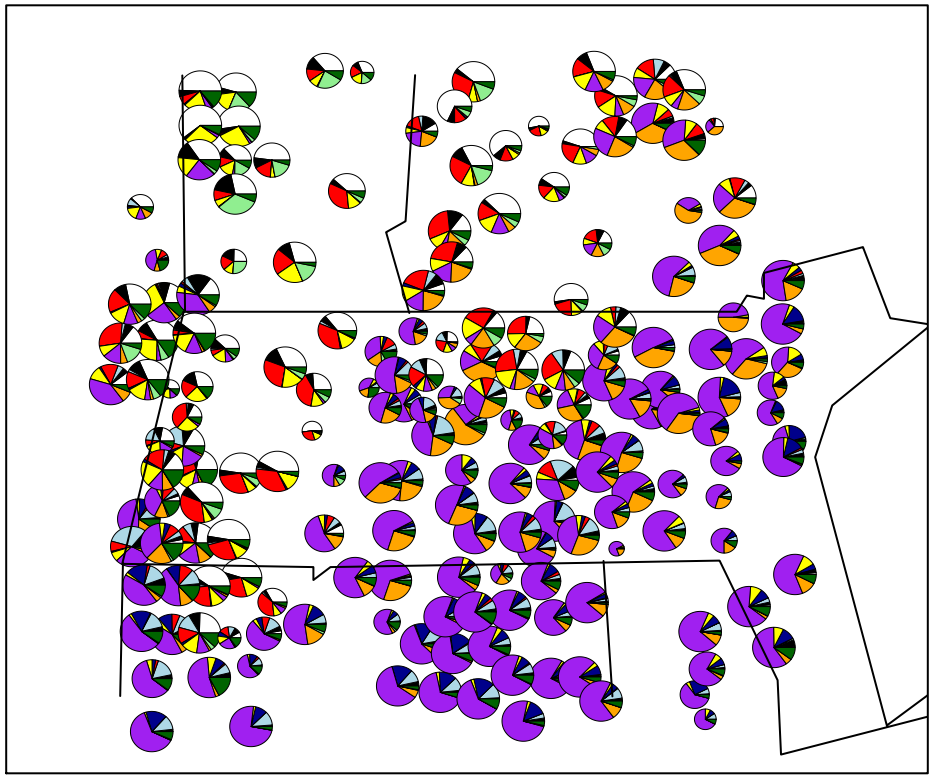
**Pollen composition**




36 ponds, 113-582 grains per pond

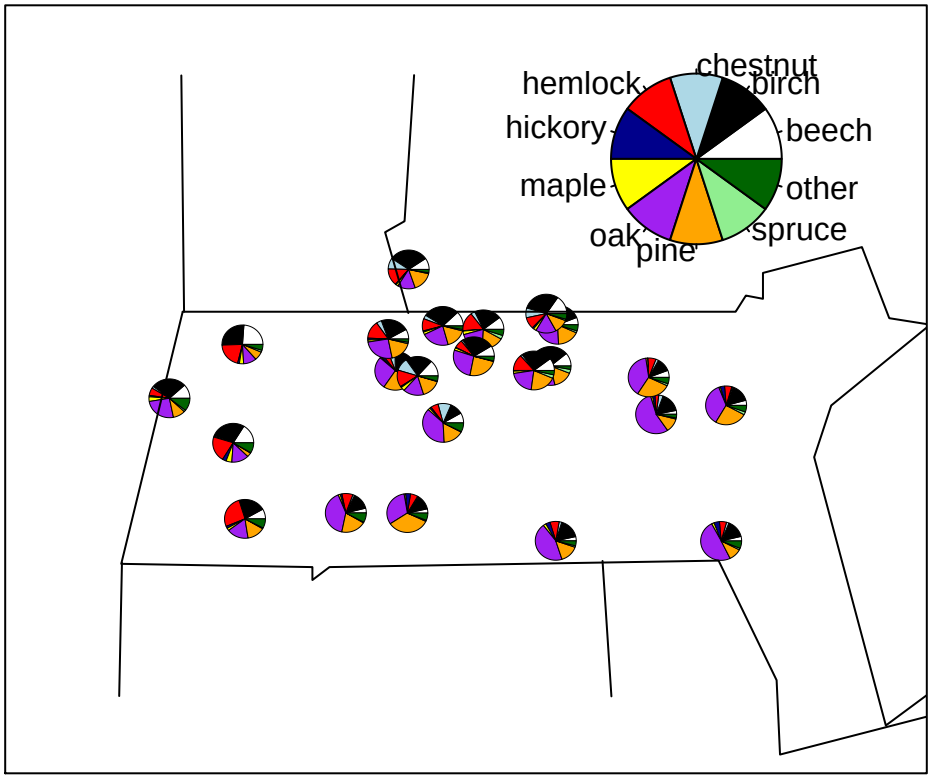
# Central New England colonial data

**Township witness tree composition**



183 towns, 26-3149 trees per town

**Pollen composition**




23 ponds, 439-621 grains per pond

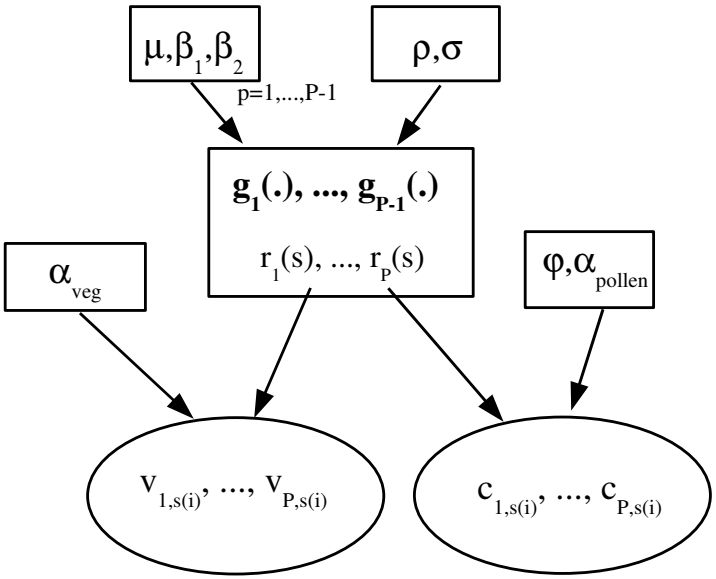
R help: “Pie charts are a very bad way of displaying information.”

# Goals

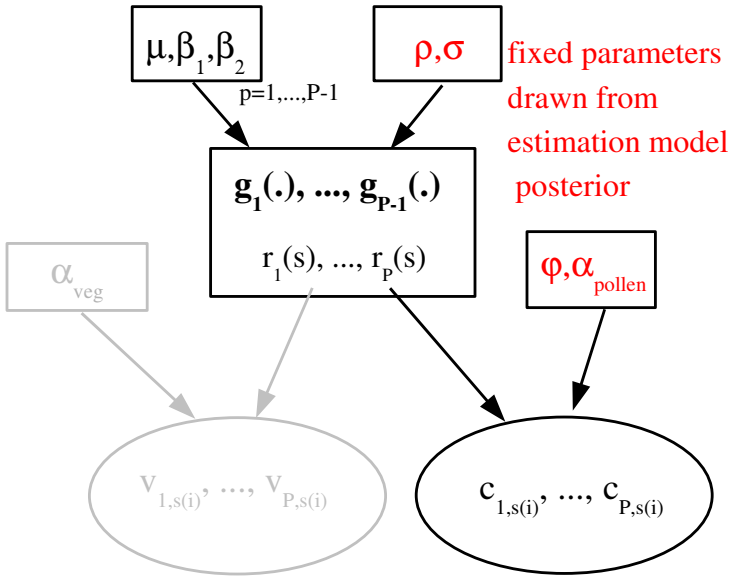
- Understand the relationship between pollen and vegetation based on modern and colonial data.
  - At what resolution are ponds a good proxy for vegetation?
  - How noisy is the relationship between the pollen record and vegetation?
- Estimate and compare vegetation in the colonial and modern eras.
  - 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, $P - 1 = 9$ 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):

$$r_p(s) = \frac{\exp(g_p(s))}{1 + \sum_{k=1}^{9} \exp(g_k(s))}; \qquad \sum_p r_p(s) = 1$$

Processes efficiently represented on a 16 by 16 grid:

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

$\Psi$ is the Fourier basis matrix

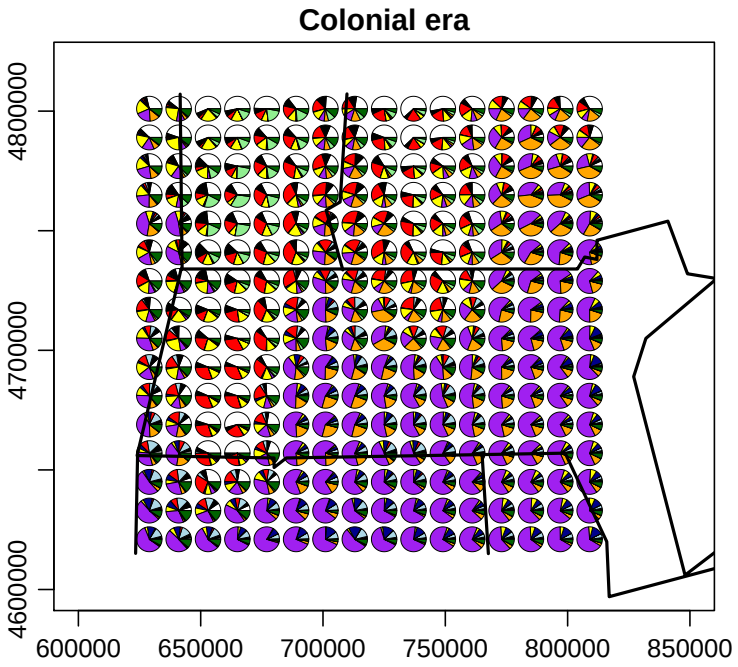
$\text{V}(\rho, \nu)$ is a diagonal variance matrix based on the spectral density of the Matern $(\rho, \nu)$ correlation function

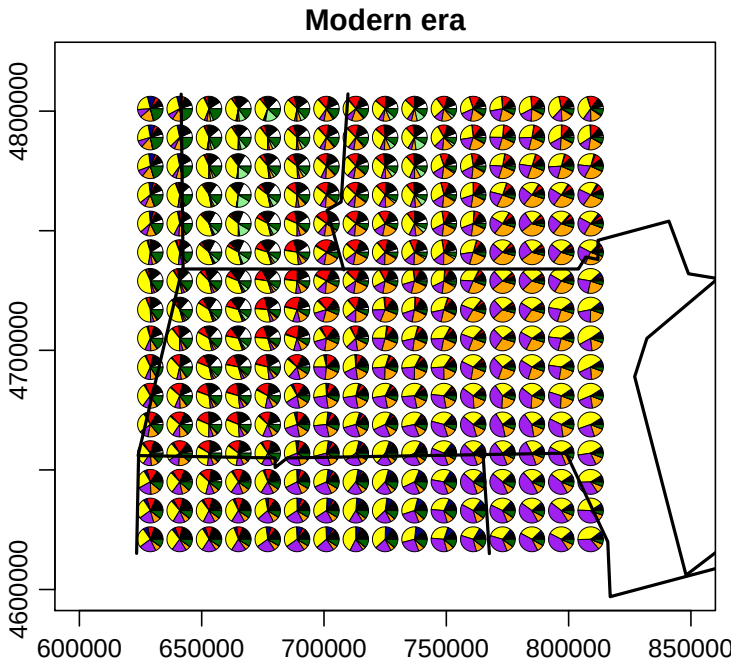
One $\rho$ and one $\sigma^2$ common to all taxa seem sufficient when covariates included

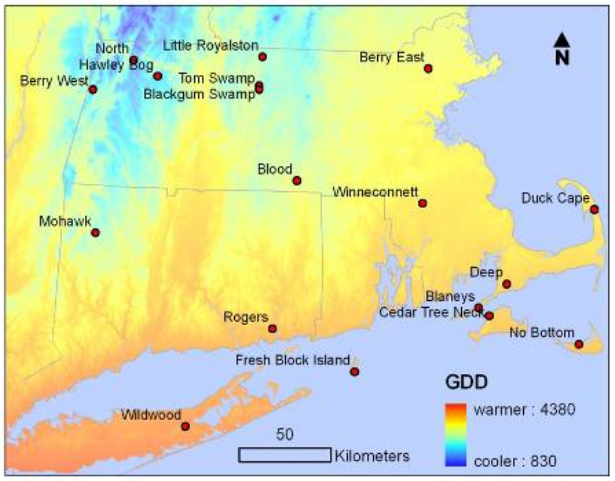
# Model (2): Likelihood terms

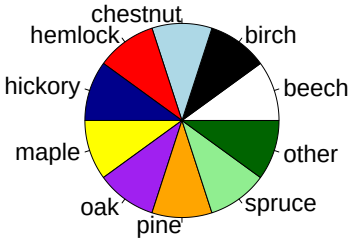
- Modern plot data (tree counts), $i = 1, \ldots, 1161$:
  - $\boldsymbol{v}_i \sim \text{Dir-multi}(n_i^{(v)}, \alpha_{\text{veg}} \boldsymbol{r}(\boldsymbol{s}(\boldsymbol{i})))$
  - $\alpha_{\text{veg}}$ is extra-multinomial heterogeneity, giving a Dirichlet mixture of multinomials
    $\boldsymbol{r}(\boldsymbol{s}(\boldsymbol{i}))$ 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
- Pollen data (pollen grain counts from 23 ponds at a fixed time), $i = 1, \ldots, 22$:
  - $\boldsymbol{c}_i \sim \text{Dir-multi}(n_i^{(c)}, \alpha_{\text{pollen}} \boldsymbol{\phi} \cdot \boldsymbol{r}(\boldsymbol{s}(\boldsymbol{i})))$
  - $\boldsymbol{\phi}$ are taxa-specific pollen to vegetation scaling factors
    - account for pollen production and dispersal variability between taxa

# Results (1): Vegetation based on vegetation data



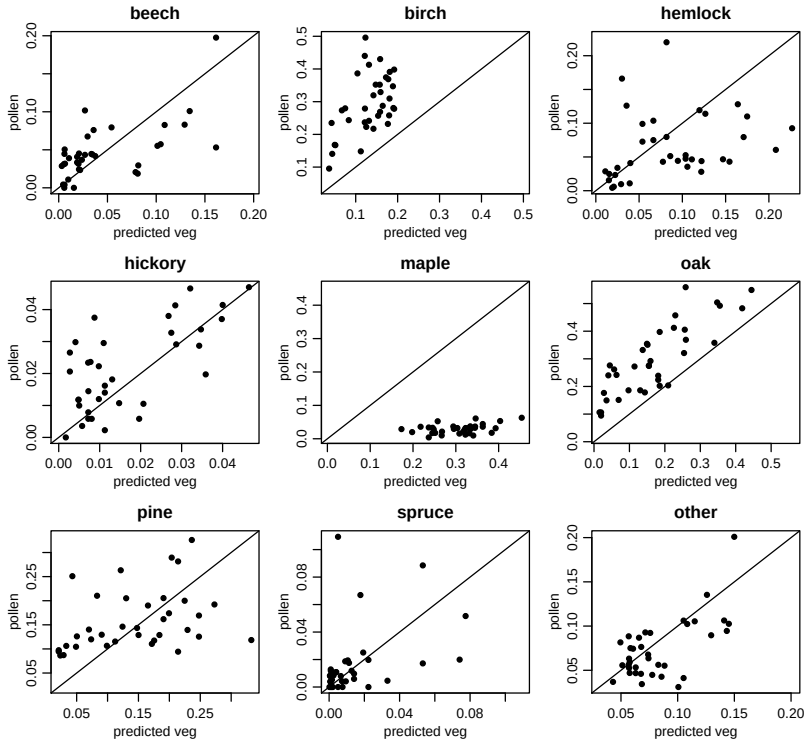



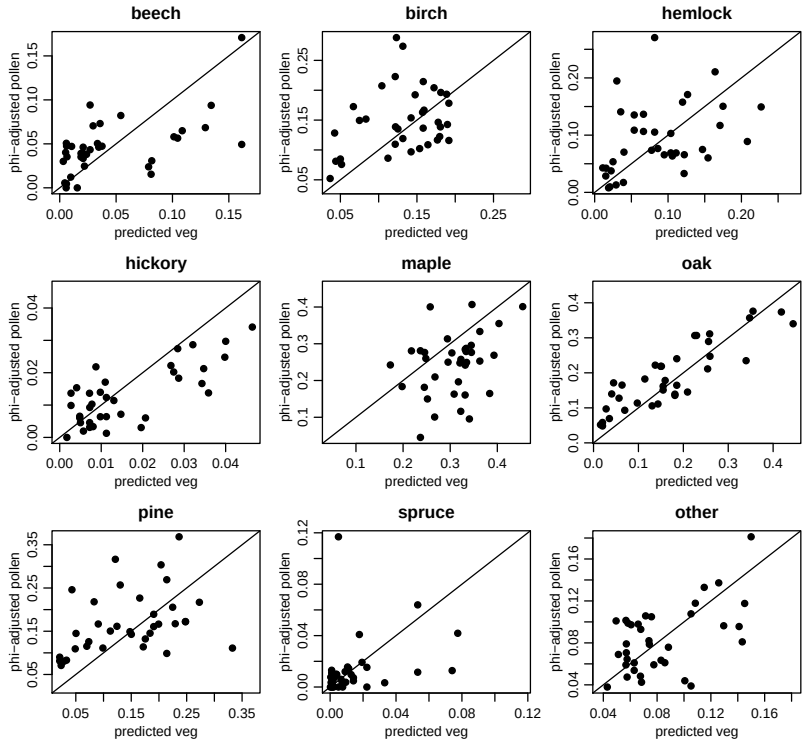

Map of growing degree days (courtesy of David Foster, Harvard Forest): note correspondence of spatial patterns

# Results(2): Pollen as a vegetation proxy

Unadjusted pollen-vegetation relationship

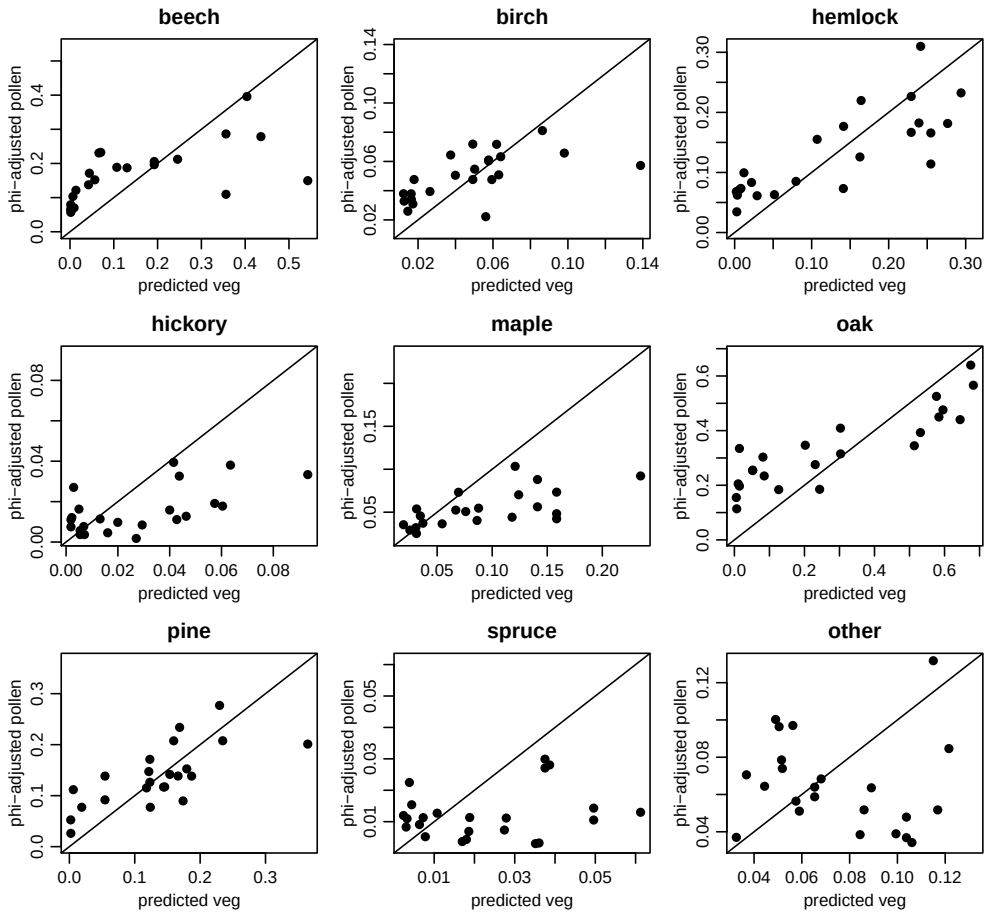



Adjusted pollen-vegetation relationship



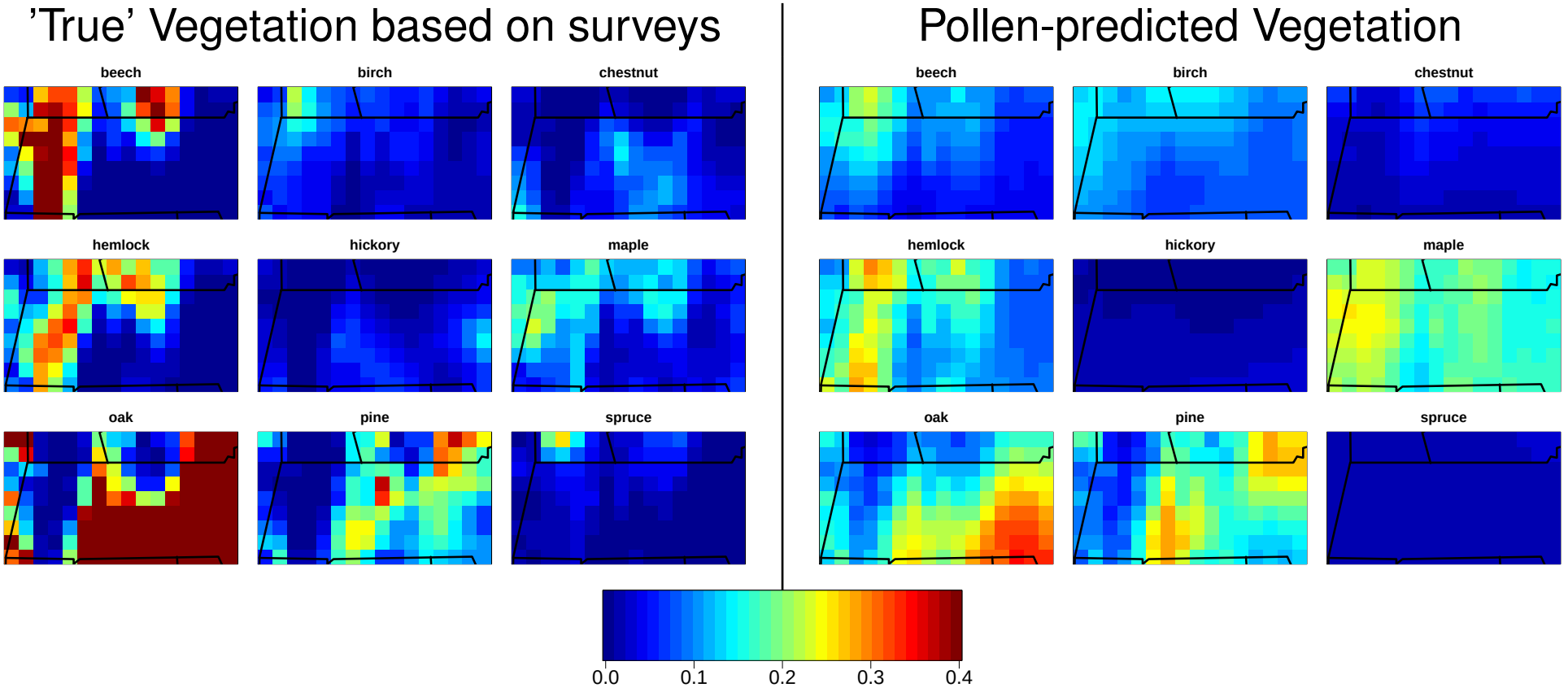

$\phi$ parameters scale pollen to vegetation. After adjustment, most ponds show reasonable, albeit noisy, relationships between pollen and vegetation estimated in the grid box.

# Results(3): Colonial pollen-vegetation mismatch



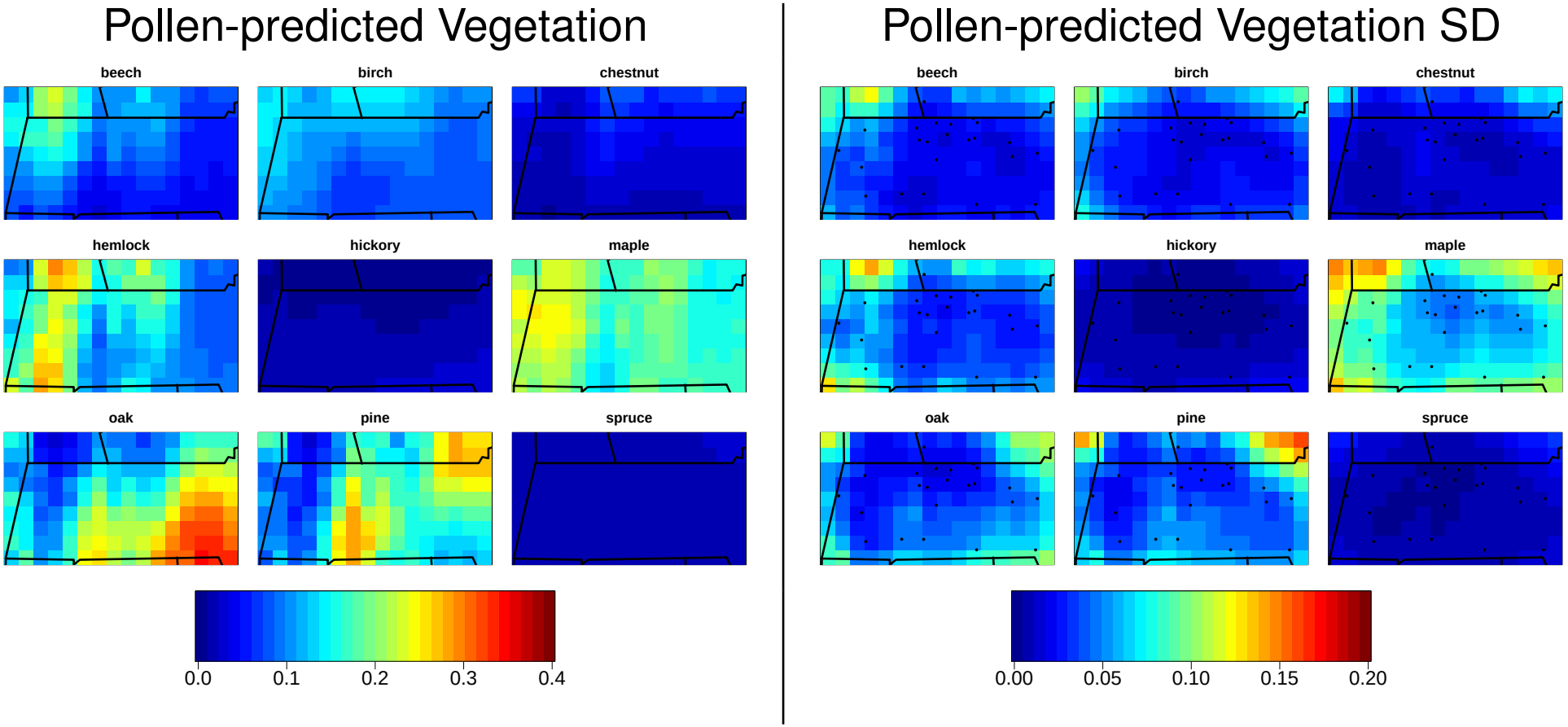

Note that for several taxa (beech, birch, hemlock, oak), pollen at low levels overestimates predicted vegetation based on colonial surveys. Perhaps the surveys missed relatively rare taxa?

# Results(4): Cross-validation: Colonial predictions using colonial pollen and modern parameter values



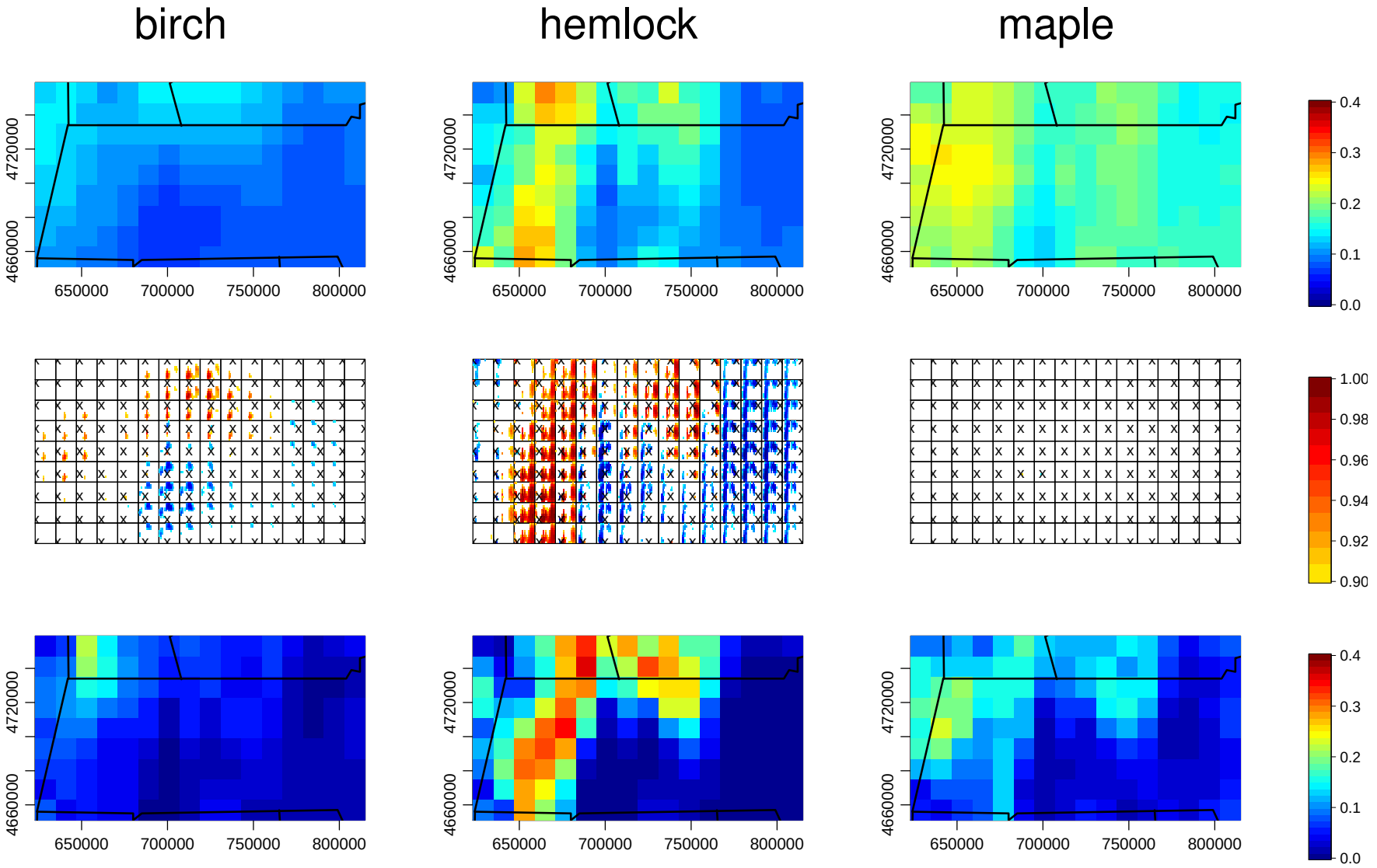

Model seems to predict well in terms of spatial patterns, though absolute abundances are off in some taxa/locations.

# Results(5): Graphical inference: posterior uncertainty



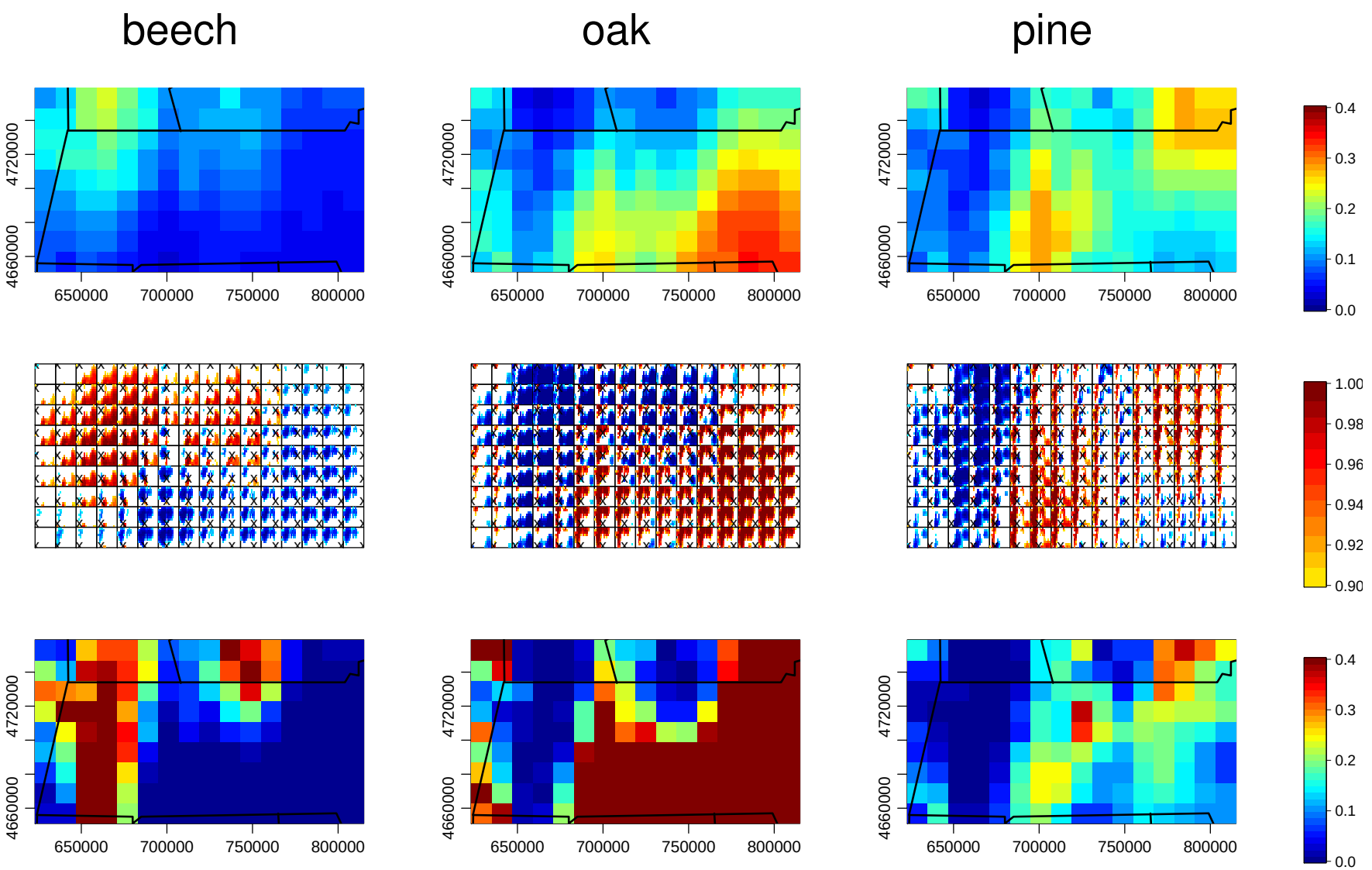

Less common taxa have least absolute uncertainty; maple with little pollen representation relative to its abundance in vegetation is hardest to predict. Uncertainty increases away from ponds.

# Results(5): Graphical inference: feature certainty



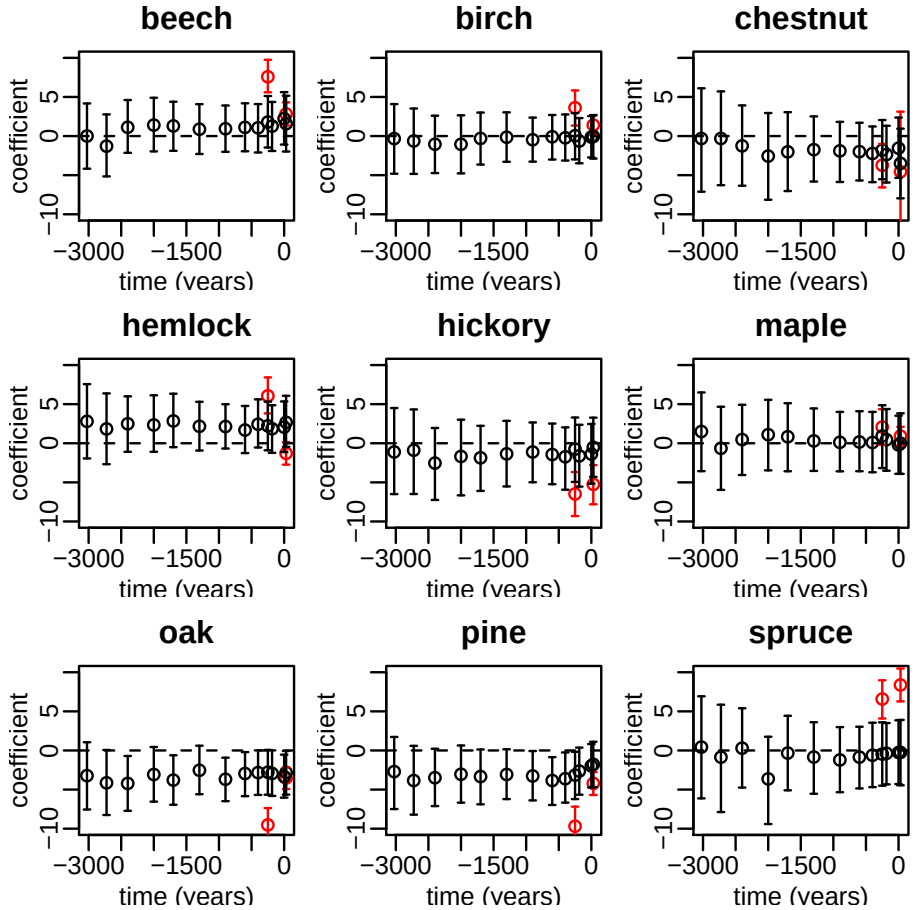

Top row: pollen predictions for colonial era; bottom row: estimates based on surveys.
Middle row shows pairwise posterior probabilities that one location has more or less of the taxa than another location, with each cell showing the contrasts of the location marked with an 'x' with all other locations.

# Results(5): Graphical inference: feature certainty

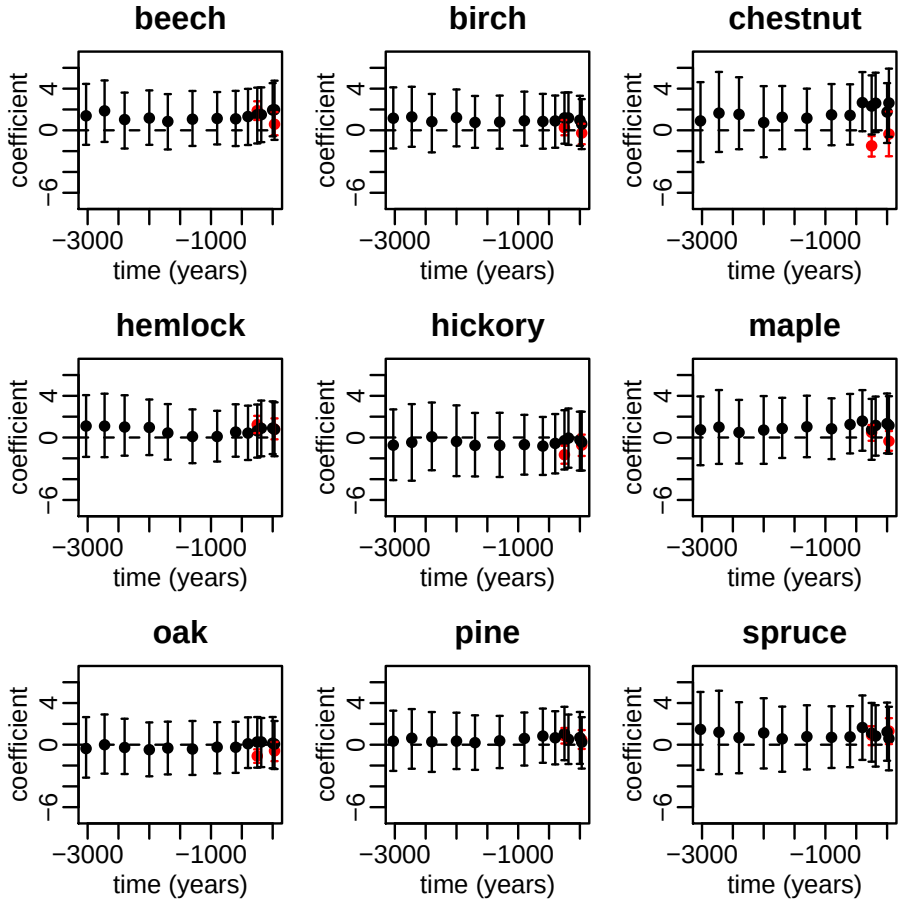


# Results(6): Covariate effects through time

Elevation

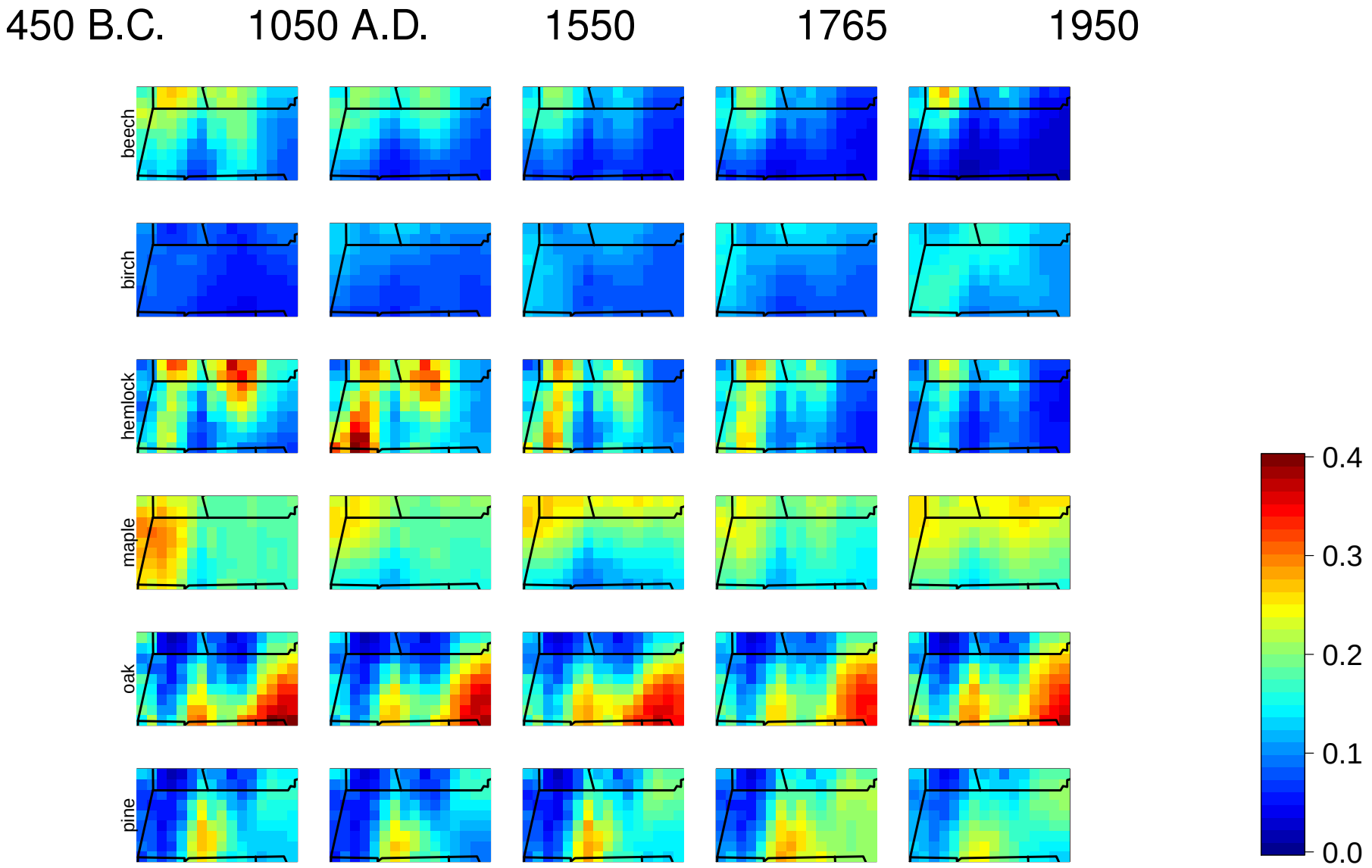



Latitude



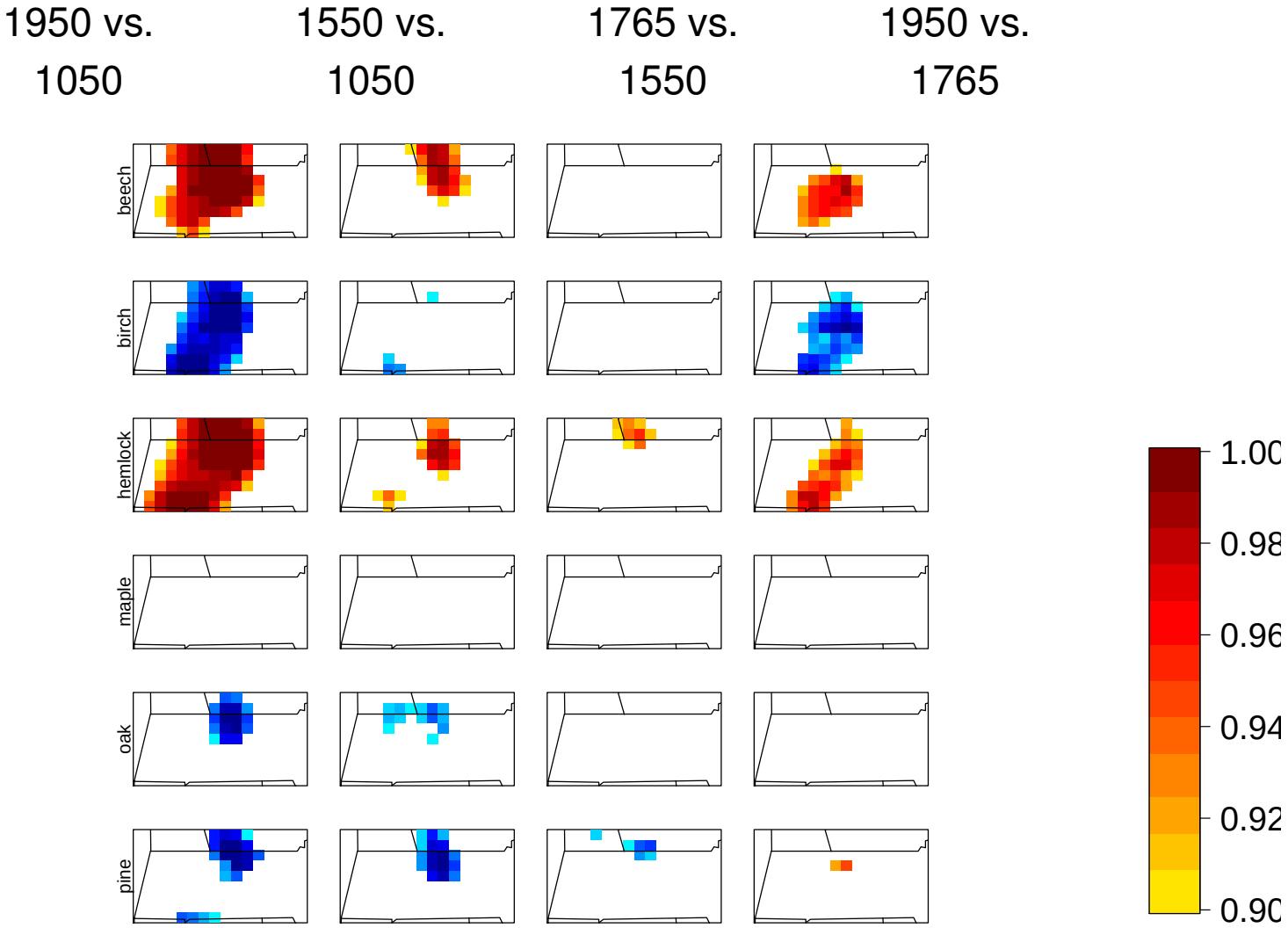

Covariate effects appear consistent through time (albeit with high uncertainty) although estimates based on vegetation data (in red for colonial and modern data) are more pronounced.

# Results(7): Prediction in time from pollen




Patterns are fairly similar over time. Can we detect changes over time?

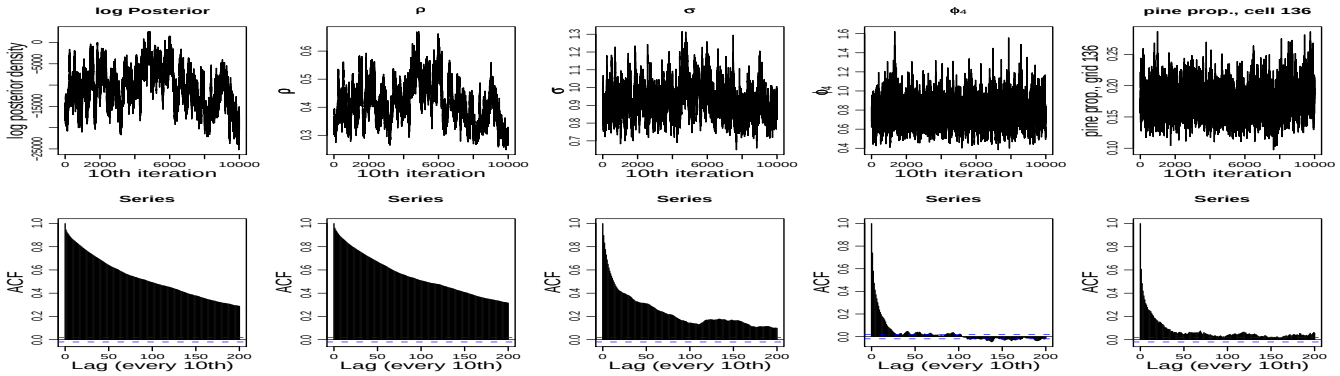
# Results(8): Significance of changes over time




Changes over time grid cell by grid cell are less certain than spatial heterogeneity. Here red indicates that the later time period has more of that taxa in the grid cell, with posterior probability given in legend, while blue indicates the earlier time period has more.

# MCMC performance

- MCMC mixing is rather slow; initial modern estimation run shown here:




- 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)
- Need longer runs and need prediction runs with propogation of hyperparameter $\{\phi, \alpha_{\text{pollen}}, \rho, \sigma\}$ uncertainty

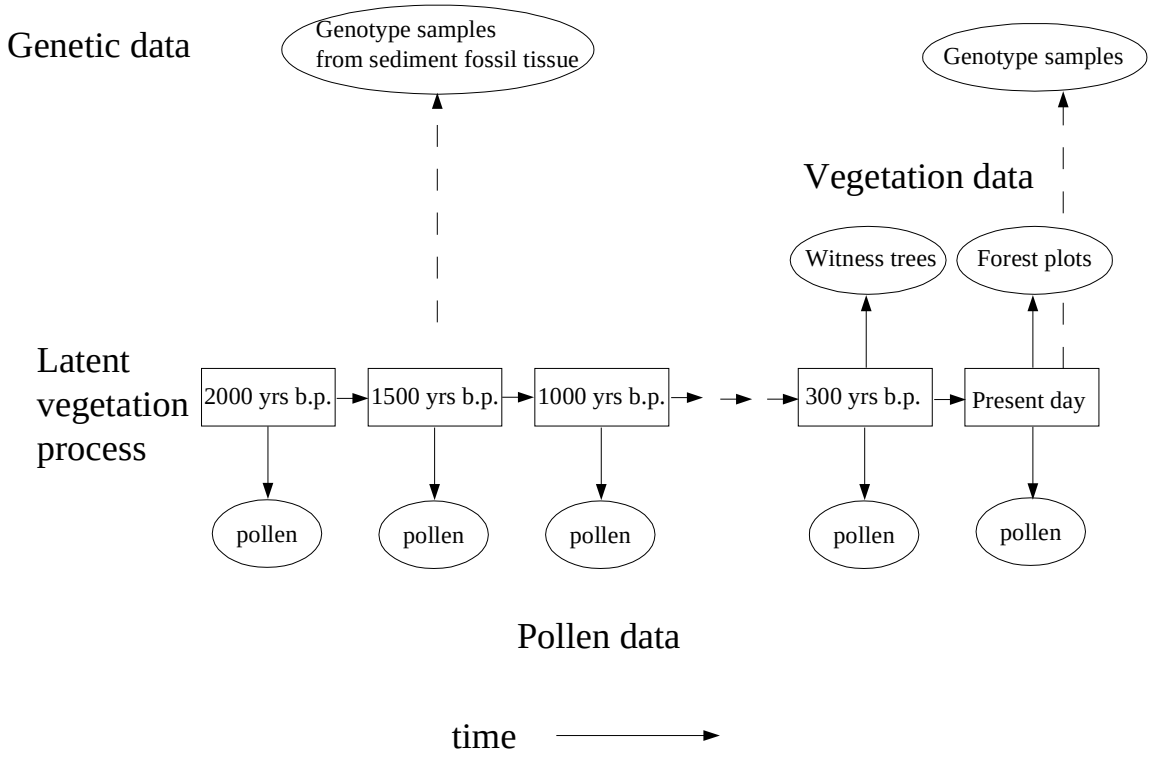
# Current challenges

- Some ponds poorly reflect ground-truth vegetation - why?
  - Relevant covariates to explain pond anomalies or pond-grid cell heterogeneity?
  - Covariates need to remain relevant over time
  - long-distance pollen dispersal?
- Sediment mixing can cause pollen to be misaligned in time
  - example of chestnut pollen appearing in modern surface sediments
  - less concern back in time as our time resolution is longer
- How much can we say definitively about changes over time?
- In estimation runs, should we allow pollen likelihood to influence composition estimates, $g_p(\cdot)$?
  - pollen data may better inform local vegetation if no plots nearby, but
  - danger of overfitting and underestimating grid-pollen heterogeneity
  - using only vegetation data for composition estimates may reduce issue of model misspecification for pollen component (Yucel & Zaslavsky 2005, cut function in BUGS)
- How to incorporate time smoothly?

# Next steps

- Continued model validation and sensitivity analysis
- Incorporation of less ad hoc approach to temporal structure:
  - smooth proportions for each pond over time and then model spatial structure at second stage: exact model form under development
- 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

# Future problem structure



# 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
- $\Psi$ 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)