SAND2023-05793T

Final Exam Presentation

Mauricio Campos

UIUC

June 30, 2023

Motivation

Contents

Integrating Different Data Sources Using a Bayesian Hierarchical Model to Unveil Glacial Refugia

In collaboration with Bo Li, Feng Sheng Hu, Joseph Napier and Guillaume de Lafontaine

Mauricio Campos

UIUC

June 30, 2023

Outline

Motivation

Motivation

- Rapid anthropogenic climate change creates a challenging scenario for the survivability of species.
- Understanding the biotic response of species during the climate variation in the paleorecord may prove essential moving forward.
- It was previously believed that species migrated south to warmer climates and later repopulated the north once ice sheets receded (Deevey, 1949; Van der Hammen et al., 1971; Davis, 1976) but new evidence points to the presence of "cryptic refugia" that might've played a bigger role than previously believed (McLachlan et al., 2005; Stewart et al., 2010; Mosblech et al., 2011).

Cryptic Refugia

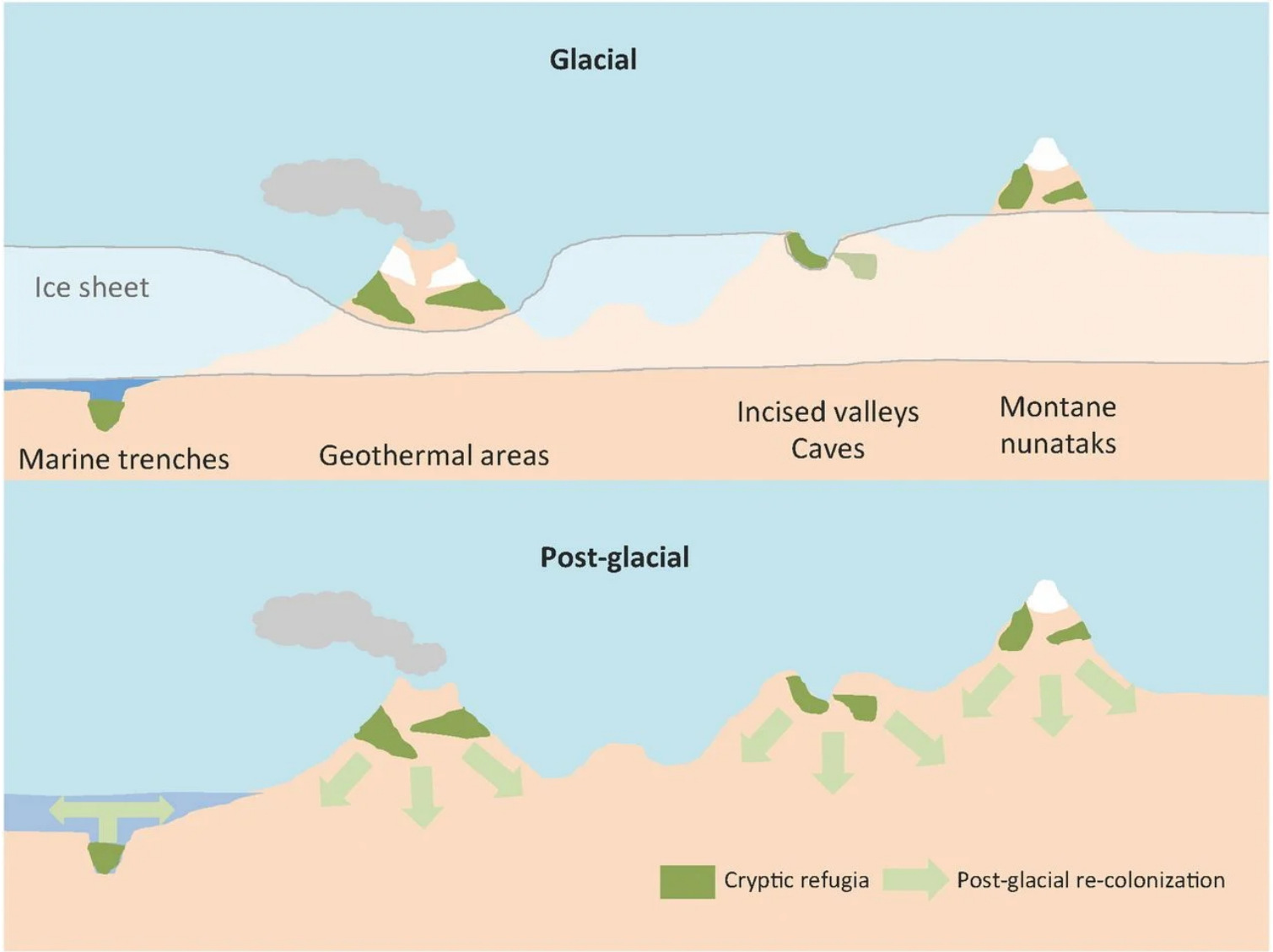


Figure: Role of cryptic refugia in post-glacial species migration. Taken from Pointing et al., 2014.

Problem

- Beringia is a region that has been featured prominently in the literature of glacial refugia.
- Recent studies suggest that a number of arcto-boreal species survived the Last Glacial Maximum (LGM) in the region.
- This study will focus mainly on *alnus viridis* and *picea glauca*, which are common constituents of fossil pollen records in Eastern Beringia.
- There are different types of seemingly conflicting data sources which we wish to integrate in some manner.

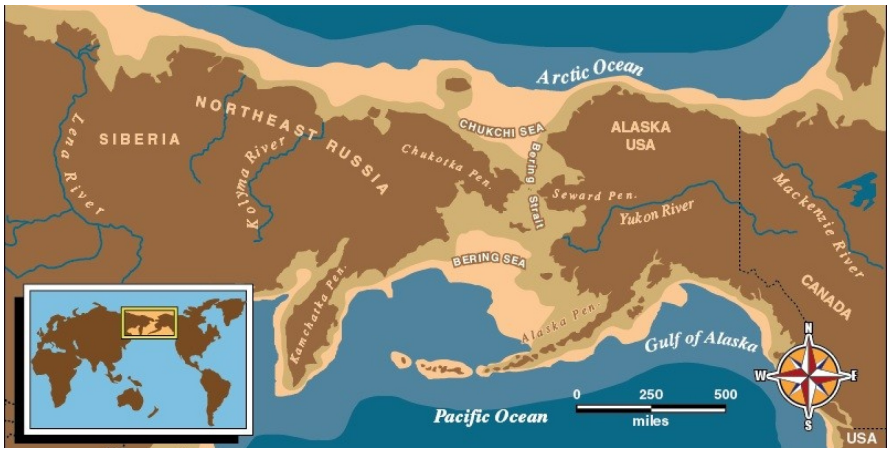


Figure: Beringia during LGM

Figure: Green Alder (*Alnus viridis*)

Data Sources: Fossil Records

- The presence of a fossil dating back to the LGM is direct evidence that the site used to be refugia.
- Plant fossils are very rare compared to animal ones, but pollen fossils can still be recovered from lake-sediment cores.

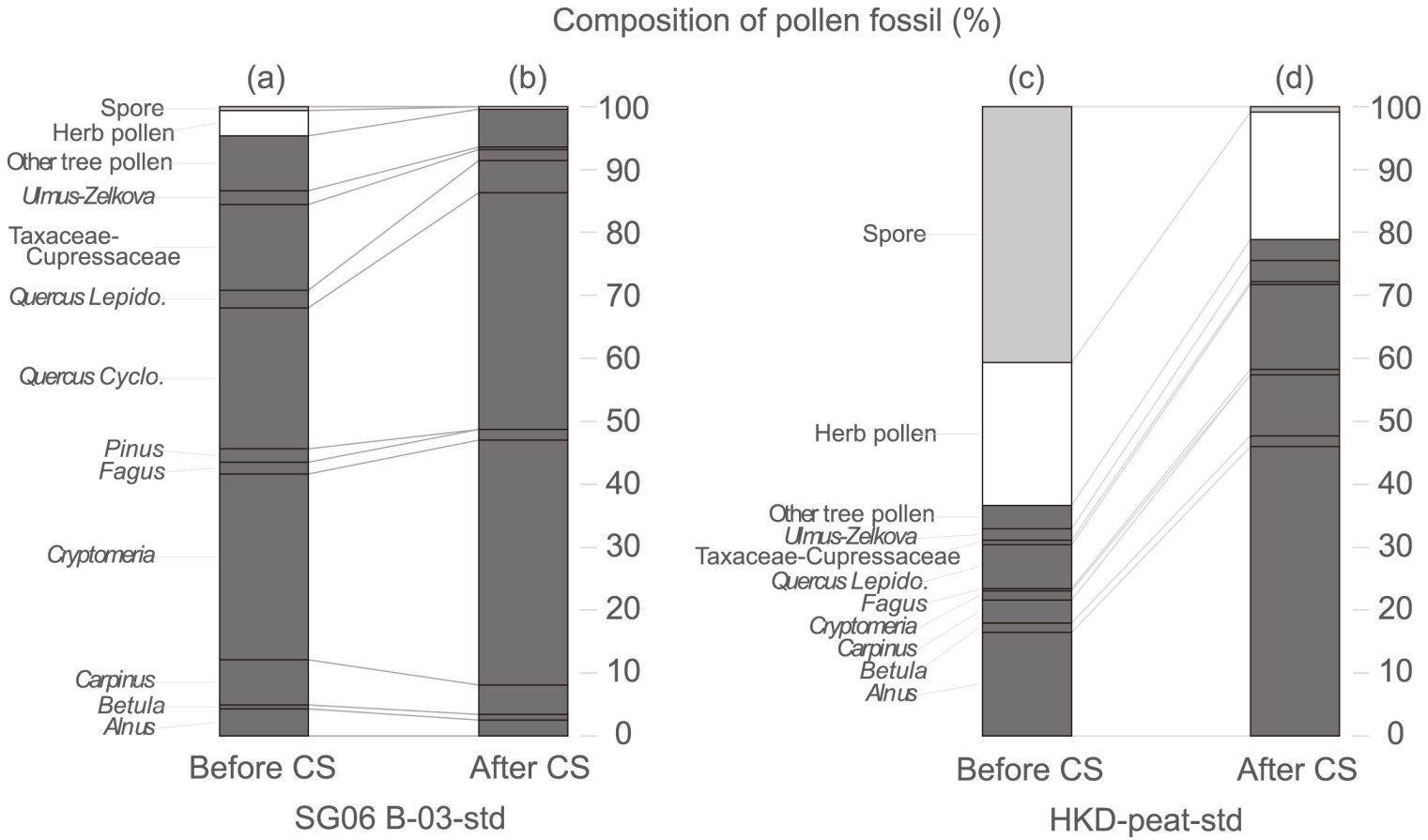


Figure: Taken from Yamada et al., 2021

Data Sources: Phylogeography

- From modern-day DNA samples we can infer the past evolutionary scenarios that generated the observed modern-day genetic lineages.
- Each site where samples are collected will be represented by it's genetic lineage composition.

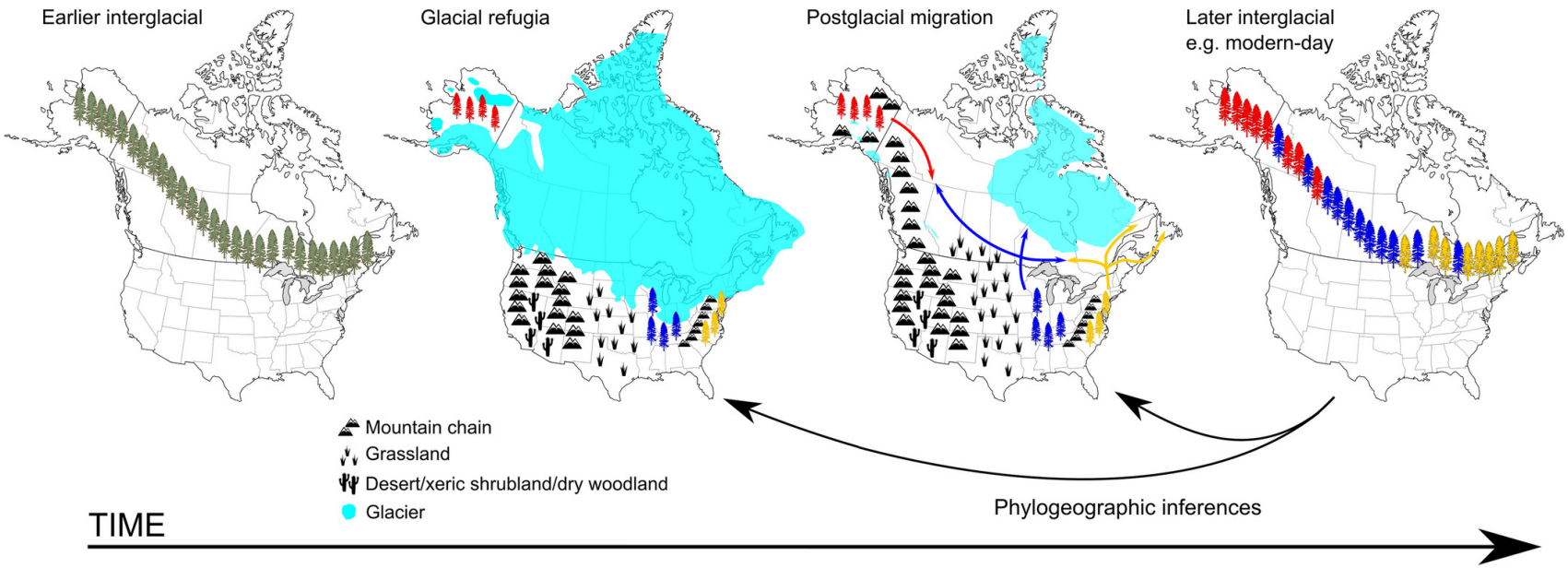
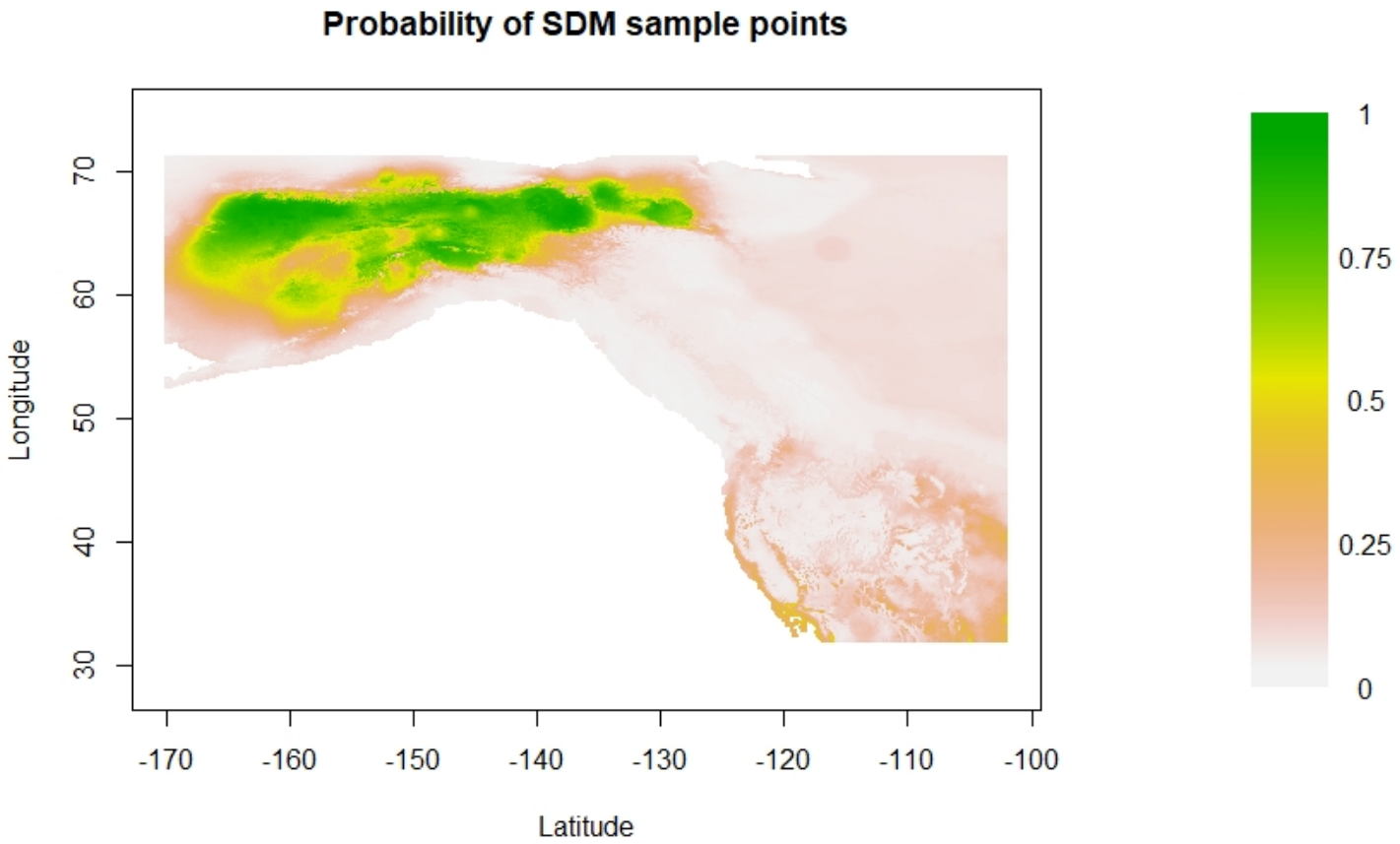


Figure: Taken from De Lafontaine et al., 2018

Data Sources: Species Distribution Models

- SDMs combine modern-day ocurrence of the species with climate variables to then reconstruct where the species could have survived in the past.
- Provides insight on regions where climate conditions might have been suitable for species survival but provides no direct evidence of presence.



Previous Studies

- Most studies focus mostly on one source of evidence and rarely combine them.
- SDM is often used as a filter that determines where to conduct analysis on pollen or genetic data (Espíndola et al., 2012; Napier et al., 2019).
- Common statistical methods:
  - MANOVA (Knowles and Alvarado-Serrano, 2010; Brown and Knowles, 2012)
  - Likelihood-based hypothesis testing (Lemmon and Lemmon, 2008; Lemey et al., 2009; Marske et al., 2012)
  - Approximate Bayesian Calculations (Gao et al., 2012; Tsuda et al., 2016; Aoki et al., 2019)
  - Bayesian Hierarchical Models (Marion et al., 2012; Pagel y Schurr, 2012; Schurr et al., 2012)

Bayesian Hierarchical Model

- Have been used successfully to integrate different sources of data in paleoclimate and paleocological studies (e.g. Li et al., 2010; Urban et al., 2013).
- The flexible modeling framework allows to take into account the unique characteristics of each data type (Clark, 2005).
- Advances in computation power as well as alternatives to MCMC, such as INLA (Rue et al., 2009), have made it possible for the estimation to be timely and efficient.

Data

Data

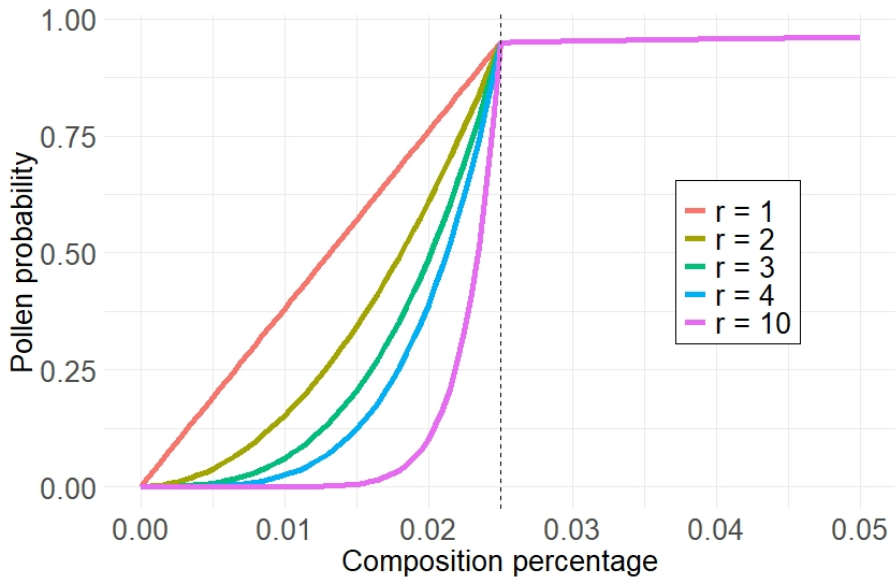
- We have evidence from three different sources:
  1. Species Distribution Models
  2. Genetic Data
  3. Fossil pollen records
- Sample sizes for each data source are drastically different.
- We will use subscripts m, g, p to denote SDM, Genetic and Pollen data, respectively. Since there are no repeated sites $n = n_m + n_g + n_p$.
- Data will have to be transformed prior to modeling so they more accurately resemble probabilities.

Pollen

For a given site $\boldsymbol{s}$ we have the percentage of the core composed by pollen fossil, $c(\boldsymbol{s})$. The higher $c(\boldsymbol{s})$ then the higher should be the probability of refugia according to pollen, $p_p(\boldsymbol{s})$.

$$p_p(\boldsymbol{s}) = \begin{cases} \left(\frac{c(\boldsymbol{s})}{\tau_p}\right)^r \gamma_p & \text{if } c(\boldsymbol{s}) \leq \tau_p \\ \{2c(\boldsymbol{s})\}^{\log(\gamma_p)/\log(2\tau_p)} & \text{if } c(\boldsymbol{s}) > \tau_p. \end{cases}$$

- $\tau_p = 0.025$ for green alder and $\tau_p = 0.01$ for white spruce (Napier et al., 2019; Warrent et al., 2016)
- $\gamma_p = 0.95$ for green alder and $\gamma_p = 0.90$ for white spruce.



Genetic Data

For a given site $\boldsymbol{s}$, the percentage of the genetic composition of the sample that belongs to a lineage l is $q_l(\boldsymbol{s})$ (i.e. $\sum_{l=1}^{L} q_l(\boldsymbol{s}) = 1$). A site that is dominated by a single lineage will be more evidence of refugia. Let $\tilde{q}(\boldsymbol{s}) = \max_l \{q_l(\boldsymbol{s})\}$.

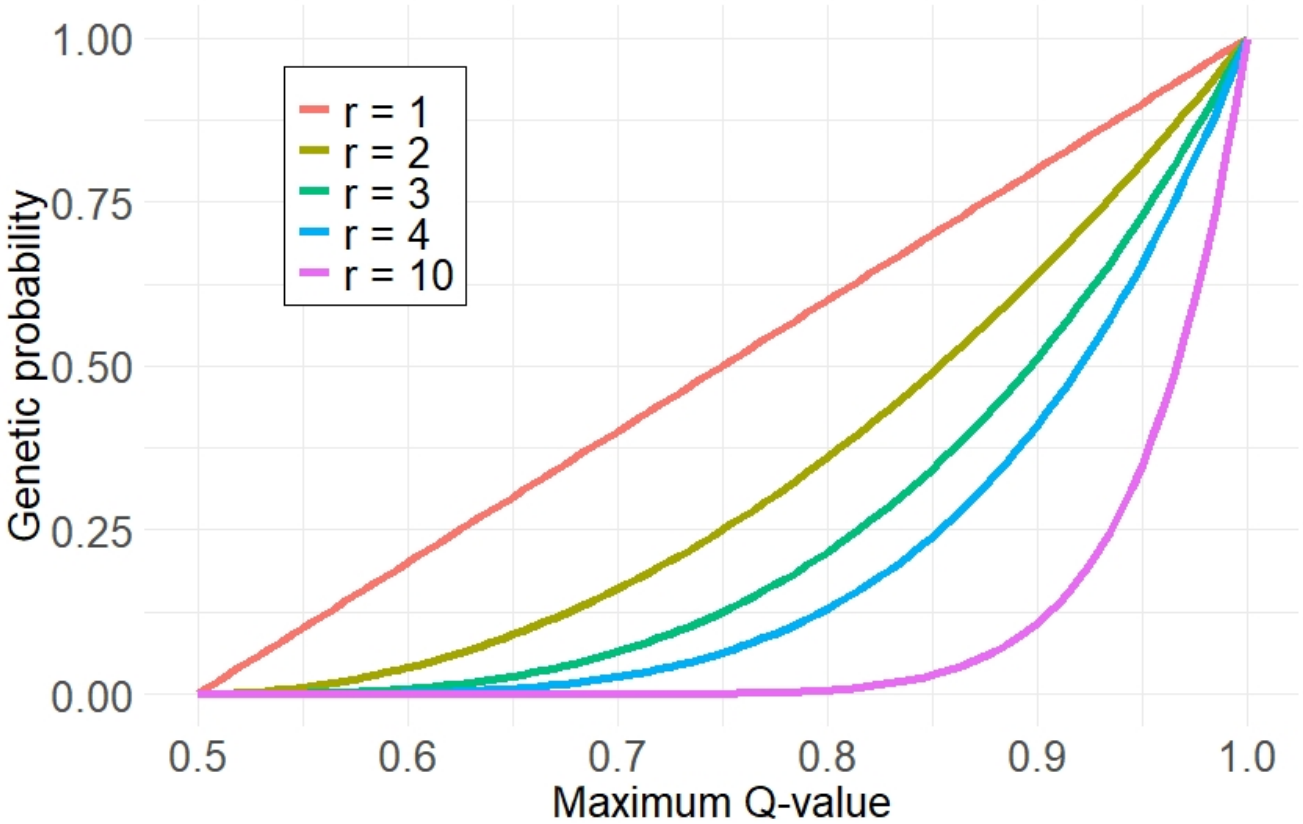
Each species has different number of lineages, L:

- Green alder: $L = 2$
- White spruce: $L = 5$

Genetic Data: Green Alder

With only 2 lineages, we check whether $\tilde{q}(\boldsymbol{s})$ is close to 0.5 or 1 and interpolate in between:

$$p_g(\boldsymbol{s}) = \{2\,(\tilde{q}(\boldsymbol{s}) - 0.5)\}^r\,.$$



Genetic Data: White Spruce

A site $\boldsymbol{s}$ will belong to the k-th lineage if $\arg\max_l \{q_l(\boldsymbol{s})\} = k, l = 1, 2, ..., 5$. This way, each lineage will be composed of m_l mutually exclusive sites. Re-define $\tilde{q}_k(\boldsymbol{s}_i) = \max_l q(\boldsymbol{s}_i), i \in \{1, 2, ..., m_k\}$ to now show the lineage of precedence.
Furthermore, define:

- $\xi_k = \max_i \tilde{q}_k(\boldsymbol{s}_i)$
- $\delta_k = \min_i \tilde{q}_k(\boldsymbol{s}_i)$
- $q_{max} = \max_k \xi_k = \max_{i',l} q_l(\boldsymbol{s}_{i'})$
- $q_{min} = \min_{i',l} q_l(\boldsymbol{s}_{i'}), i' \in 1, 2, ..., n_g$

$$p_g(\boldsymbol{s}) = \frac{q_{min}[\xi_k - \tilde{q}_k(\boldsymbol{s})] + q_{max}[\tilde{q}_k(\boldsymbol{s}) - \delta_k]}{\xi_k - \delta_k}.$$

White Spruce genetic data visualization

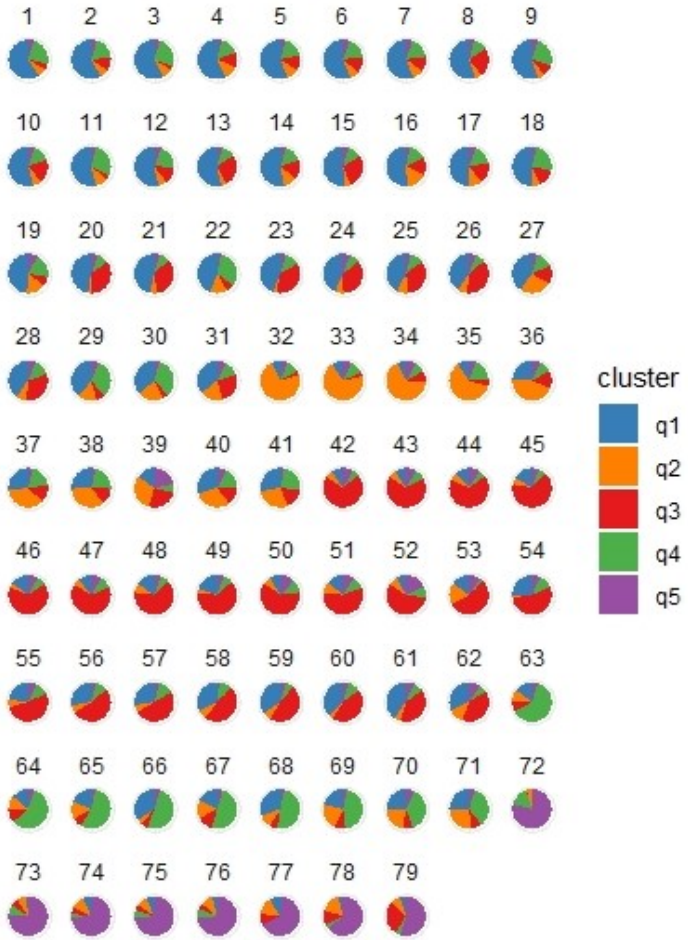


Figure: Lineage composition

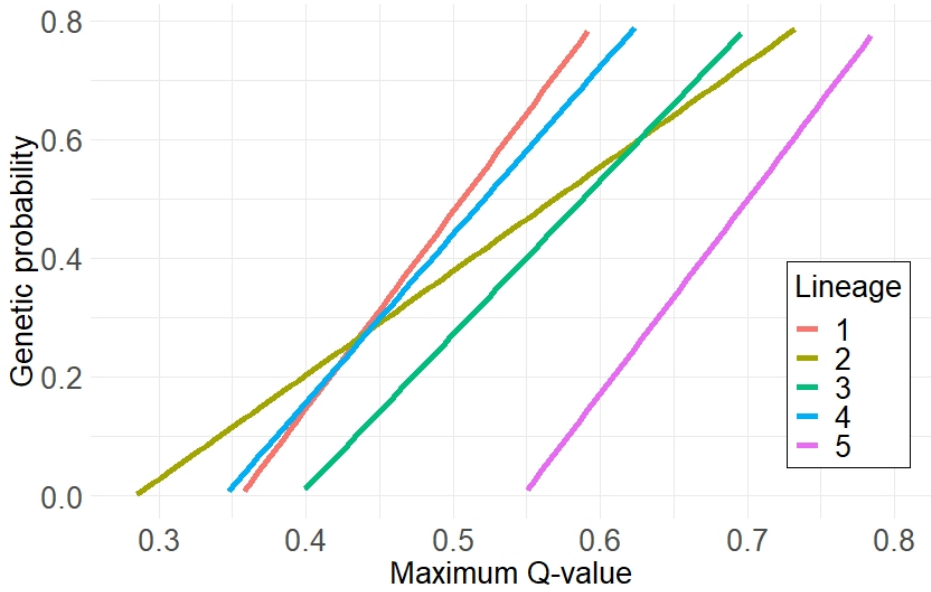
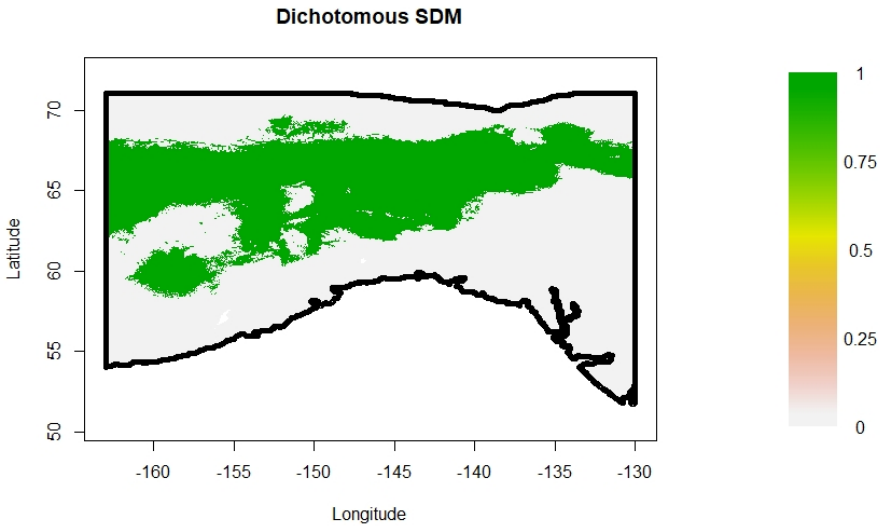


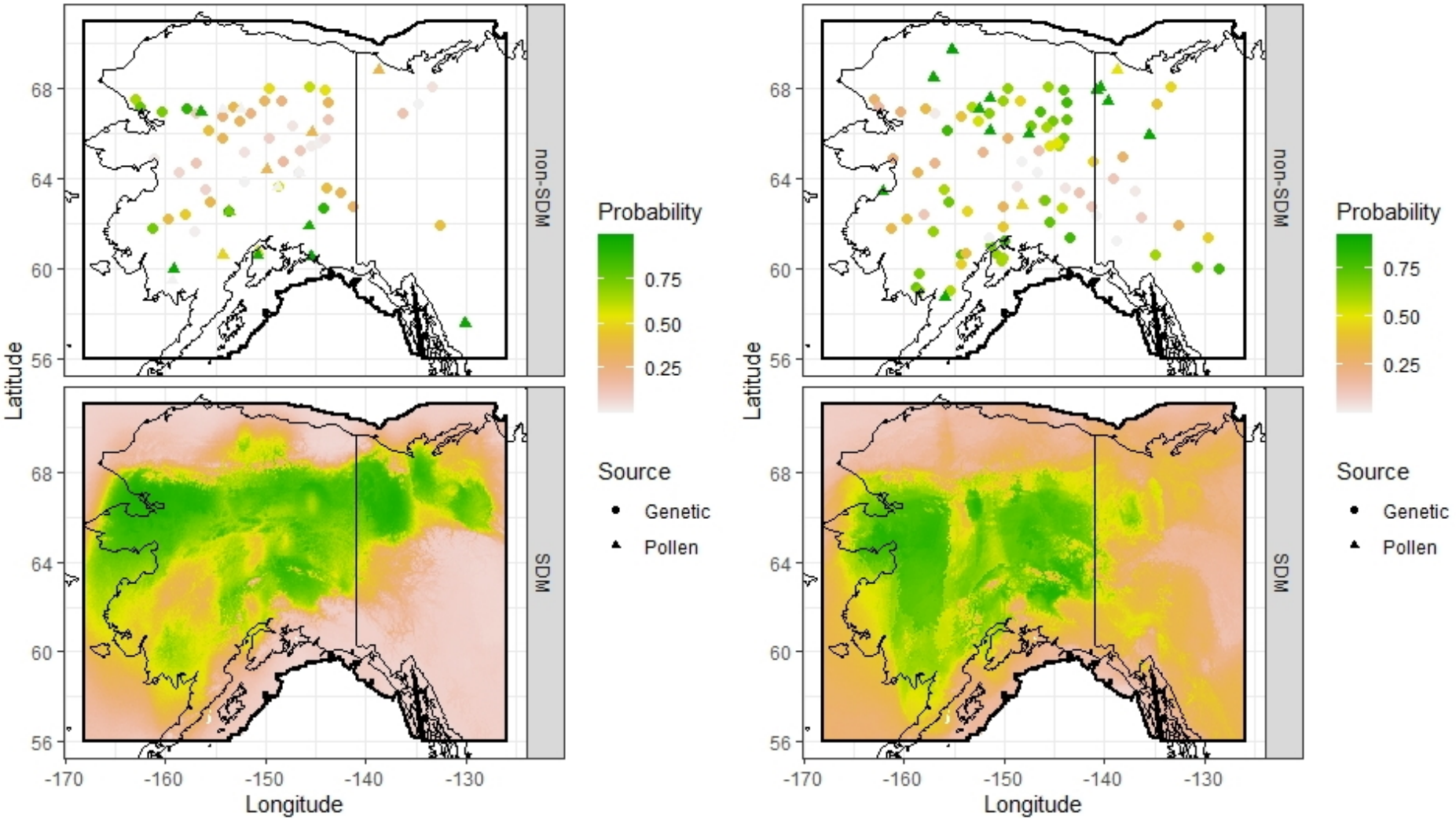
Figure: Interpolations

SDM

- Since SDM can be interpreted as the noisiest of all data, we will dichtomize the information into sites that correspond to good or bad survivability conditions.
  - We do have more certainty of where the species couldn't have been than the opposite.
- The dichotomization is done according to the True Skill Statisitc (Allouche et al., 2006), which varies for each species:
  - $\tau_m = 0.54$ for green spruce
  - $\tau_m = 0.506$ for white alder



Data Visualization



(a) (b)

Figure: Observed data from all sources for (a) green alder and (b) white spruce.

Model

Hierarchical Model: Level 1 - Data Models

Assume there is a true probability of refugia from which our observed data can be modeled as deviations from this truth. Let $\boldsymbol{P}(\boldsymbol{s})$ represent the true probability of site $\boldsymbol{s}$ being refugia. We will assume that these probabilities come from a Gaussian Process (i.e. $\mu + \boldsymbol{X}(\boldsymbol{s}) = \Phi^{-1}(\boldsymbol{P}(\boldsymbol{s}))$). Additionally, let $Y_p(\boldsymbol{s}) = \Phi^{-1}(P_p(\boldsymbol{s}))$, $Y_g(\boldsymbol{s}) = \Phi^{-1}(P_g(\boldsymbol{s}))$ and $Y_m(\boldsymbol{s}) = 1\{P_m(\boldsymbol{s}) \geq \tau_m\}$.

$$
\begin{aligned}
\text{logit}(P_m(\boldsymbol{s})) &= \alpha_m + \beta_m\{\mu + X(\boldsymbol{s})\} + Z(\boldsymbol{s}), & \boldsymbol{Z} &\sim GP\left(\boldsymbol{0}, \boldsymbol{\Sigma}(\sigma_m^2(\boldsymbol{s}), \rho_m)\right) \\
Y_p(\boldsymbol{s}) &= \alpha_p + \beta_p\{\mu + X(\boldsymbol{s})\} + \epsilon_p, & \epsilon_p &\sim N(0, \sigma_p^2), \\
Y_g(\boldsymbol{s}) &= \mu + X(\boldsymbol{s}) + \epsilon_g, & \epsilon_g &\sim N(0, \sigma_g^2),
\end{aligned}
\tag{1}
$$

The non-stationary SDM variance is modeled as $log(\sigma_m(\boldsymbol{s})) = \theta_1 + \theta_2 \cdot 1\{\boldsymbol{s} \in S_{m1}\}$ where $S_{m1} = \{\boldsymbol{s} : Y_m(\boldsymbol{s}) = 1\}$.

Hierarchical Model: Level 2 - Latent Spatial Process

We model the latent spatial process with a mean-zero Gaussian process with a Matérn covariance function with fixed smoothness $\nu = 1$.

$$X(\boldsymbol{s}) \sim GP\left(\mathbf{0}, \boldsymbol{\Sigma}(\sigma_x^2, \rho_x)\right). \tag{2}$$

Hierarchical Model: Level 3 - Priors

$$
\begin{aligned}
\mu, \alpha_m, \alpha_p &\sim N(0, 1000), \\
\beta_m, \beta_p &\sim N(1, 1000), \\
\log(1/\sigma_m^2), \log(1/\sigma_p^2) &\sim \text{LogGamma}(1, 0.00005), \\
\sigma_g &\sim \text{PC Prior s.t. } P(\sigma_g > 1) = 0.01, \\
\sigma_x^2 &\sim \text{PC Prior s.t. } P(\sigma_x > 3) = 0.01, \\
\rho_x &\sim \text{PC Prior s.t. } P(\rho_x < 1) = 0.01, \\
(\theta_1, \theta_2, \log \rho_m)^T &\sim N\left((0, 1, 0)^T, \boldsymbol{I}_3\right).
\end{aligned}
$$

The PC prior is the Penalized Complexity Prior from Simpson et al. (2017). It serves as a weakly informative prior that penalizes complexity of a hierarchical model structure.

Estimation

Bayesian Methods

MCMC methods

- MCMC methods tend to exhibit poor performance when applied to such models.
- The components of the latent field are strongly dependent of each other.
- Methods that overcome this still remain painfully slow.

Machine Learning approaches

- Most well known are Variational Bayes and Expectation Propagation.
- Find $q(\boldsymbol{x}, \boldsymbol{\theta})$ that minimizes the KL divergence of $\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$ w.r.t. $q(\boldsymbol{x}, \boldsymbol{\theta})$ for VB, or vice-versa for EP, subject to $q(\boldsymbol{x}, \boldsymbol{\theta})$ factorizing in a 'nice' way (e.g. $q(\boldsymbol{x}, \boldsymbol{\theta}) = q_{\boldsymbol{x}}(\boldsymbol{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$).
- These methods tend to underestimate/overestimate the marginal variances and do a poor job detecting the dependence between $\boldsymbol{x}$ and $\boldsymbol{\theta}$.

INLA

INLA (Rue, Martino and Chopin, 2009) is based on the following Laplace approximation:

$$\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \left. \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x}=\boldsymbol{x}^*(\boldsymbol{\theta})} \tag{3}$$

where $\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ is the Gaussian approximation to the full conditional of $\boldsymbol{x}$ and $\boldsymbol{x}^*(\boldsymbol{\theta})$ is the mode of the full conditional for $\boldsymbol{x}$, for a given $\boldsymbol{\theta}$.

The density of $x_i|\boldsymbol{\theta}, \boldsymbol{y}$ is approximated with the Gaussian marginal derived from $\tilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$, i.e.

$$\tilde{\pi}(x_i|\boldsymbol{\theta}, \boldsymbol{y}) = \mathcal{N}\{x_i; \mu_i(\boldsymbol{\theta}), \sigma_i^2(\boldsymbol{\theta})\} \tag{4}$$

where $\boldsymbol{\mu}(\boldsymbol{\theta})$ is the mean of the Gaussian approximation and $\boldsymbol{\sigma}^2(\boldsymbol{\theta})$ the corresponding marginal variances.

INLA (continued)

INLA approximation is computed in three steps:

1. Approximate the posterior marginal of $\boldsymbol{\theta}$ by using the Laplace approximation (3).
2. Compute the Laplace approximation, or the simplified version, of $\pi(x_i|\boldsymbol{y},\boldsymbol{\theta})$, for selected values of $\boldsymbol{\theta}$, to improve on the Gaussian approximation (4).
3. Combine the previous two by using numerical integration

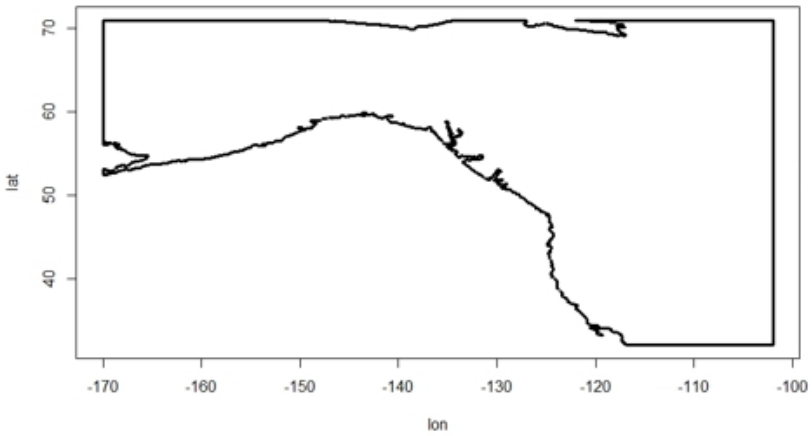
$$\tilde{\pi}(x_i|\boldsymbol{y}) = \sum_k \tilde{\pi}(x_i|\boldsymbol{\theta}_k,\boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}_k|\boldsymbol{y})\Delta_k \tag{5}$$

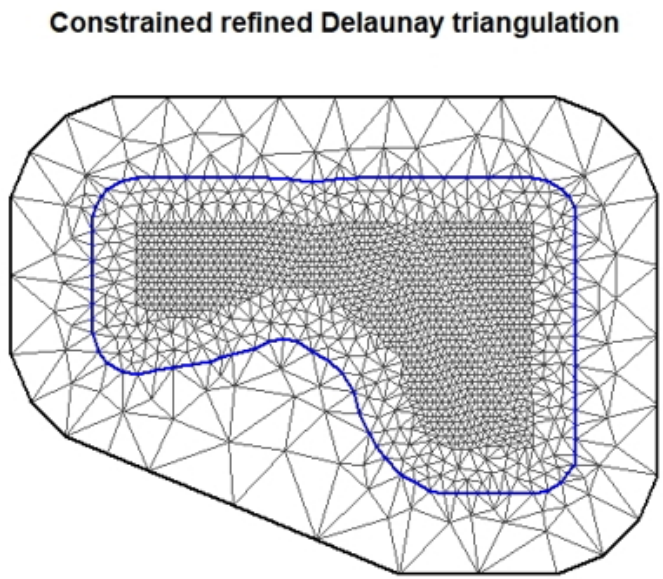
The main use of $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ from (3) is to integrate out $\boldsymbol{\theta}$ when approximating the posterior marginal of x_i. For this purpose, $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ must be explored sufficiently well to be able to select good evaluation points for the numerical integration.

Using INLA in practice

- INLA assumes $\boldsymbol{x}$ is a Gaussian Markov Random Field, which is a discretely indexed random field. However, most spatial analysis happens on continuously indexed random fields, modeled through more intuitive covariance functions (e.g. Matérn).
- Lindgren, Rue and Lindstöm (2011) use a SPDE which has a GF with Matérn covariance as its stationary solution, to construct a GMRF representation.
  - The spatial field is represented in a mesh where the GMRF is fitted at each vertex.
- All of this can be easily implemented via the R-INLA package in R. (Rue et al., 2009, 2017).

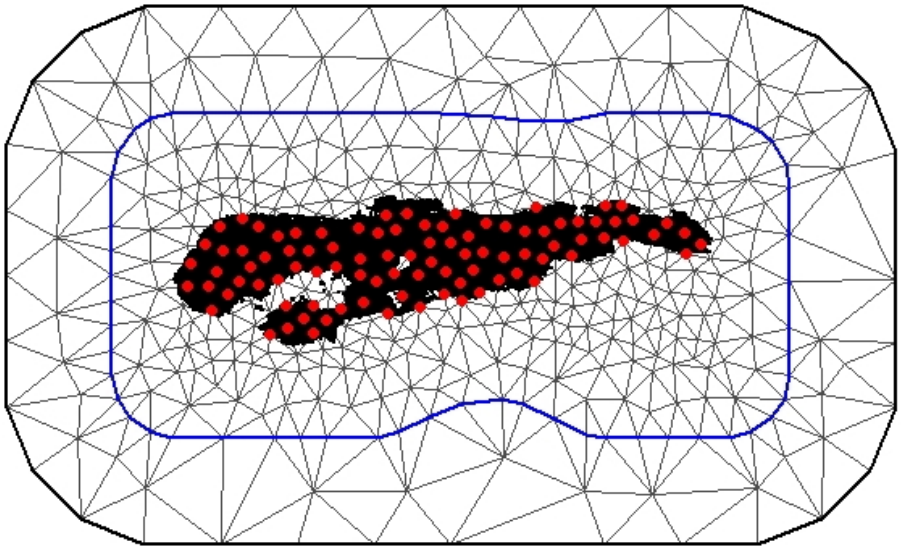
Delaunay Triangulation Mesh

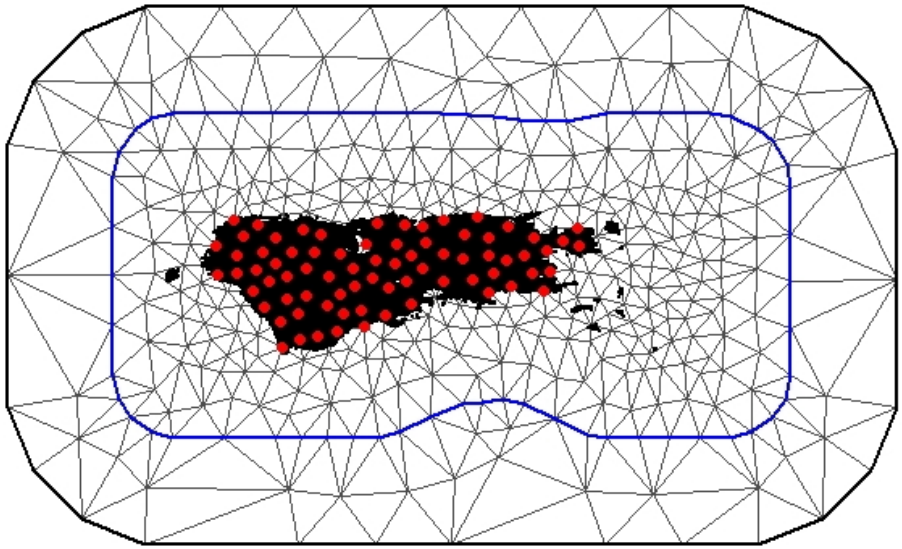




Non-Stationarity on the mesh

Since estimation is done on the mesh vertices, we must define S_{m1} on the mesh as well. For a mesh vertex v we say $v \in S_{m1}$ if $|\{s \in S_{m1} : ||s - v||_2^2 \leq r\}| > \delta$, for some r and δ.





Simulation Study

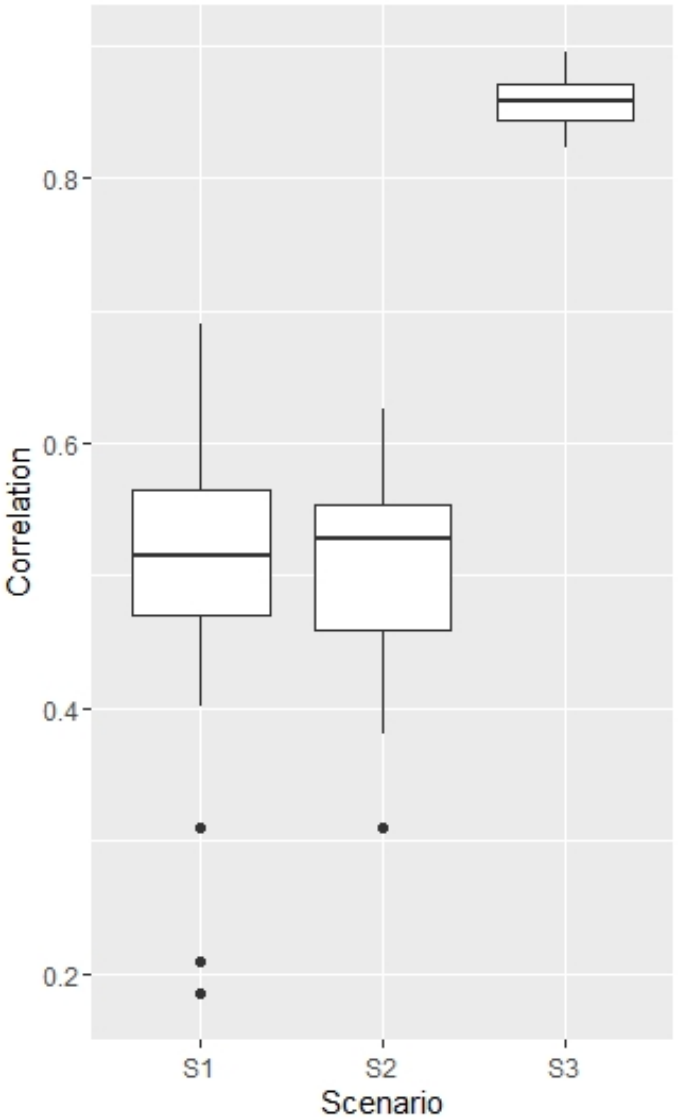
Setup

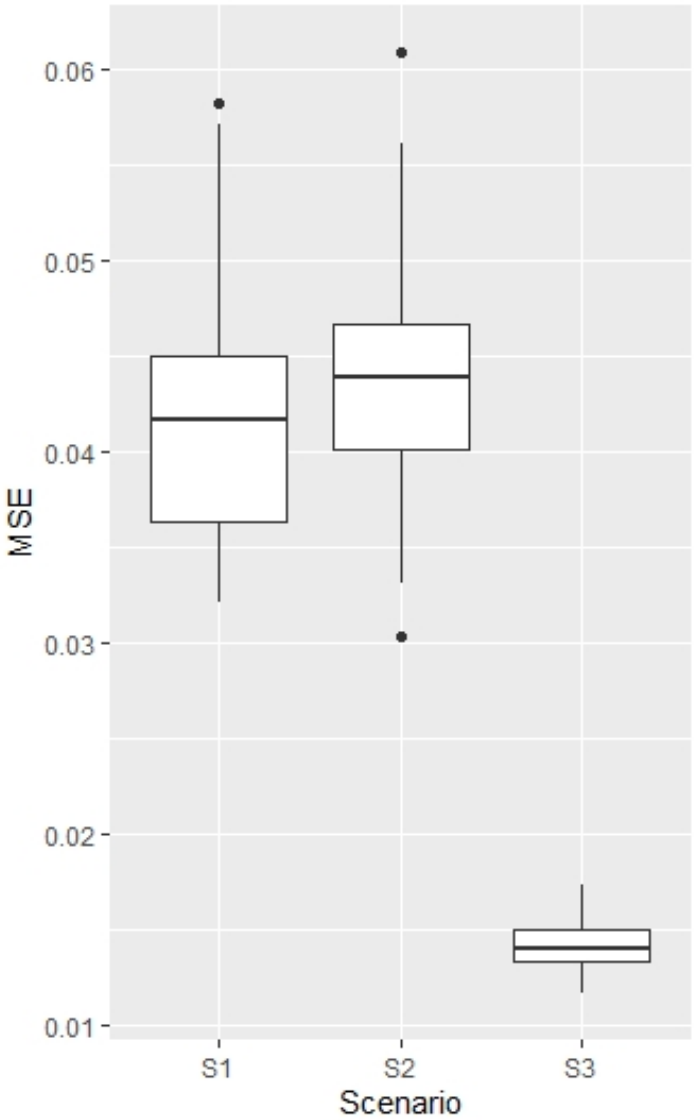
We perform simulations to assess the performance of our model in recovering the latent refugia probability $\boldsymbol{P}(\boldsymbol{s})$. We also study the effect of different sample sizes based on three different sample scenarios:

Sampling Scenario	n_m	n_p	n_g
S1	500	15	60
S2	1000	15	60
S3	1000	1000	1000

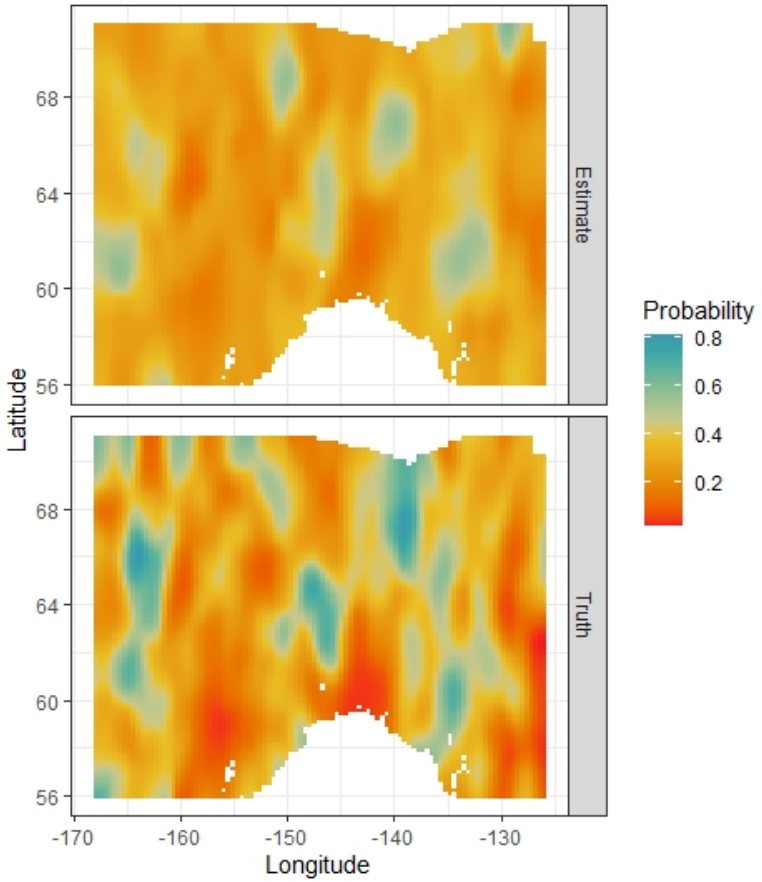
All simulations are done in the study region and data is simulated following the processes in (2) and (1). The correlation and mean square error (MSE) between $\boldsymbol{P}(\boldsymbol{s})$ and $\hat{\boldsymbol{P}}(\boldsymbol{s})$ are observed.

Simulation Results

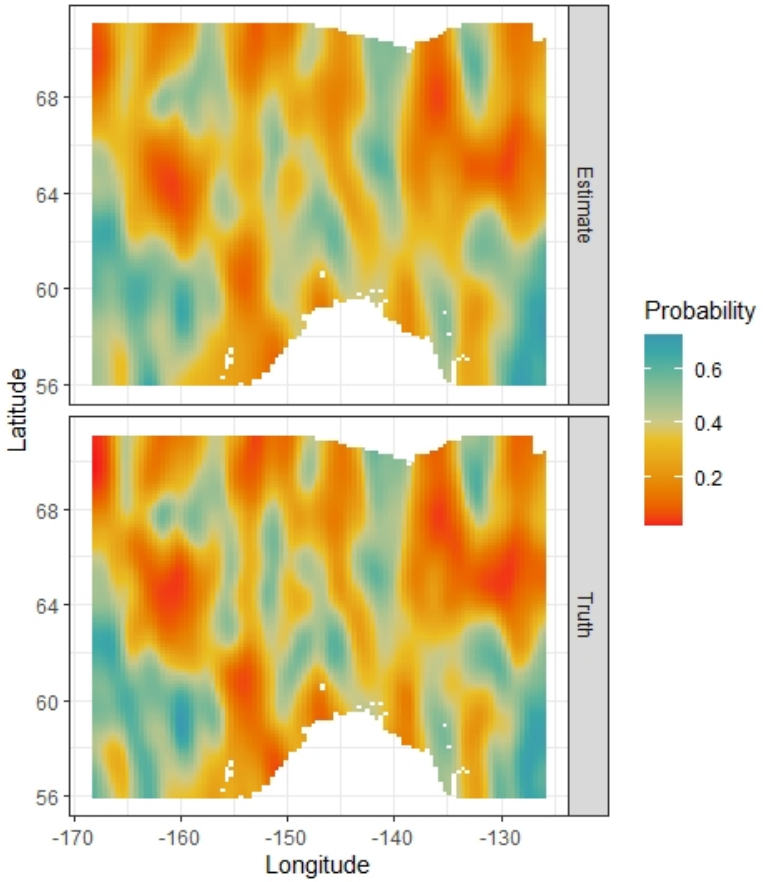




Example Reconstruction



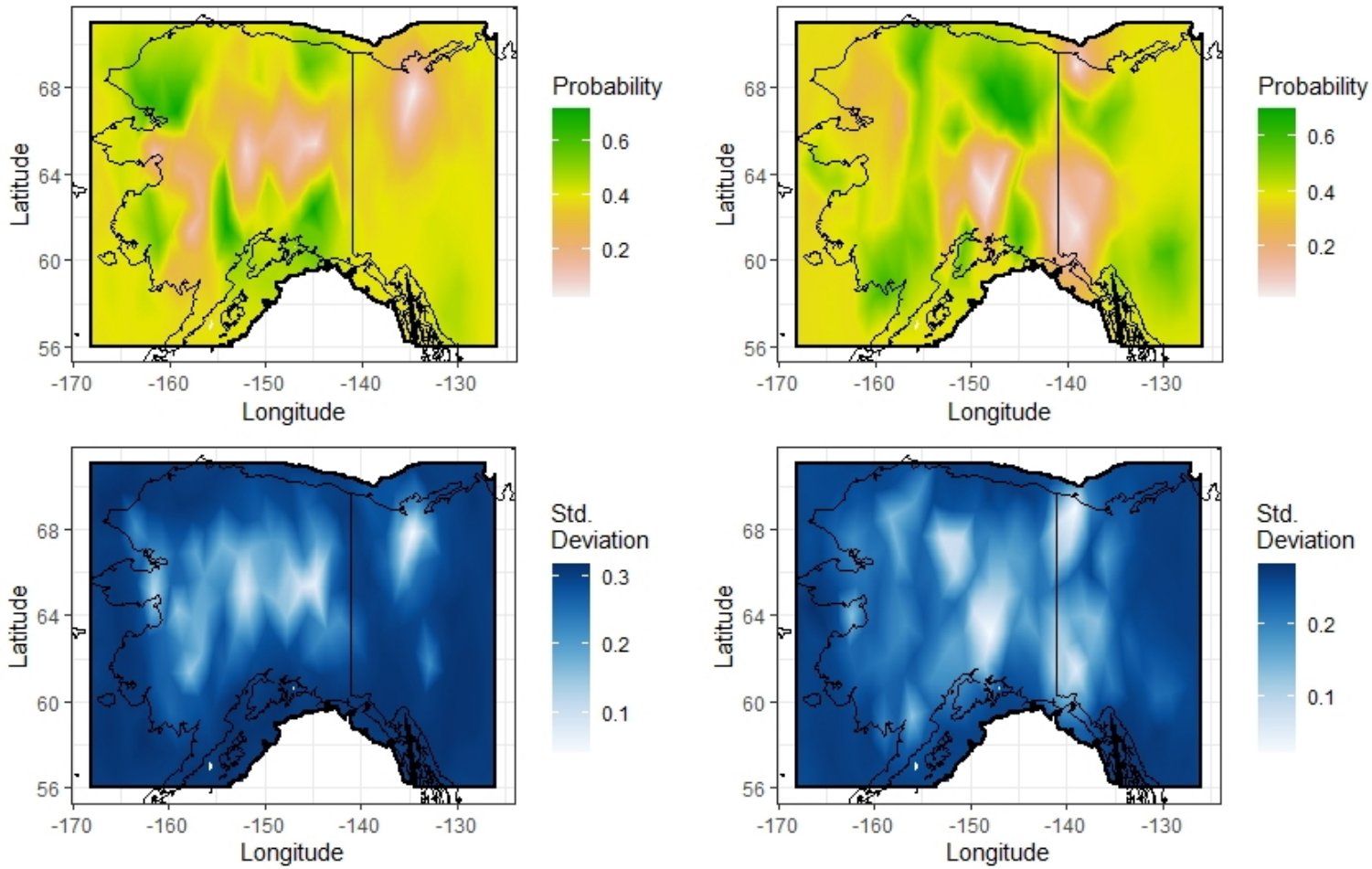
(a) S2



(b) S3

Results

Results



(a) **(b)**

Figure: Posterior mean and standard deviation of $P(\boldsymbol{s})$ of (a) green alder and (b) white spruce.

Conclusion

- We propose an innovative Bayesian hierarchical model to utilize the diverse pieces of evidence collected in paleoecology to uncover cryptic refugia.
- The simplicity and flexibility of the model plus the computational convenience offered by INLA allow for researchers to quickly and efficiently implement the model with their data sources.
- The regions highlighted in our results can be used to inform future field expeditions for the further study of cryptic refugia in Eastern Beringia.

Estimation of Solar-Induced Chlorophyll Fluorescence Yield of Various Vegetation Land Types Using a Spatially Varying Coefficient Model

In collaboration with Bo Li and Kaiyu Guan

Mauricio Campos

UIUC

June 30, 2023

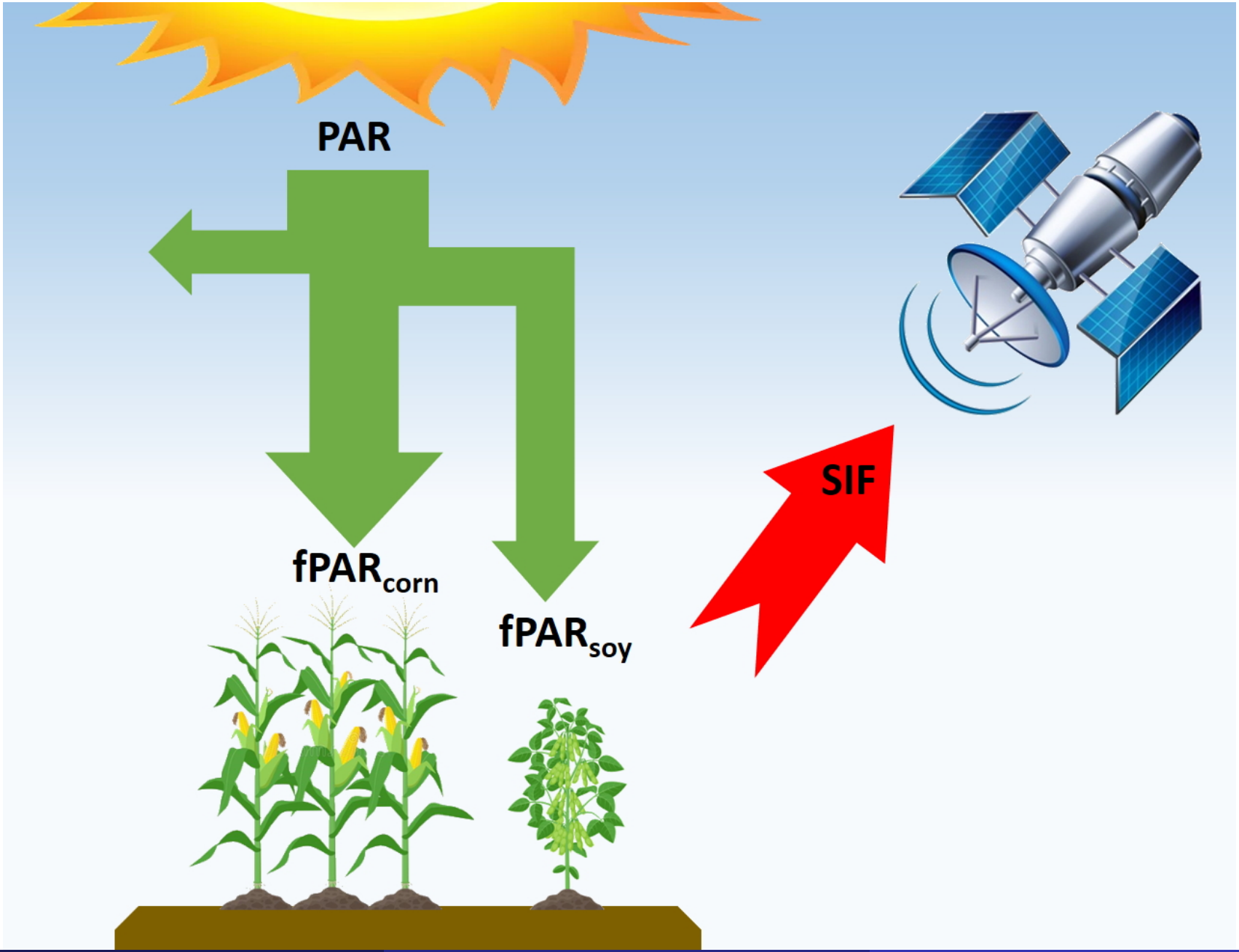
Outline

Motivation

Motivation

- Gross Primary Productivity (GPP) is defined as the amount of carbon assimilated by plants through photosynthesis. If quantified correctly it can vastly aid our understanding of ecosystem functions as well as the impact of climate change to crop yields.
- Recent advances in satellite spectroscopy have made it possible to estimate photosynthesis with the flux of chlorophyll fluorescence emitted by the canopy (Frankenberg et al., 2011; Guanter et al., 2007; Joiner et al., 2011; Meroni et al., 2009)
- Guan et al. (2015) recently used space-borne measurements to develop an approach of using Solar Induced Fluorescence (SIF) as a general tool for crop yield monitoring.

Solar Induced Fluorescence



Decomposing SIF

- SIF can be expressed as follows (Li et al., 2018; Yoshida et al., 2015):

$$\text{SIF} = \sum_{c} f\text{PAR}_c \times \text{PAR} \times \text{SIF}_{\text{yield } c}$$

  - PAR: photosynthetically active radiation (i.e. light that can be used for photosynthesis)
  - fPAR: fraction of PAR absorbed by vegetation canopy
  - $\text{SIF}_{\text{yield}}$: emitted SIF per photon absorbed
  - The sub-index c stands for different land types.

Problem

- Each site $\boldsymbol{s}$ consists of a satellite image that contains information of several land types. The proportion of pixels pertaining to each land type ($\pi_c(\boldsymbol{s})$) can be included in the following way:

$$\text{SIF}(\boldsymbol{s}) = \sum_c \pi_c(\boldsymbol{s}) \times f\text{PAR}_c(\boldsymbol{s}) \times \text{PAR}(\boldsymbol{s}) \times \text{SIF}_{\text{yield } c}(\boldsymbol{s})$$

Problem

- Each site $\boldsymbol{s}$ consists of a satellite image that contains information of several land types. The proportion of pixels pertaining to each land type ($\pi_c(\boldsymbol{s})$) can be included in the following way:

$$\text{SIF}(\boldsymbol{s}) = \sum_c \overbrace{\pi_c(\boldsymbol{s}) \times f\text{PAR}_c(\boldsymbol{s}) \times \text{PAR}(\boldsymbol{s})}^{\text{known!}} \times \text{SIF}_{\text{yield } c}(\boldsymbol{s})$$

Problem

- Each site $\boldsymbol{s}$ consists of a satellite image that contains information of several land types. The proportion of pixels pertaining to each land type ($\pi_c(\boldsymbol{s})$) can be included in the following way:

$$\text{SIF}(\boldsymbol{s}) = \sum_c \overbrace{\pi_c(\boldsymbol{s}) \times f\text{PAR}_c(\boldsymbol{s}) \times \text{PAR}(\boldsymbol{s})}^{\text{known!}} \times \overbrace{\text{SIF}_{\text{yield } c}(\boldsymbol{s})}^{\text{unknown}}$$

Problem

- Each site $\boldsymbol{s}$ consists of a satellite image that contains information of several land types. The proportion of pixels pertaining to each land type ($\pi_c(\boldsymbol{s})$) can be included in the following way:

$$\text{SIF}(\boldsymbol{s}) = \sum_c \overbrace{\pi_c(\boldsymbol{s}) \times f\text{PAR}_c(\boldsymbol{s}) \times \text{PAR}(\boldsymbol{s})}^{\text{known!}} \times \overbrace{\text{SIF}_{\text{yield } c}(\boldsymbol{s})}^{\text{unknown}}$$

- Notice we can interpret this problem as a linear regression with different coefficients for each site:

$$y(\boldsymbol{s}) = \sum_{k=1}^{p} x_k(\boldsymbol{s})\beta_k(\boldsymbol{s})$$

  - $y(\boldsymbol{s}) = \text{SIF}(\boldsymbol{s})$
  - $x_k(\boldsymbol{s}) = \pi_c(\boldsymbol{s}) \times f\text{PAR}_c(\boldsymbol{s}) \times \text{PAR}(\boldsymbol{s})$
  - $\beta_k(\boldsymbol{s}) = \text{SIF}_{\text{yield } c}(\boldsymbol{s}) \geq 0.$

Model

Spatially Varying Coefficient Model

- A model with p land types, whose coefficients vary with each of the n sites will have np coefficients to estimate.
- Preferably, for each land type, the coefficients in close proximity will be more similar.
- There are several techniques that addresses this problem:
  - Spatial Expansion Method (SEM) (Casetti, 1972, 1997)
  - Geographically Weighted Regression (GWR) (Brunsdon et al., 1996; Fotheringham et al., 2002)
  - Flexible-band GWR (Fotheringham et al., 2017)
  - Bayesian SVC (Gelfand et al., 2003; Finley and Banerjee, 2020)
  - Spatially Clustered Coefficients (Li and Sang, 2019)

Model

This is a regression model where each site $\boldsymbol{s}_i$ has it's own set of p coefficients, one for each covariate, that may vary from site to site:

$$y(\boldsymbol{s}_i) = \sum_{k=1}^{p} x_k(\boldsymbol{s}_i)\beta_k(\boldsymbol{s}_i) + \epsilon(\boldsymbol{s}_i), \qquad \epsilon(\boldsymbol{s}_i) \overset{iid}{\sim} N(0, \sigma^2)$$

subject to

$$\sum_{k=1}^{p} \sum_{i,j \in E} \mathcal{P}(\beta_k(\boldsymbol{s}_i) - \beta_k(\boldsymbol{s}_j)) < \tau \qquad \text{and} \quad \boldsymbol{\beta} \geq \mathbf{0}$$

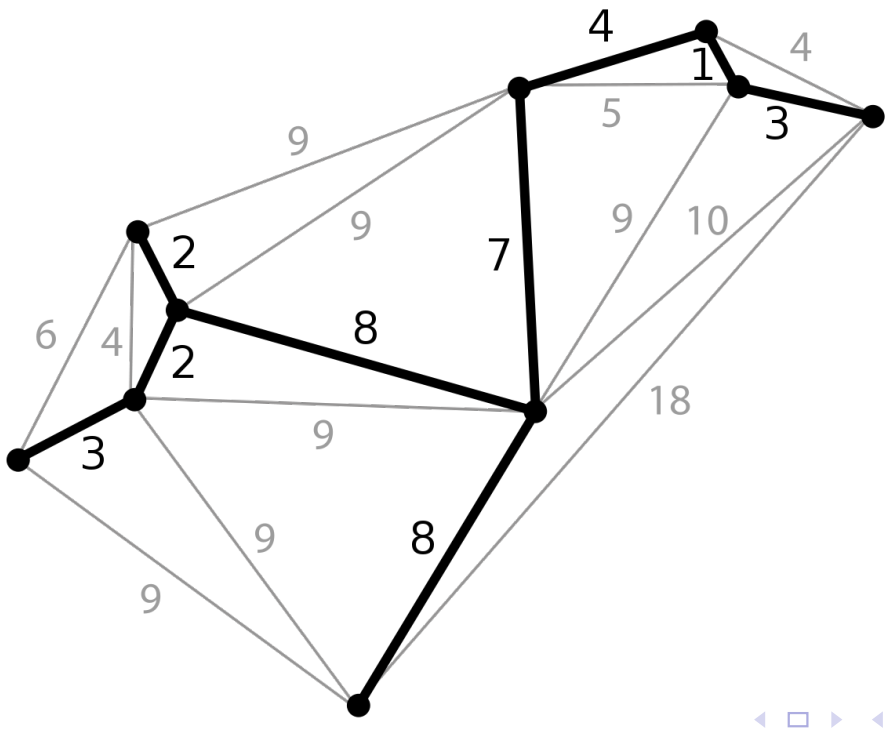
where E is an edge set belonging to a graph of the sites, $\mathcal{P}(\cdot)$ is a penalty function and τ is a threshold that determines spatial closeness.

Penalty, Edge Set and Threshold selection

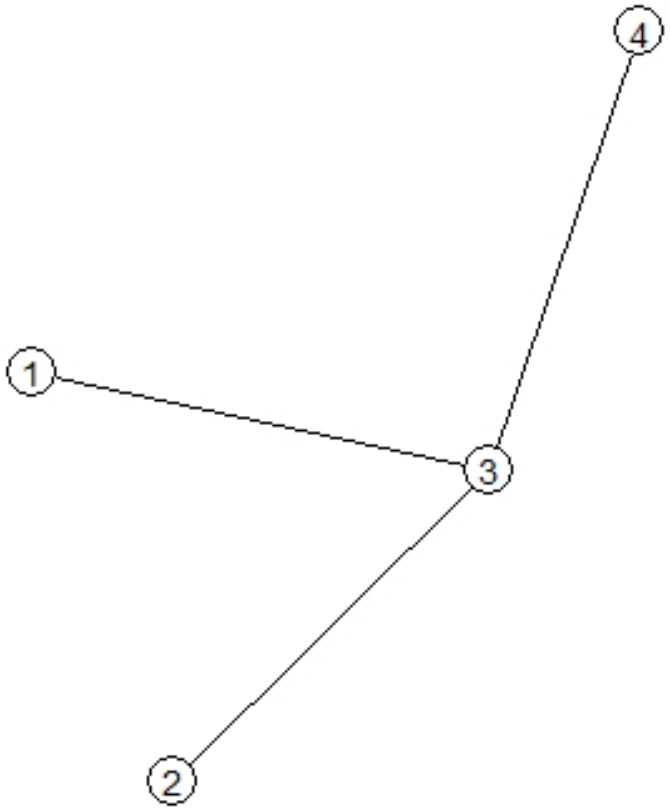
- Penalty:
  - Li and Sang (2019) use an L1 penalty loss and obtain clustered coefficients. We will use L2 instead.
- Threshold:
  - Using some information criteria method we can tune for this treshold.
  - Cross-validation is not an option!
- Edge set:
  - Following Li and Sang, we will use the edge set from the *minimum spanning tree* if we take the sites as vertices and their euclidean distance as edge weights.

Minimum Spanning Tree (MST)

- A *spanning tree* is an undirected, acylical subgraph that connects all vertices.
- An MST is a spanning tree that minimizes the total edge weight. It may not necessarily be unique.
- An MST with n vertices will have $n - 1$ edges.
- This will give us a computationally efficient edge set to apply the penalty function.



Example of the edge set matrix



$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}$$

$n = 4$

Estimation

Estimation

- Let us define:
  - $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 & \cdots & \boldsymbol{X}_p \end{bmatrix}$, where $\boldsymbol{X}_k = \text{diag}(\boldsymbol{x}_k)$
  - $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1^T & \boldsymbol{\beta}_2^T & \cdots & \boldsymbol{\beta}_p^T \end{pmatrix}^T$
- Using an L2 penalty, we can easily estimate an unconstrained $\boldsymbol{\beta}$ by minimizing the following objective function:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{k=1}^{p} \|\boldsymbol{A}\boldsymbol{\beta}_k\|_2^2 \right\}$$

- Closed-form solution for $\boldsymbol{\beta}$ given by:

$$\hat{\boldsymbol{\beta}} = \left[\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p \otimes (\boldsymbol{A}^T\boldsymbol{A}) \right]^{-1} \boldsymbol{X}^T\boldsymbol{y}$$

Constrained estimation

- Recall coefficients have to be non-negative, meaning:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}\geq\mathbf{0}} \left\{ \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{k=1}^{p} \|\boldsymbol{A}\boldsymbol{\beta}_k\|_2^2 \right\} \\ &= \arg\min_{\boldsymbol{\beta}\geq\mathbf{0}} \left\{ \frac{1}{2} \boldsymbol{\beta}^T \left[\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p \otimes (\boldsymbol{A}^T\boldsymbol{A}) \right] \boldsymbol{\beta} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} \right\}\end{aligned}$$

- This is a standard quadratic programming problem, for a given λ.
- We tune λ using AIC (corrected for small sample sizes).

Known coefficients

There are sites where $y(\boldsymbol{s}) = 0$. Coupled with the non-negativity constraint of the coefficients, this immediately implies $\beta_j(\boldsymbol{s}) = 0, \forall j = 1, ..., p$.

Data

Data

- Data consists of 6205 sites observed over the span of 112 days (Spring-Fall), all relating to the surrounding areas of Urbana-Champaign.
  - Model will be fitted to each day independently of the others.
- There are four different vegetation types of interest:
  - Corn
  - Soy
  - Grass
  - Forest

Daily sample sizes

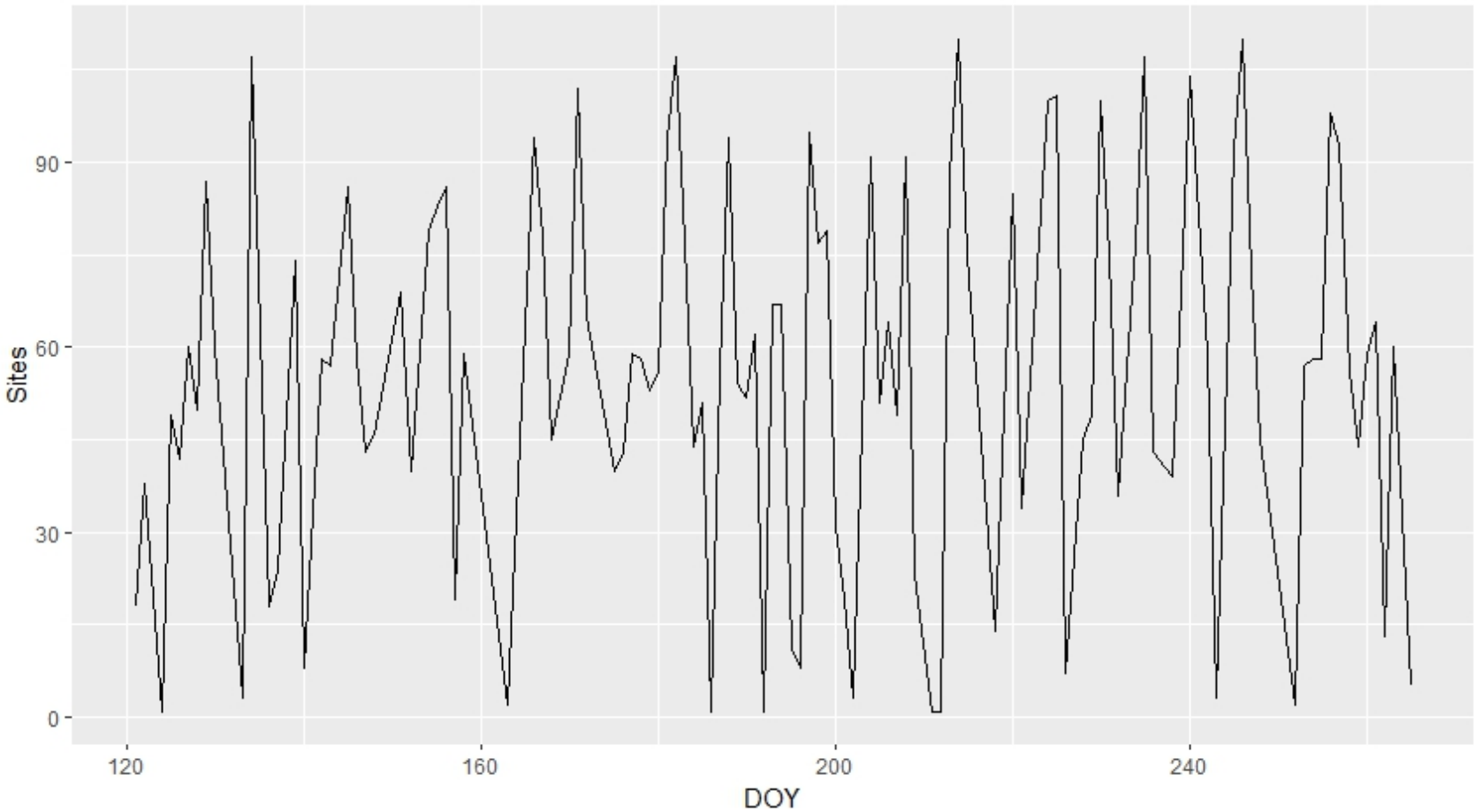


Figure: Number of sites/footprints per day of the year (DOY)

SIF and covariates

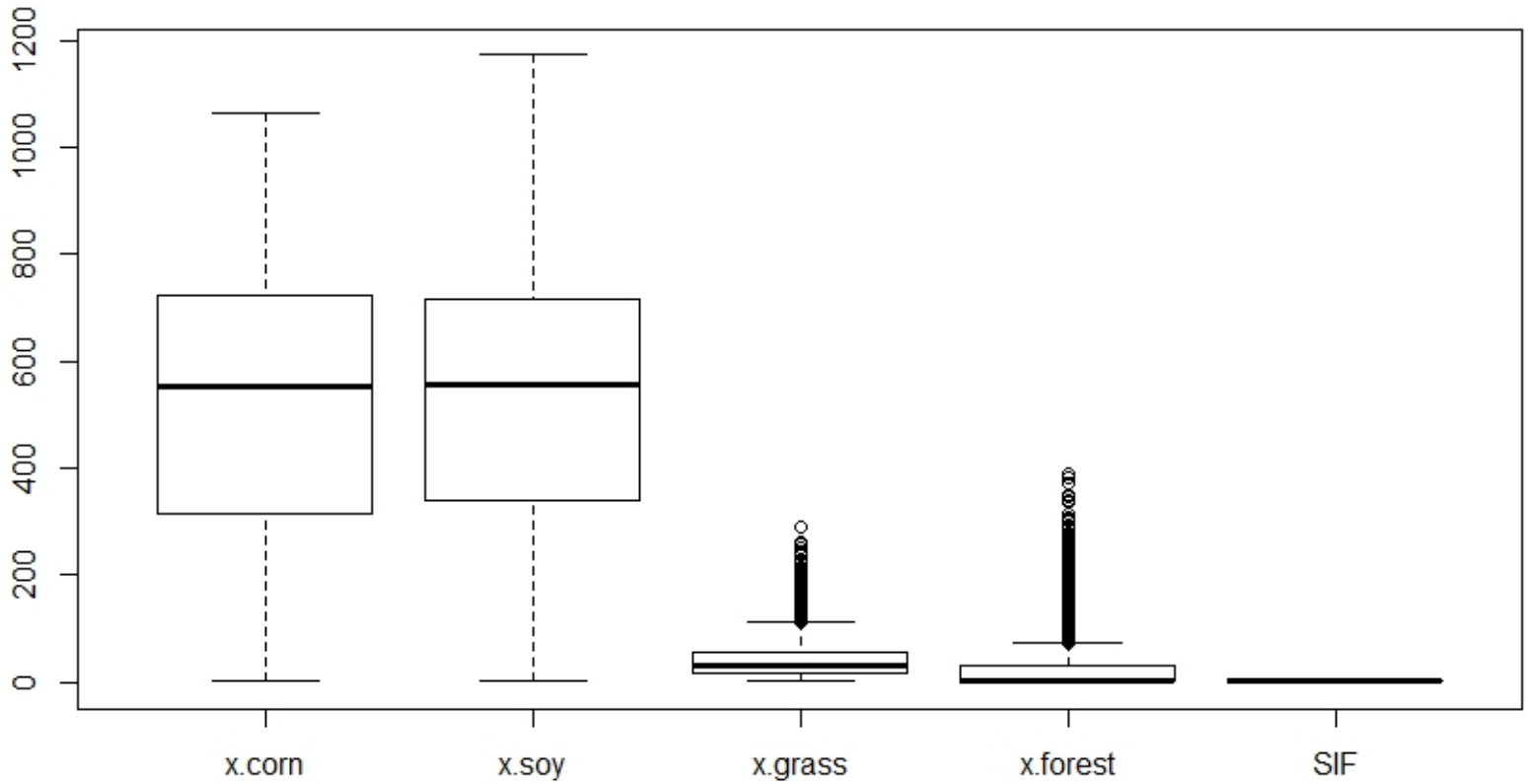


Figure: Boxplots for the observed SIF (y) and land-type-specific covariates

Correlations between SIF and covariates

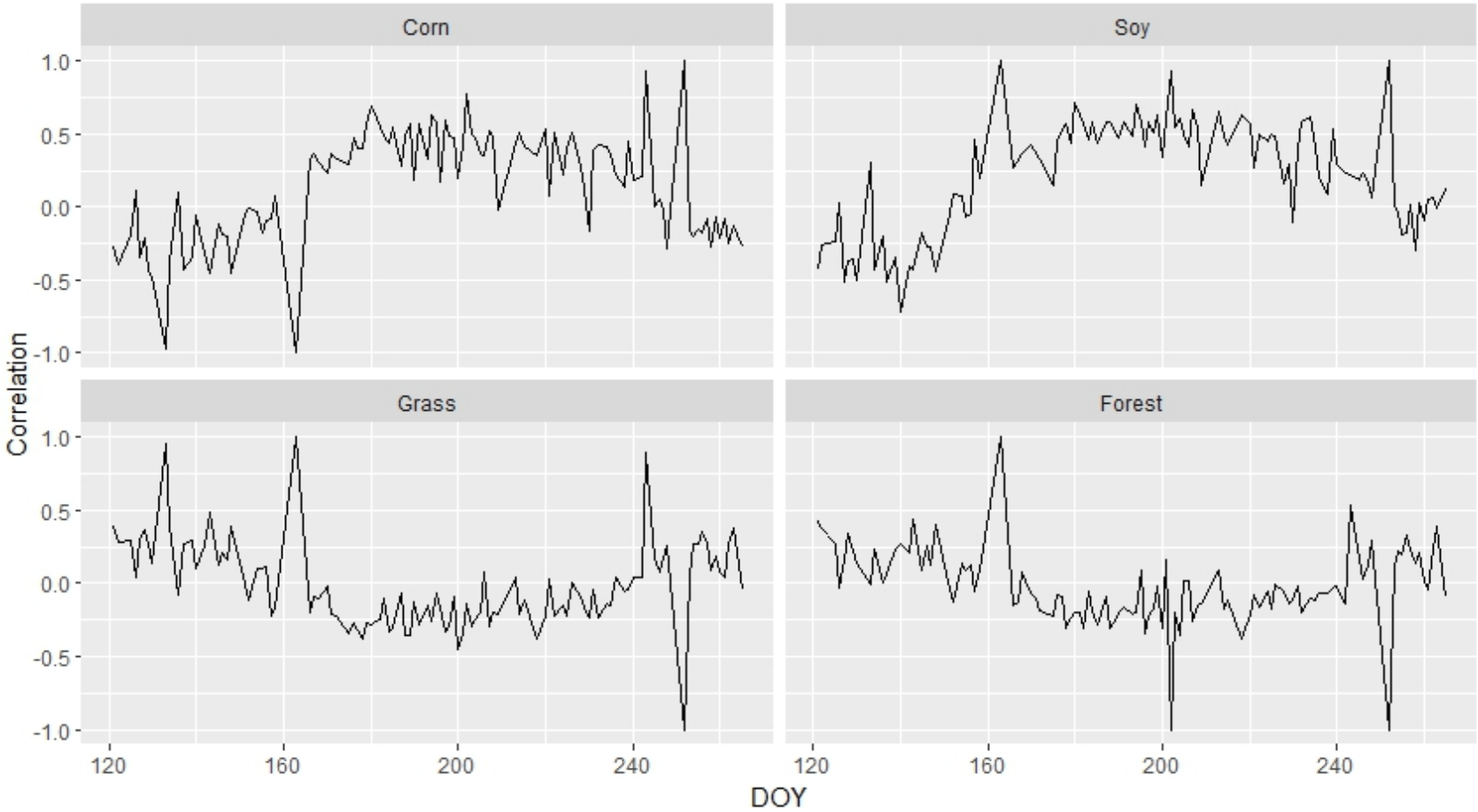
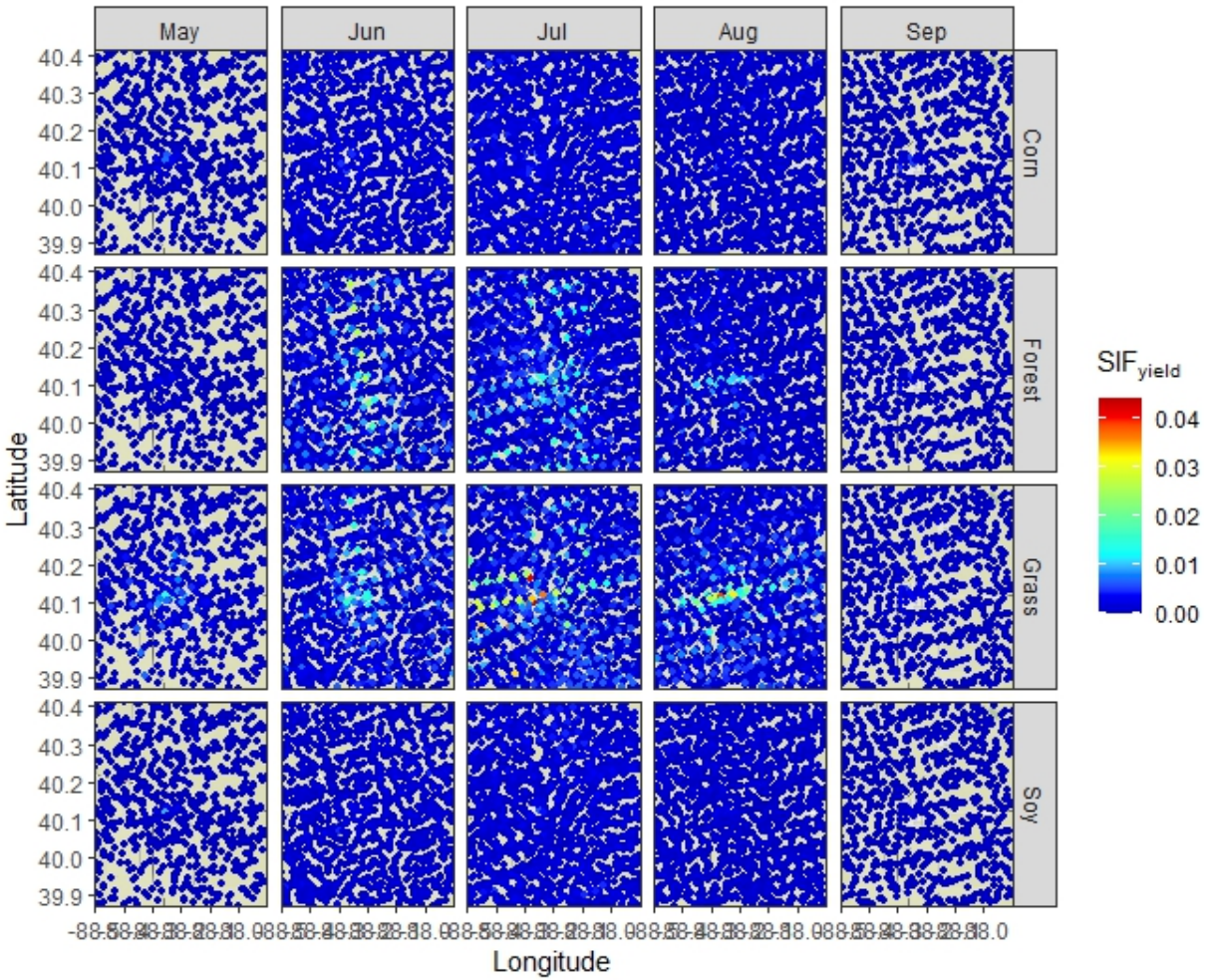


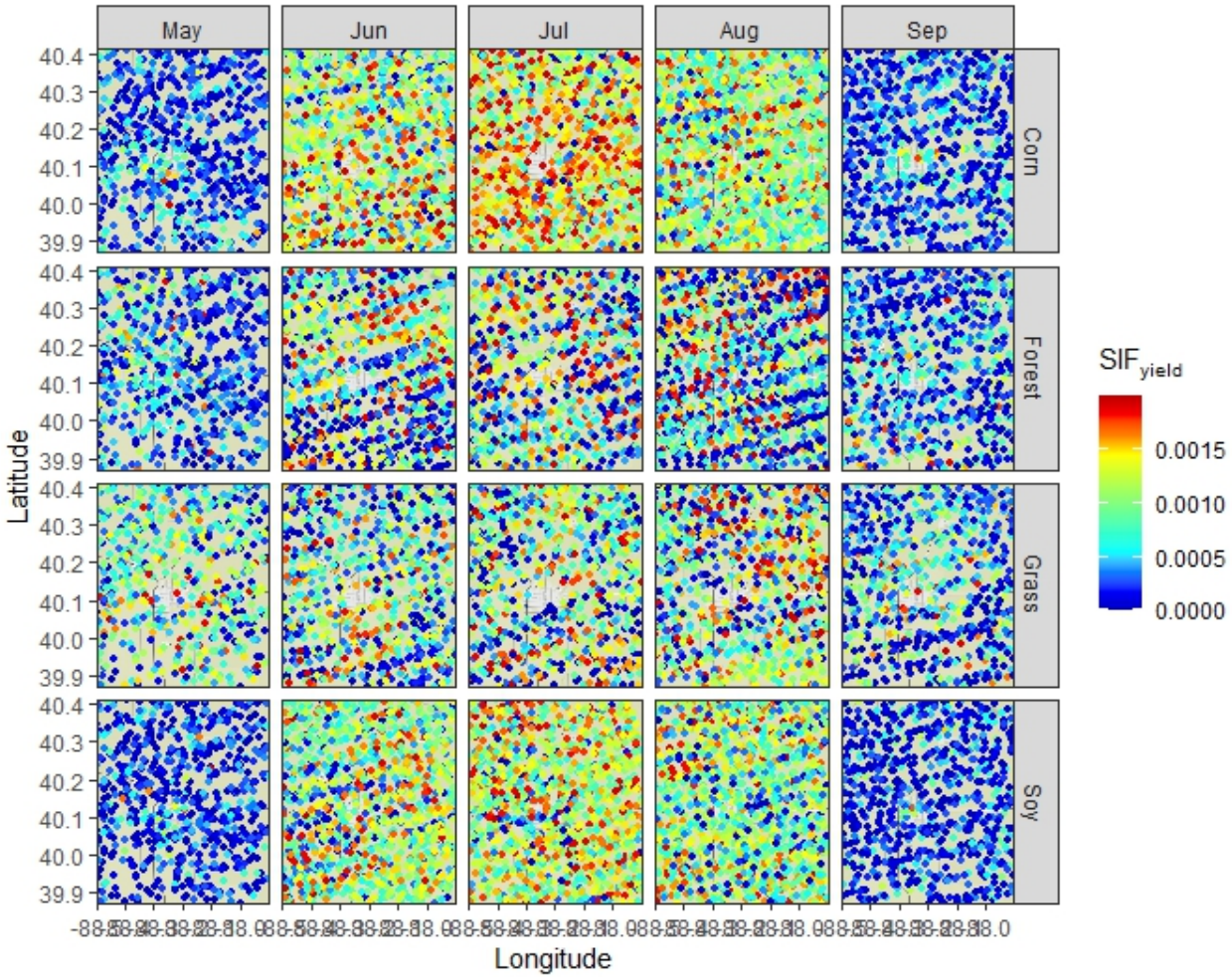
Figure: Daily correlations between the covariates and SIF for each different type of vegetation

Results

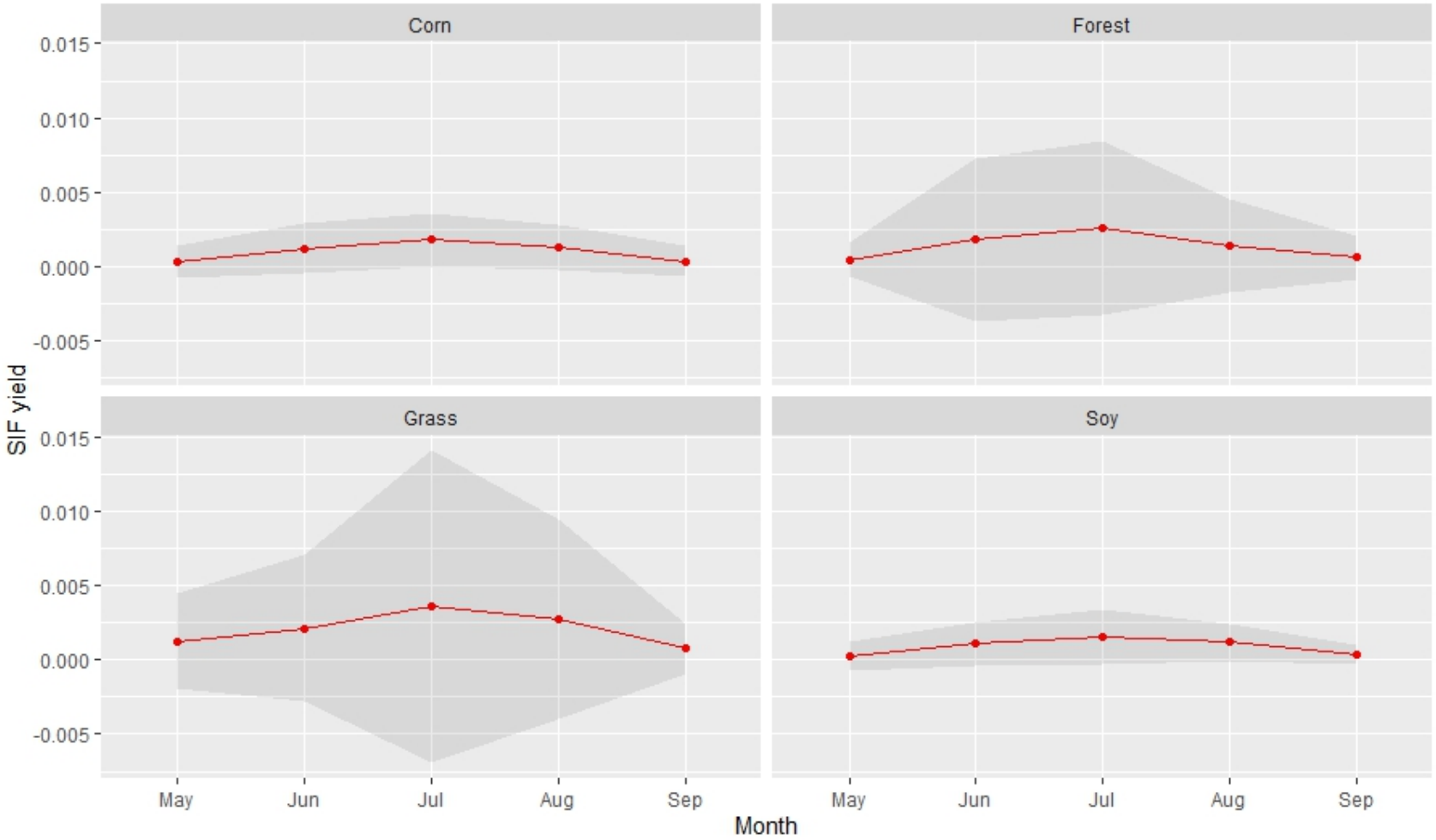
Estimated SIF_{yield} for each land type, per month



Estimated SIF_{yield} for each land type, per month - small values



Median SIF_{yield} for each land type, per month



Conclusions

- Our spatially varying coefficient model successfully estimates a smooth spatial field that faithfully resembles the SIF decomposition.
- Results are obtained in much less time when compared to the currently used methods to obtain SIF yield.
- Currently the spatial dependence is muddied by the lack of temporal modelling, so we could try to add some temporal dependence.

Data Fusion of Temperature Datasets Using INLA

In collaboration with Audrey McCombs, Justin Li, Gabriel Huerta and Lyndsay Shand

Mauricio Campos

UIUC

June 30, 2023

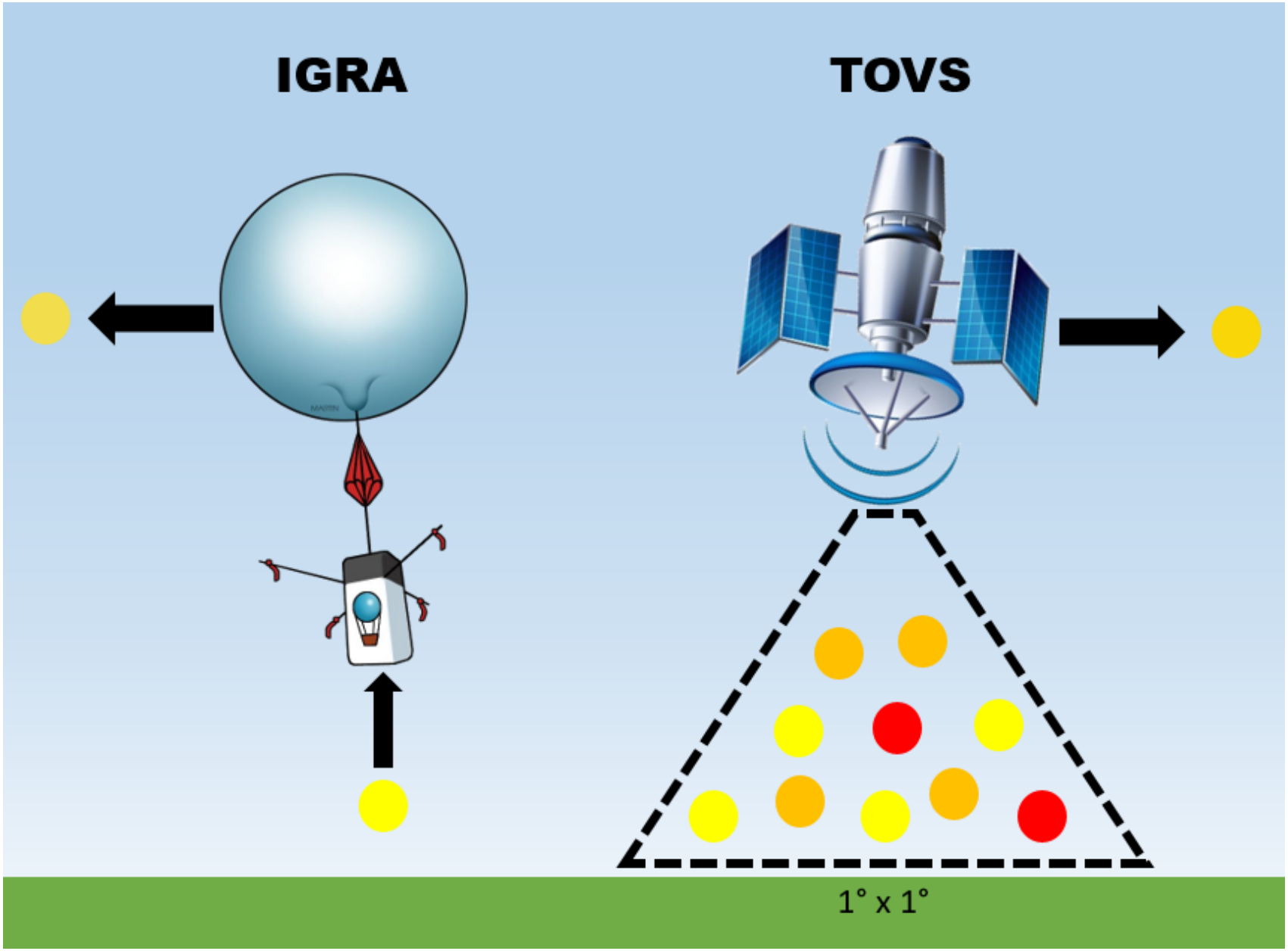
Outline

Motivation

Motivation

- To study the impact Mt. Pinatubo had in global temperature, it is imperative to have complete observed datasets for the years before and after the eruption.
- For these period of times, data can be obtained from the Integrated Global Radiosonde Archive (IGRA) and TIROS Operational Vertical Sounder (TOVS), but they come in different spatial resolutions and don't always cover the entire globe.
- We are met with interpolating global temperatures daily from 1990 to 1993 while dealing with the change of support problem.

Differing Spatial Resolution



Change of Support Problem

- The change of support problem is concerned with inference about the values of a variable at points or blocks different from those at which it has been observed.
  - Gotway and Young (2002) provide a comprehensive review of statistical methods for combining incompatible spatial data.
- Additionally, Gaussian Processes have become an indispensable tool for spatial analysis, but, in the 'big data' era, becomes computationally infeasible for modern spatial data. Modern methods exploit low-rank structures and/or multi-core and multi-threaded computing environments to facilitate computation (see Heaton et al., 2018).
- Moraga et al. (2017) present a Bayesian model for combining data obtained at point and areal resolutions using INLA and SPDE.

Data

Point Data

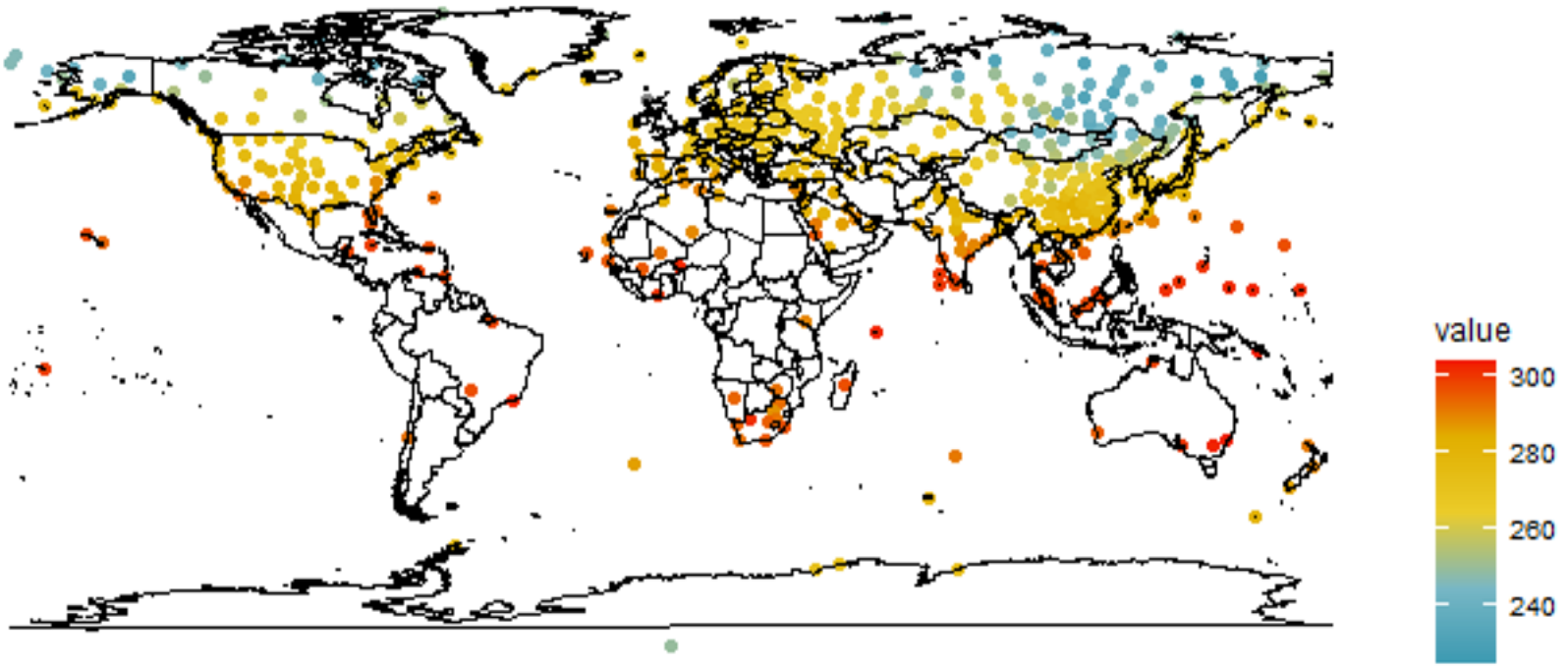


Figure: Point data from IGRA for January 1st, 1990

Areal Data

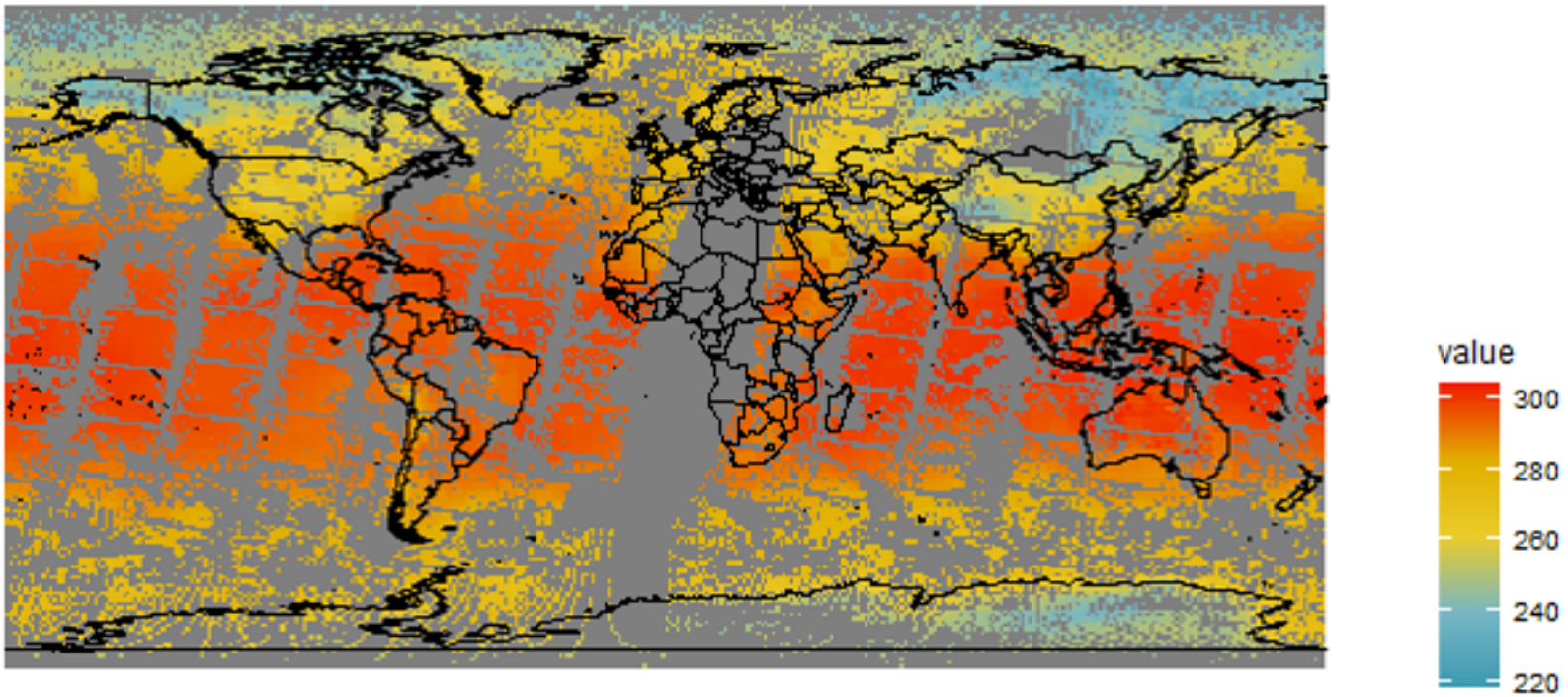


Figure: Areal data from TOVS for January 1st, 1990

Model

Gaussian Process

Let $S(\boldsymbol{x})$, $\boldsymbol{x} \in D \subset \mathbb{R}^2$, be a spatially continuous variable modeled using a Gaussian Process

$$S(\boldsymbol{x}) \sim GP(\mathbf{0}, \boldsymbol{\Sigma})$$

We observe point and areal data coming from the same process:

- Point data, $\boldsymbol{x}_i \in D$:

$$y(\boldsymbol{x}_i) = \boldsymbol{\mu}(\boldsymbol{x}_i) + S(\boldsymbol{x}_i) + \varepsilon(\boldsymbol{x}_i), \quad \varepsilon(\boldsymbol{x}_i) \sim N(0, \sigma^2)$$

- Areal data, $A_j \subset D$:

$$y(A_j) = \frac{1}{|A_j|} \int_{A_j} (\mu(\boldsymbol{x}) + S(\boldsymbol{x}))dx$$

COSP under INLA and SPDE

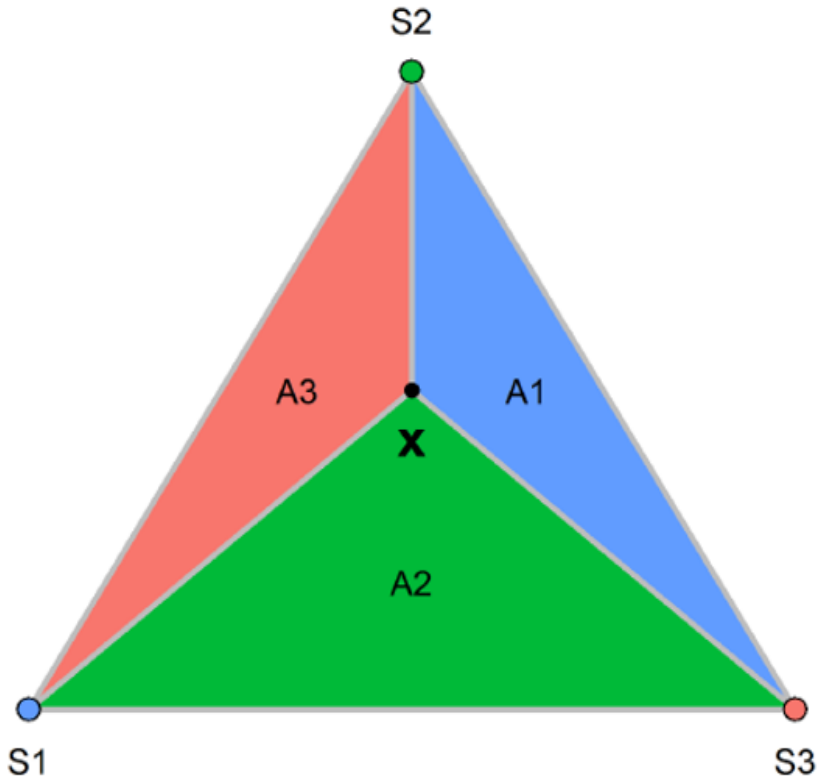
Recall that in the SPDE approach, the GMRF is fitted at each vertex of the mesh: $\{S_g\}, g = 1, 2, ..., G$

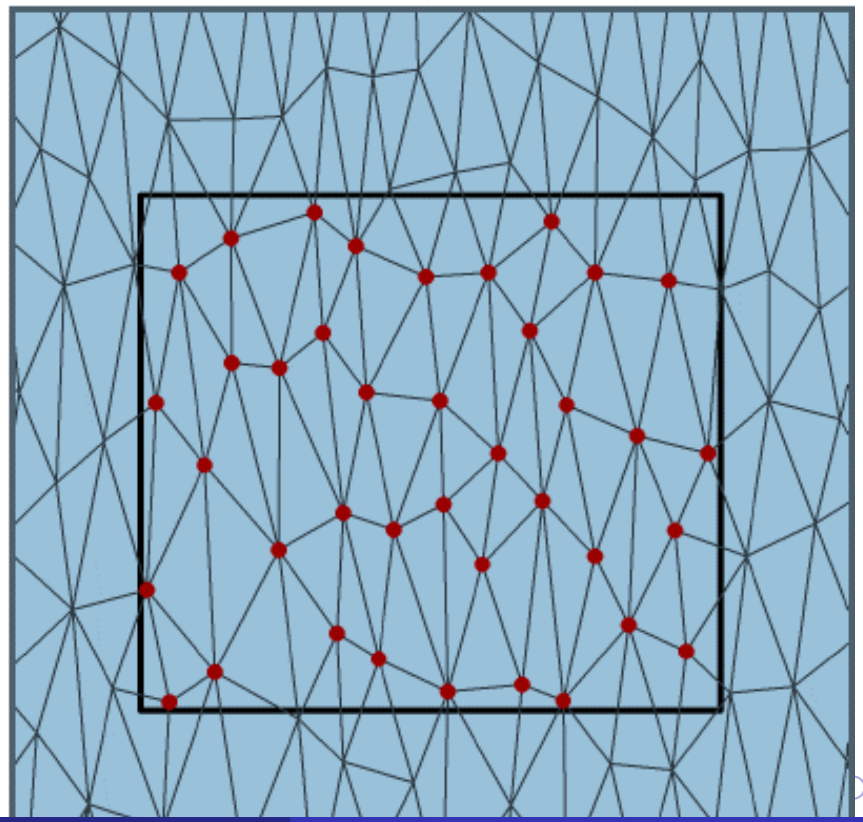
Point data:

$$S(\boldsymbol{x}) \approx \frac{A_1}{A} S_1 + \frac{A_2}{A} S_2 + \frac{A_3}{A} S_3$$

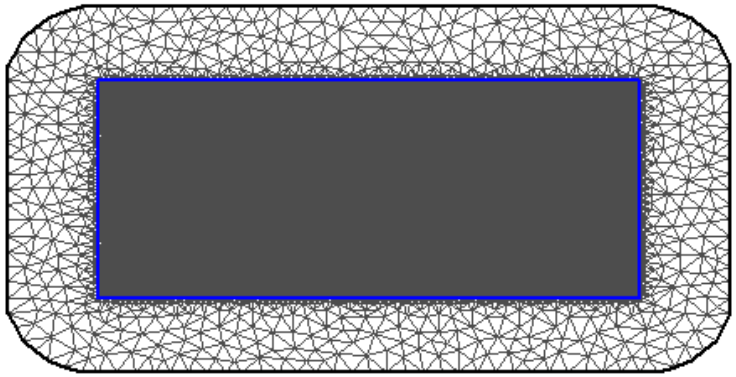
Areal data:

$$S(B) \approx \frac{1}{m} \sum_{g \in B} S_g$$

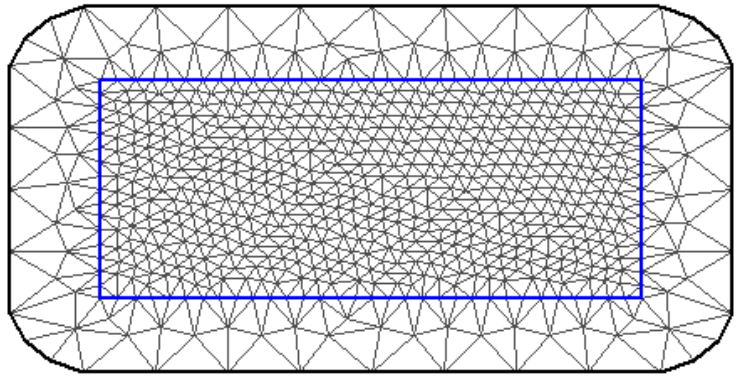




Differences in Mesh



(a)



(b)

Figure: Delaunay triangulation of the study region (inside the blue polygon) using a (a) dense mesh and a (b) sparse mesh. The dense mesh has 18066 vertices, whereas the sparse mesh consists of only 708, much less than the total sample size (16302).

Alternative Model

Under the same Gaussian Process $S(\boldsymbol{x})$, we now model areal data similar to point data but with a different error structure:

- Point data:

$$Y_p(\boldsymbol{x}) = \mu(\boldsymbol{x}) + S(\boldsymbol{x}) + \epsilon_p, \quad \epsilon_p \sim N(0, \sigma_p^2)$$

- Areal data:

$$Y_a(\boldsymbol{x}) = \mu(\boldsymbol{x}) + S(\boldsymbol{x}) + \epsilon_a, \quad \epsilon_a \sim N(0, \sigma_a^2)$$

Spatially Varying Variance

Similar to before, we will use the expanded class of covariance models given access to by the SPDE approach that allows for a spatially varying variance:

$$\log(\sigma(\boldsymbol{x})) = \theta_1 + \theta_2 \cdot \text{scale}(x_2)^2$$

where x_2 refers to latitude and the scale$(\cdot)$ function centers the latitude around it's mean and scales it to have a variance of 1.
The smoothness is fixed at $\nu = 1$ and range is left constant throughout space.

Priors

For simplicity, we will assume $\mu(\boldsymbol{x}) = \mu$. The priors used are as follows:

$$\begin{aligned} \mu &\sim N(0, 1000), \\ \log(1/\sigma_p^2), log(1/\sigma_a^2) &\sim \text{LogGamma}(1, 0.00005), \\ (\theta_1, \theta_2, \log \rho)^T &\sim N\left((0, 0, \log(142))^T, 100 \cdot \boldsymbol{I}_3\right). \end{aligned}$$

Simulation Study

Sampling Scenarios

The goal of the simulation study is to assess how estimations are affected by different sampling scenarios inspired by the real data:

Scenario	Areal Data	Point Data
1	All data	Sampled at random
2	All data	Sampled in land
3	Missing data at random	Sampled at random
4	Missing data in stripes	Sampled at random
5	Missing data in stripes	Sampled in land

Additionally, sample sizes are split two ways:

- 50-50: $n_a = n_p = 2048$
- 95-05: $n_a = 2048, n_p = 102$

Simulated data

The true field is simulated as:

$$S(\boldsymbol{x}) \sim GP(0, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma}$ follows a Matérn covariance function with $\sigma^2 = 1, \nu = 1, \rho = 142$.

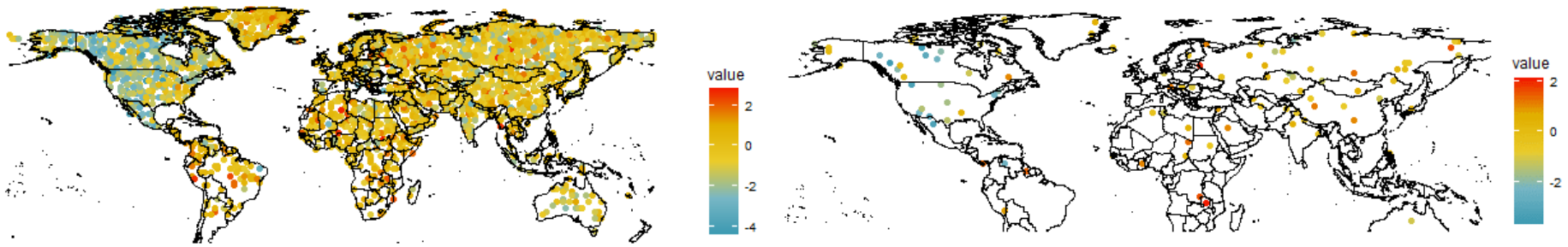
- Point data:

$$y(\boldsymbol{x}) = S(\boldsymbol{x}) + \varepsilon(\boldsymbol{x}), \quad \varepsilon(\boldsymbol{x}) \sim N(0, 1)$$

- Areal data:

$$y(A) = \frac{1}{K} \sum_{k \in A} S(\boldsymbol{x}_k)$$

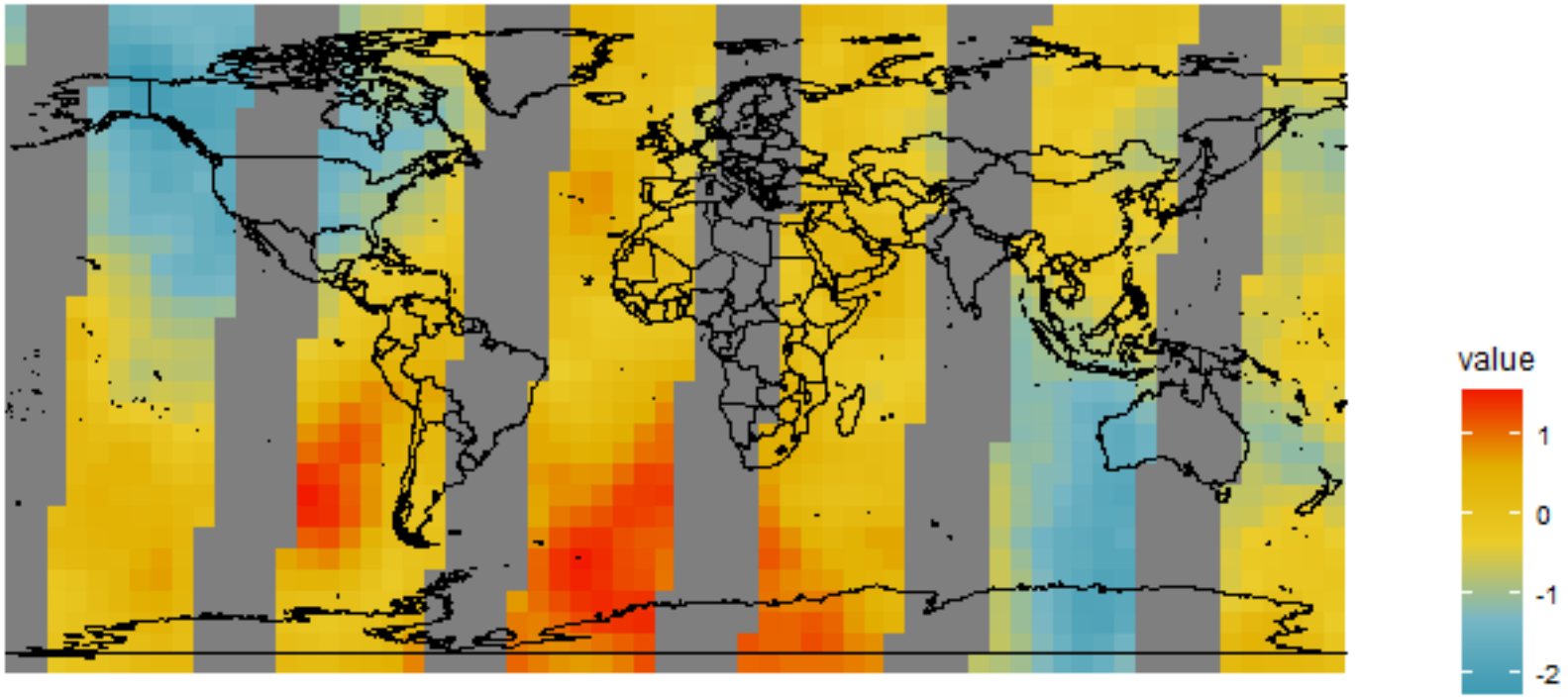
Example of simulated point data (sampled in land)



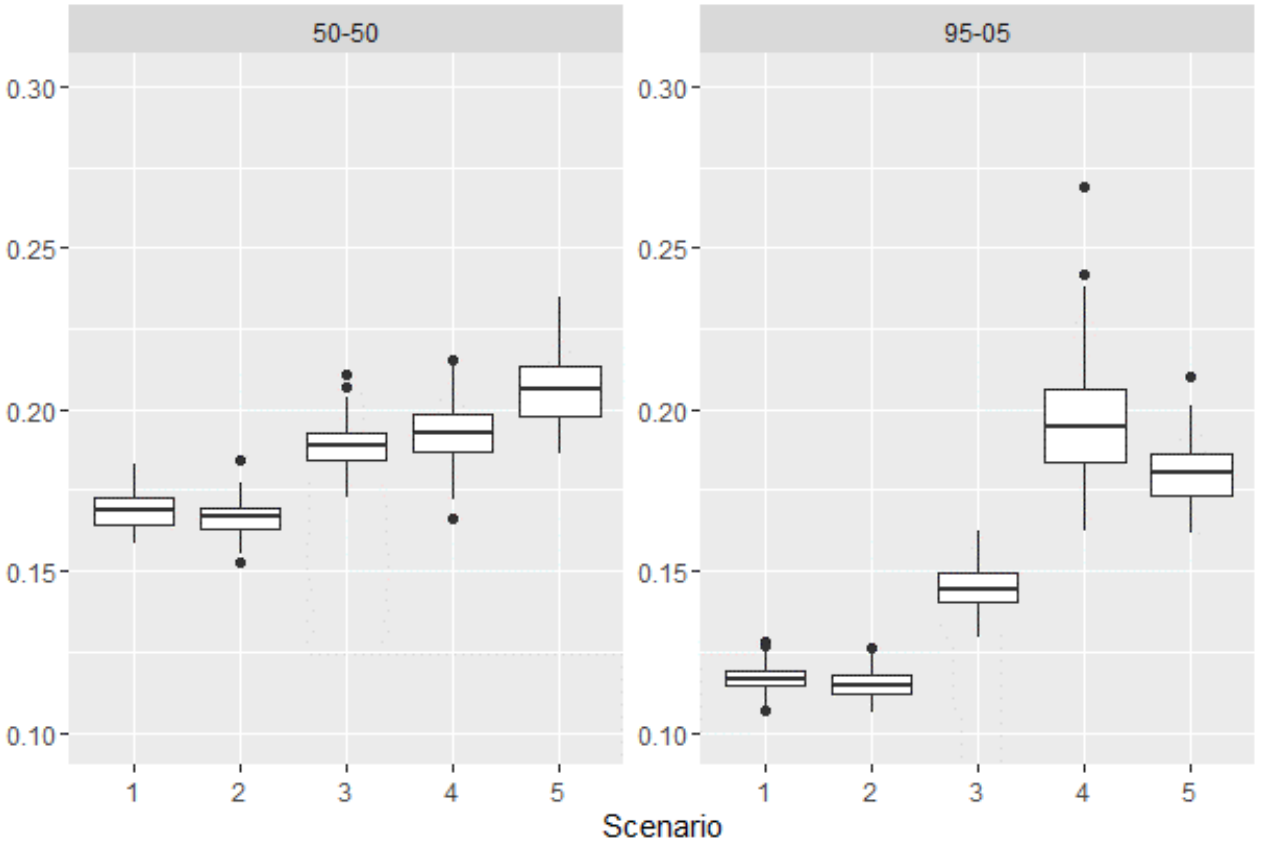
(a) **(b)**

Figure: (a) 50-50 split and (b) 95-05

Example of simulated areal data (Missing in stripes)

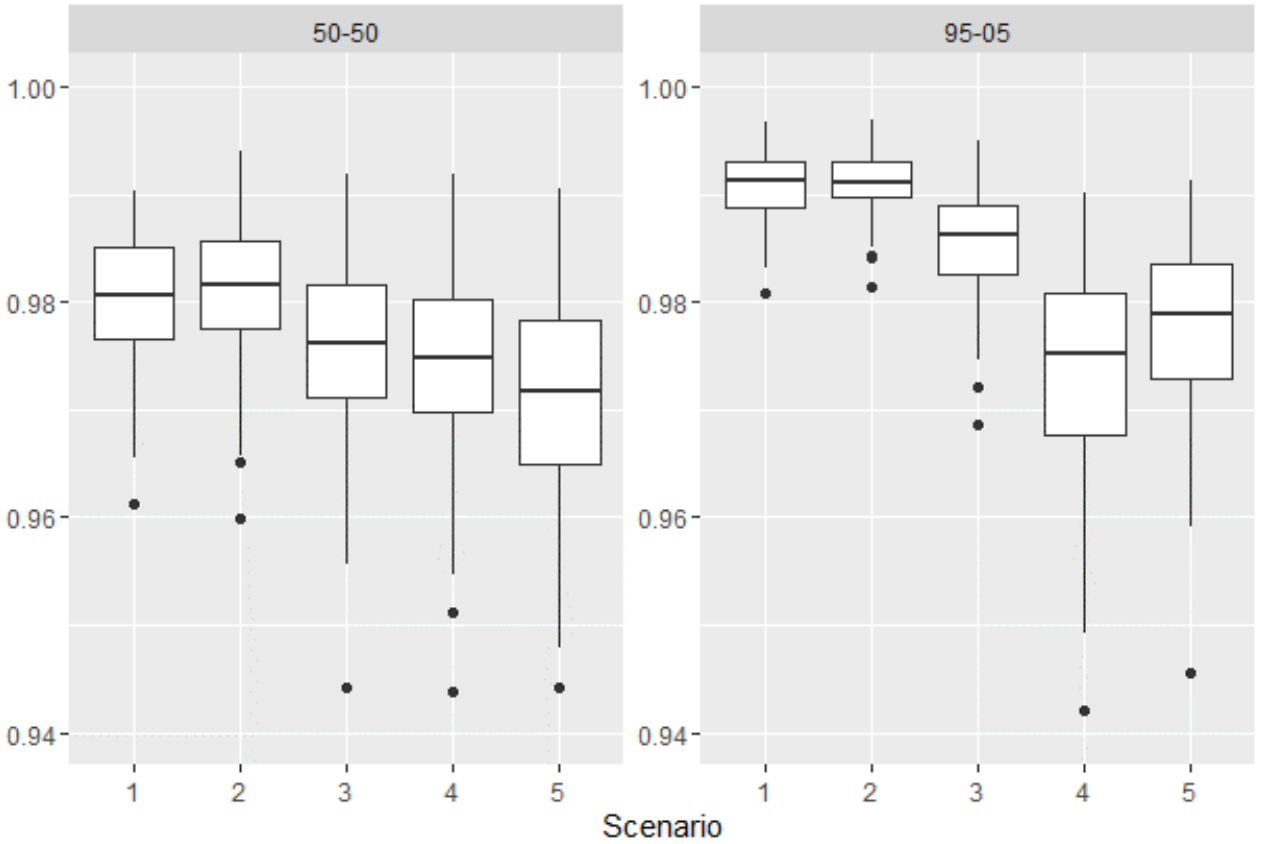


Results: Root Mean Square Error



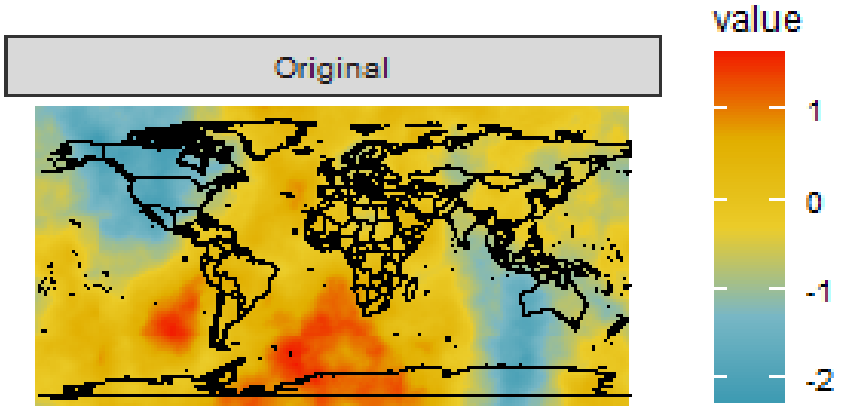
1. no missing areal, random points
2. no missing areal, in land points
3. areal missing at random, random points
4. areal missing in stripes, random points
5. areal missing in stripes, in land points

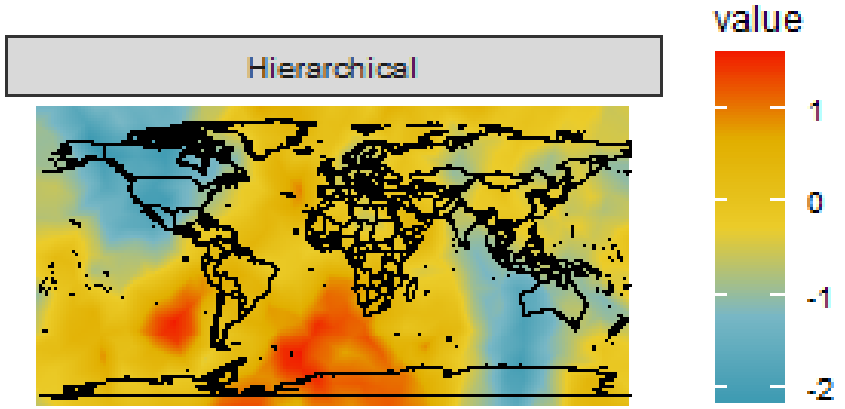
Results: Correlation

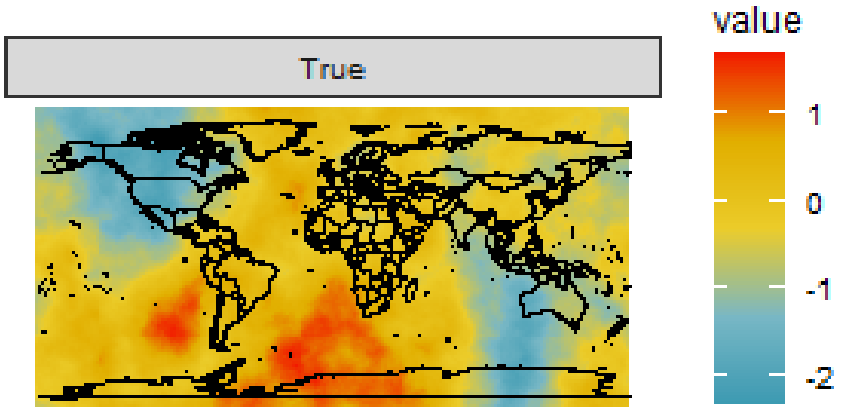


1. no missing areal, random points
2. no missing areal, in land points
3. areal missing at random, random points
4. areal missing in stripes, random points
5. areal missing in stripes, in land points

Difference in estimations between models

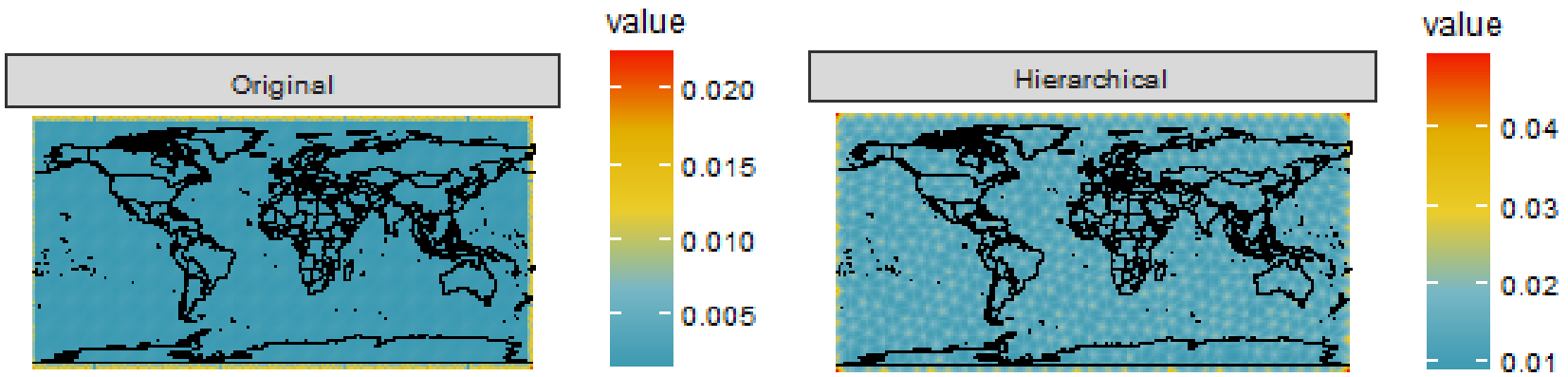






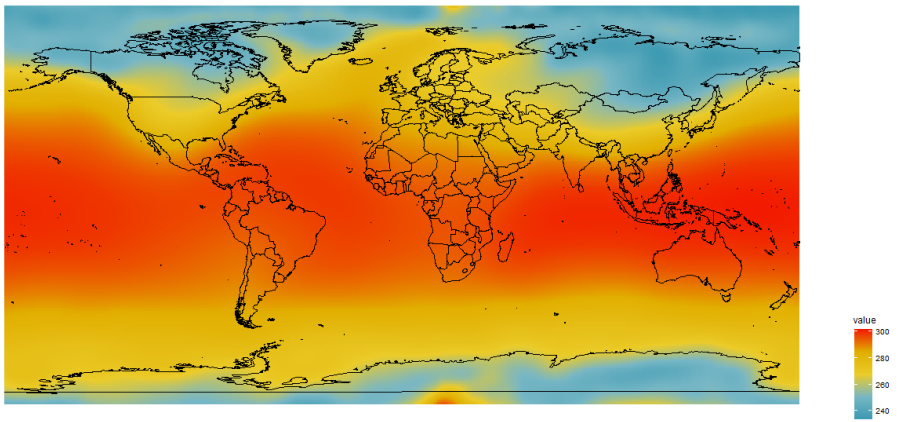
Prediction Error
Original: 0.0261
Hierarchical: 0.0759

Estimated uncertainty

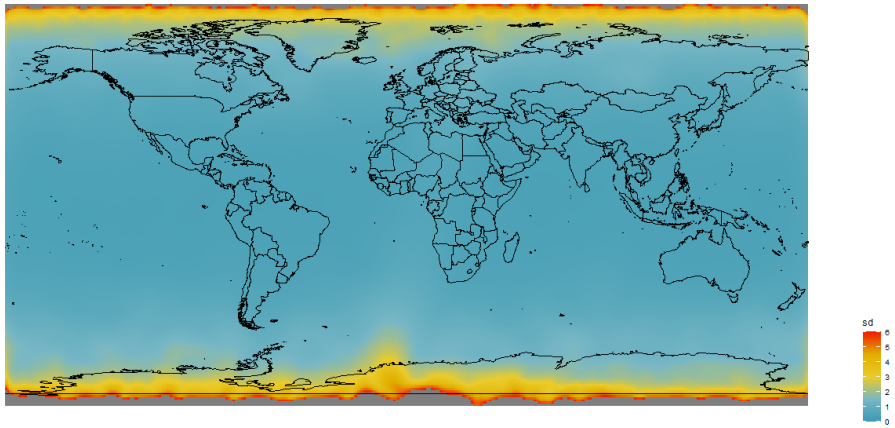


Results

Results from original model

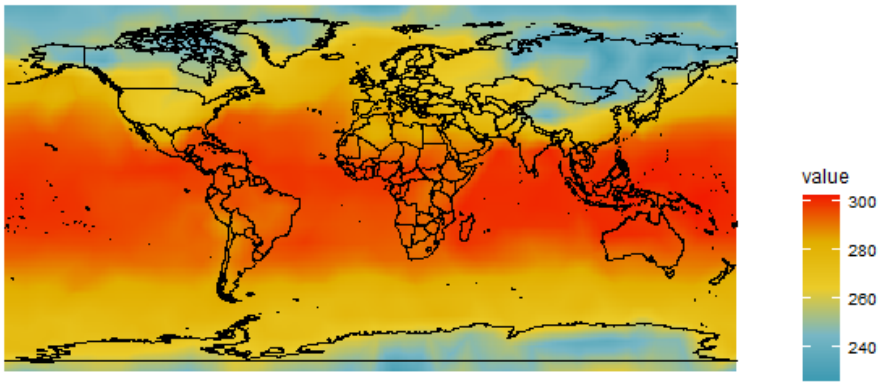


(a) Estimated field (posterior means)

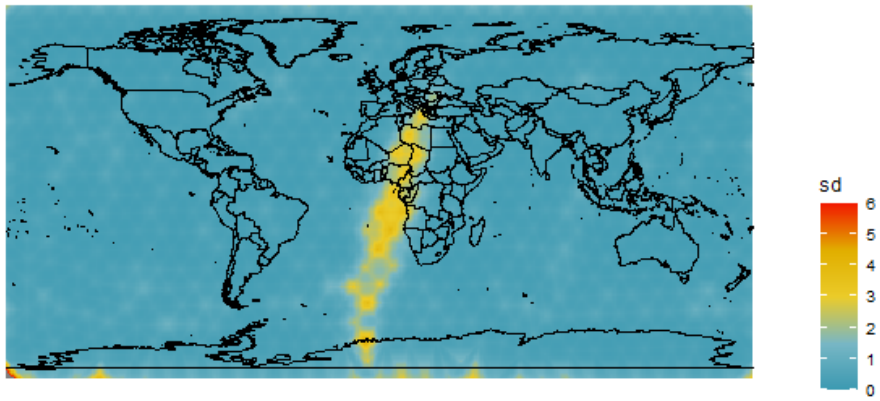


(b) Estimated uncertainty (posterior standard deviations)

Results from alternative model



(a) Estimated field (posterior means)



(b) Estimated uncertainty (posterior standard deviations)

Conclusions

- INLA is capable of accurately combining areal and point data through the SPDE approach. However, this is heavily dependent on the mesh.
  - Accuracy over runtime?
- Mesh is also constructed on a 2D surface as opposed to a sphere.
- Can implement to rest of the days, as well as other datasets of interest such as aerosol optical depth (AOD).
- This is part of a bigger project that is trying to find the best way to handle the change of support problem in terms of both accuracy and computation time when it comes to integrating observed datasets.

Thank you! Questions?