Spatial meshing and simplified manifold preconditioning for scaling models of multivariate multi-type spatial data

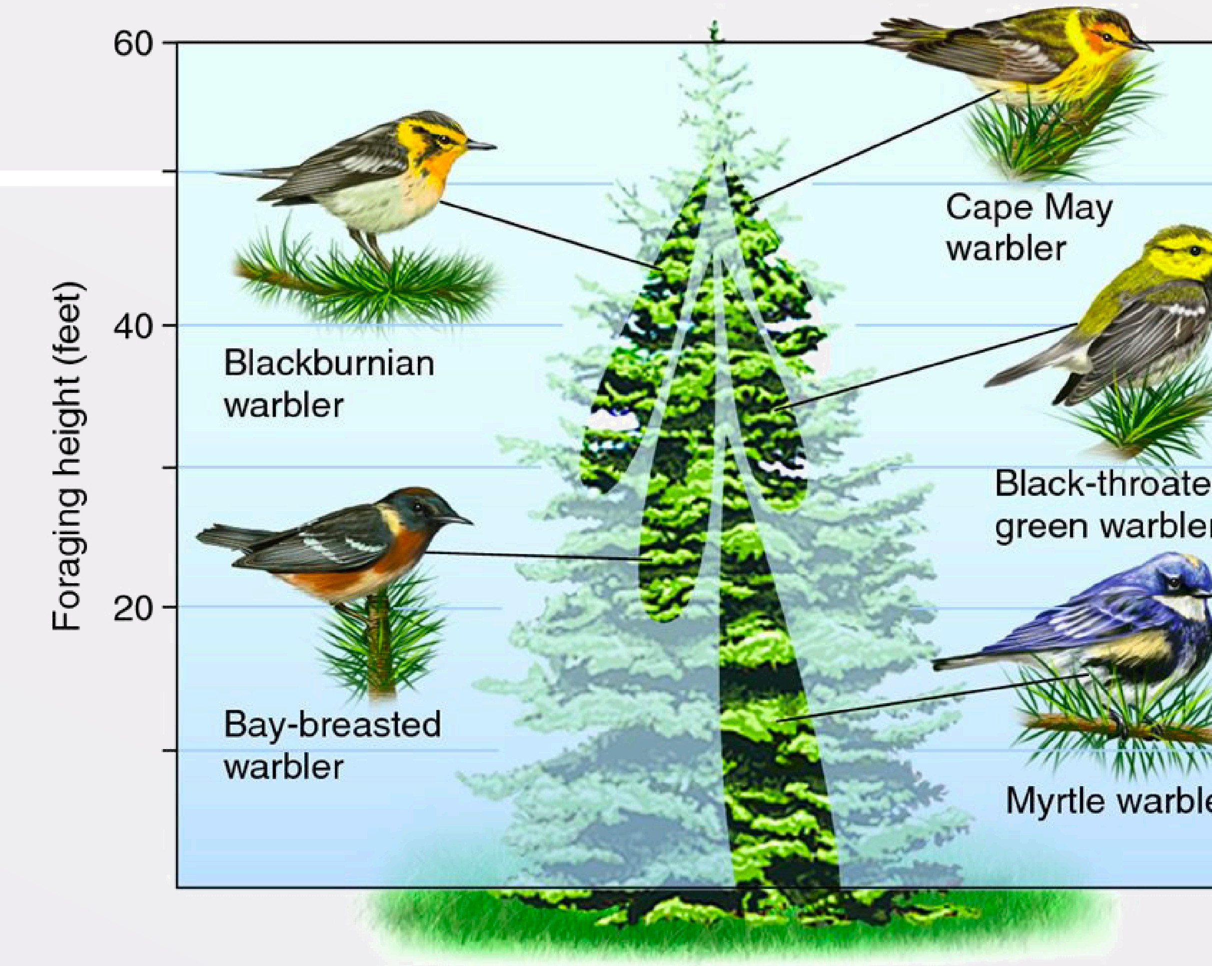
Michele Peruzzi
Assistant Professor (starting Fall 2023)
University of Michigan
Department of Biostatistics

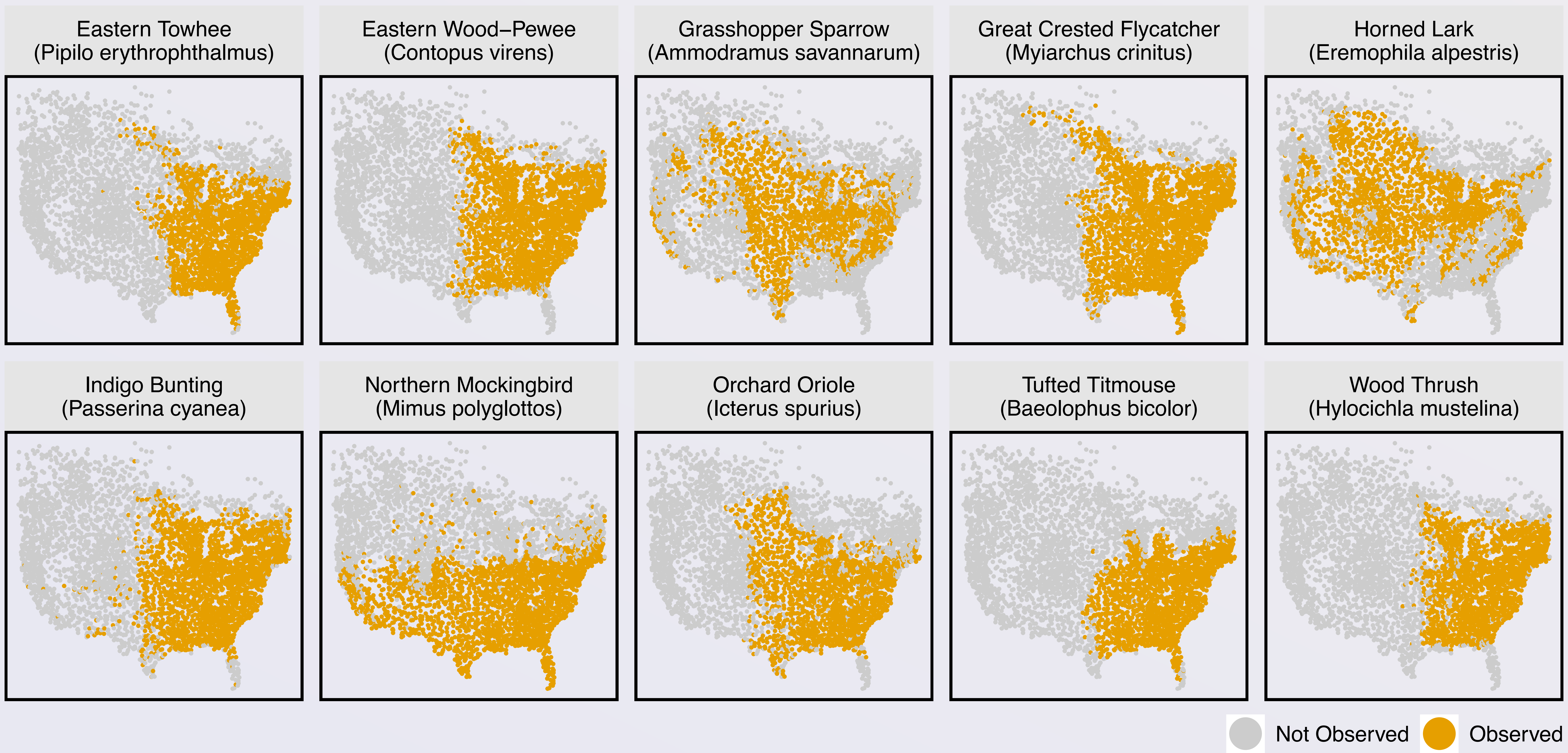
presenting joint work with:
David Dunson (Duke)
Sudipto Banerjee (UCLA)
Andrew Finley (MSU)

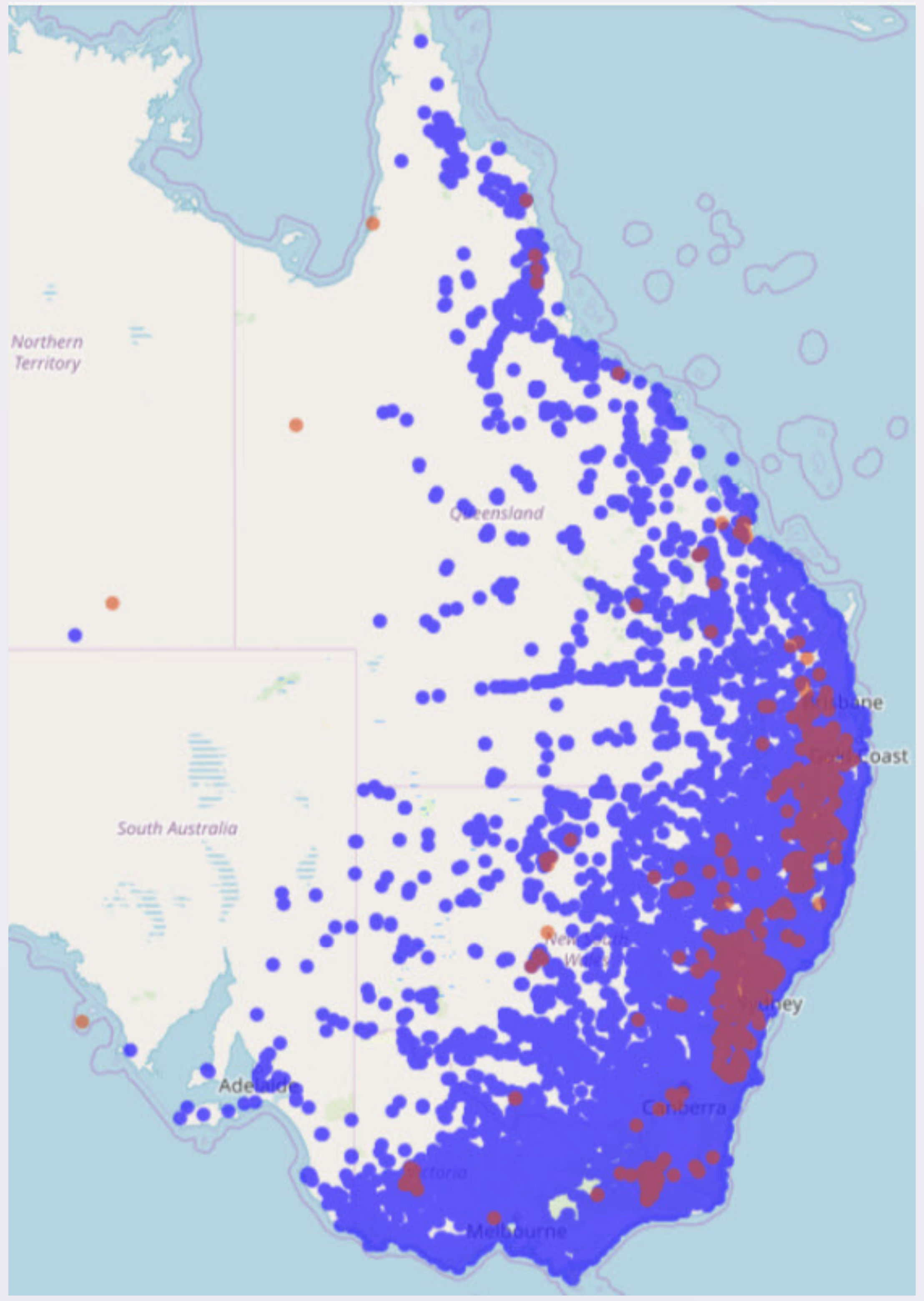
Large scale multivariate spatial and spatiotemporal data

Species' occurrence & Communities

- Spatial niches and shared resources
- Highly multivariate, non-Gaussian data
- Misalignment?







Multivariate spatial model with latent effects

Suppose we observe data at ***n*** locations (coordinates / pixels / voxels)

Probability model for outcome *j*

$$y_j(\boldsymbol{\ell}) \mid \eta_j(\boldsymbol{\ell}), \tau_j \sim P_j(\eta_j(\boldsymbol{\ell}), \tau_j)$$

$$\boldsymbol{\ell} \in \mathcal{D} \subset \Re^d \quad j = 1, \ldots, q$$

Linear predictor $\eta_j(\boldsymbol{\ell}) = \boldsymbol{x}_j(\boldsymbol{\ell})^\top \boldsymbol{\beta}_j + w_j(\boldsymbol{\ell})$

features or observed covariates

latent/random/unobserved effects for dependent data

example: Gaussian outcome $y_j(\boldsymbol{\ell}) = \eta_j(\boldsymbol{\ell}) + \varepsilon_j(\boldsymbol{\ell}) \quad \varepsilon_j(\boldsymbol{\ell}) \sim N(0, \tau_j^2)$

example: multivariate GLM with latent spatial factors

example: multi-species N-mixture models for large scale spatial abundance data

Multivariate spatial model with latent effects

Suppose we observe data at ***n*** locations (coordinates / pixels / voxels)

Probability model for outcome *j* $\quad y_j(\boldsymbol{\ell}) \mid \eta_j(\boldsymbol{\ell}), \tau_j \sim P_j(\eta_j(\boldsymbol{\ell}), \tau_j)$

$$\boldsymbol{\ell} \in \mathcal{D} \subset \Re^d \quad j = 1, \ldots, q$$

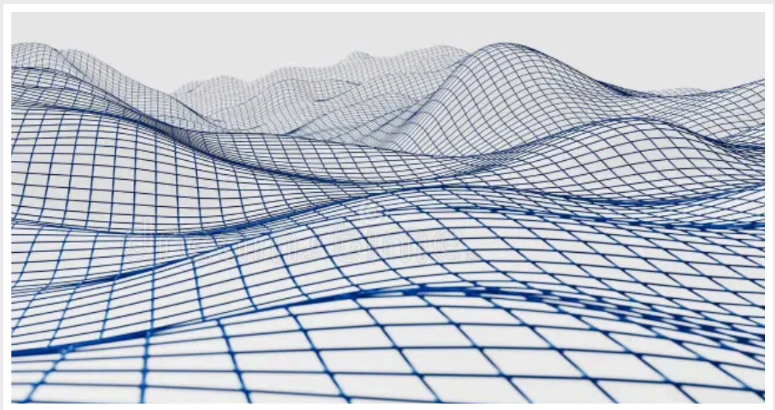
Linear predictor $\quad \eta_j(\boldsymbol{\ell}) = \boldsymbol{x}_j(\boldsymbol{\ell})^\top \boldsymbol{\beta}_j + w_j(\boldsymbol{\ell})$

$\boldsymbol{x}_j(\boldsymbol{\ell})^\top \boldsymbol{\beta}_j$: features or observed covariates

$w_j(\boldsymbol{\ell})$: latent/random/unobserved effects for dependent data

example: Gaussian outcome $\quad y_j(\boldsymbol{\ell}) = \eta_j(\boldsymbol{\ell}) + \varepsilon_j(\boldsymbol{\ell}) \quad \varepsilon_j(\boldsymbol{\ell}) \sim N(0, \tau_j^2)$

vector of random effects at location $\boldsymbol{\ell}$ $\quad \boldsymbol{w}(\boldsymbol{\ell}) = \begin{bmatrix} w_1(\boldsymbol{\ell}) \\ \vdots \\ w_q(\boldsymbol{\ell}) \end{bmatrix}$



Gaussian Process (GP) prior for $\boldsymbol{w}(\boldsymbol{\ell})$ at ***any*** set of locations $\boldsymbol{L} = \{\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_n\}$

$$\begin{cases} \boldsymbol{w}(\cdot) \sim GP(\mathbf{0}, \boldsymbol{C}_{\boldsymbol{\theta}}) \\ \boldsymbol{w}_L \sim N(\mathbf{0}, \boldsymbol{C}_L) \end{cases}$$

Scaling computations with Gaussian Processes

Gaussian Predictive Process & Inducing points

Quiñonero-Candela and Rasmussen, 2005; Snelson and Ghahramani, 2007; Banerjee et al. 2008; Banerjee et al. 2010; Guhaniyogi et al. 2011; Finley, Banerjee, and Gelfand 2012; Low et al., 2015; Ambikasaran et al., 2016; Huang and Sun, 2018; Geoga et al., 2020

Exploit data structure

Gilboa et al., 2015; Moran and Wheeler, 2020; Loper et al., 2020

Fixed Rank Kriging

Cressie and Johannesson 2008

Multi-resolution approximations

Gramacy and Lee 2008; Fox and Dunson 2012; Katzfuss 2017

Covariance Tapering

Furrer, Genton, and Nychka 2006; Kaufman, Schervish, and Nychka 2008; Bevilacqua et al., 2019

Independent partitioning

Sang and Huang 2012; Stein 2014

Composite likelihood

Bai et al., 2012; Eidsvik et al., 2014; Bevilacqua and Gaetan, 2015

Gaussian Random Markov Fields

Cressie 1993; Rue 2001; Rue and Held 2005

Vecchia's approximation & extensions

Vecchia 1988; Stein et al. 2014; Gramacy and Apley 2015; Datta et al. 2016; Guinness 2018; Heaton et al. 2019; Katzfuss and Guinness 2019; Quiroz et al., 2019; Schafer et al. 2021

Meshed Gaussian processes

Probability model for outcome j $\quad y_j(\boldsymbol{\ell}) \mid \eta_j(\boldsymbol{\ell}), \tau_j \sim P_j(\eta_j(\boldsymbol{\ell}), \tau_j)$

$$\boldsymbol{\ell} \in \mathcal{D} \subset \Re^d \quad j = 1, \dots, q$$

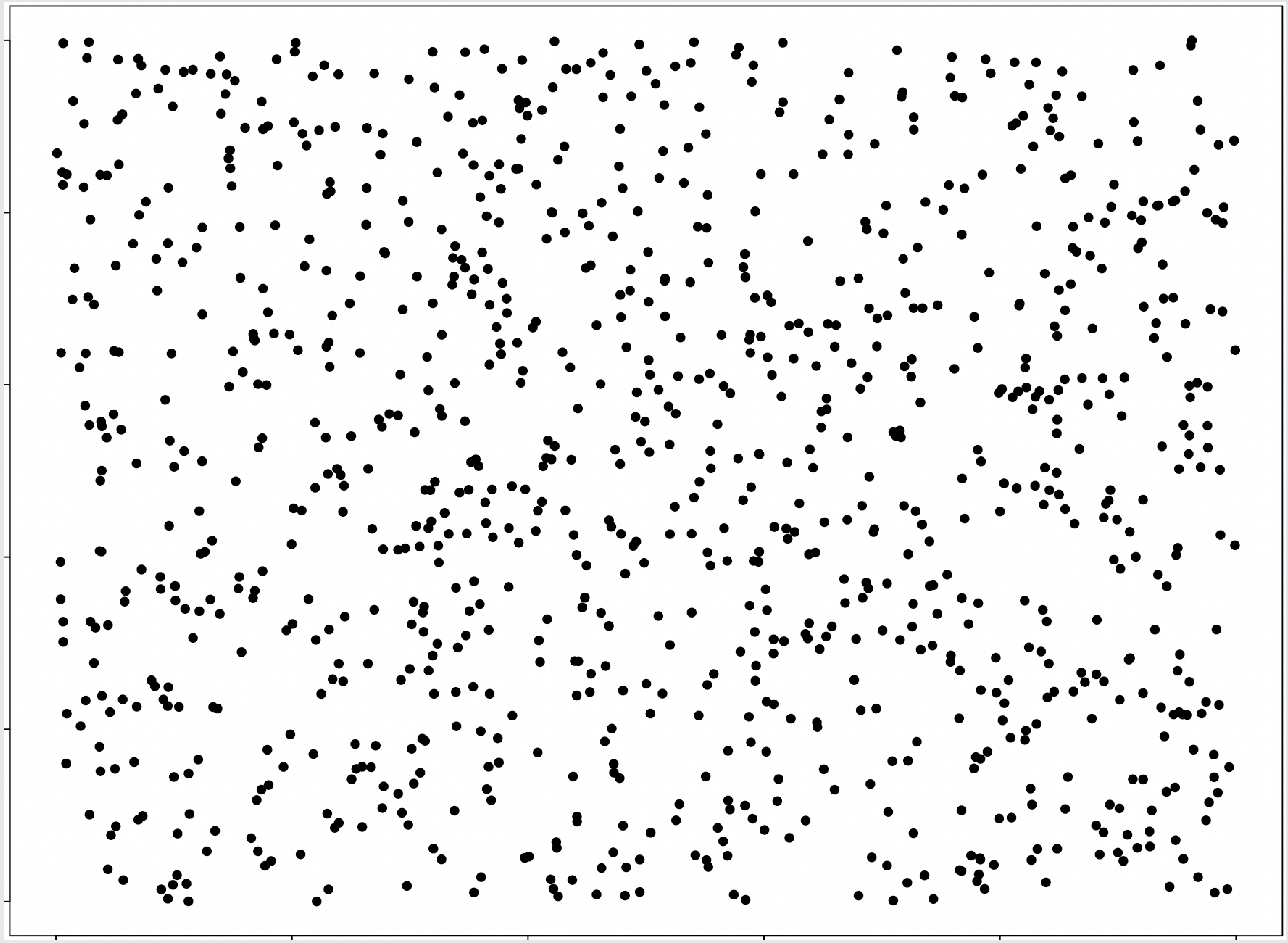
Linear predictor $\quad \eta_j(\boldsymbol{\ell}) = \boldsymbol{x}_j(\boldsymbol{\ell})^\top \boldsymbol{\beta}_j + w_j(\boldsymbol{\ell})$

Prior for spatial random effects $\quad \boldsymbol{w}(\cdot) \sim \text{meshed}GP(\mathbf{0}, \boldsymbol{C}_{\boldsymbol{\theta}}, \mathcal{G}, \boldsymbol{T})$

M. Peruzzi, S. Banerjee & A.O. Finley (2022)
Highly Scalable Bayesian Geostatistical Modeling via Meshed Gaussian Processes on Partitoned Domains.
Journal of the American Statistical Association *117(538): 969-982.*
https://www.tandfonline.com/doi/full/10.1080/01621459.2020.1833889

Meshed Gaussian processes

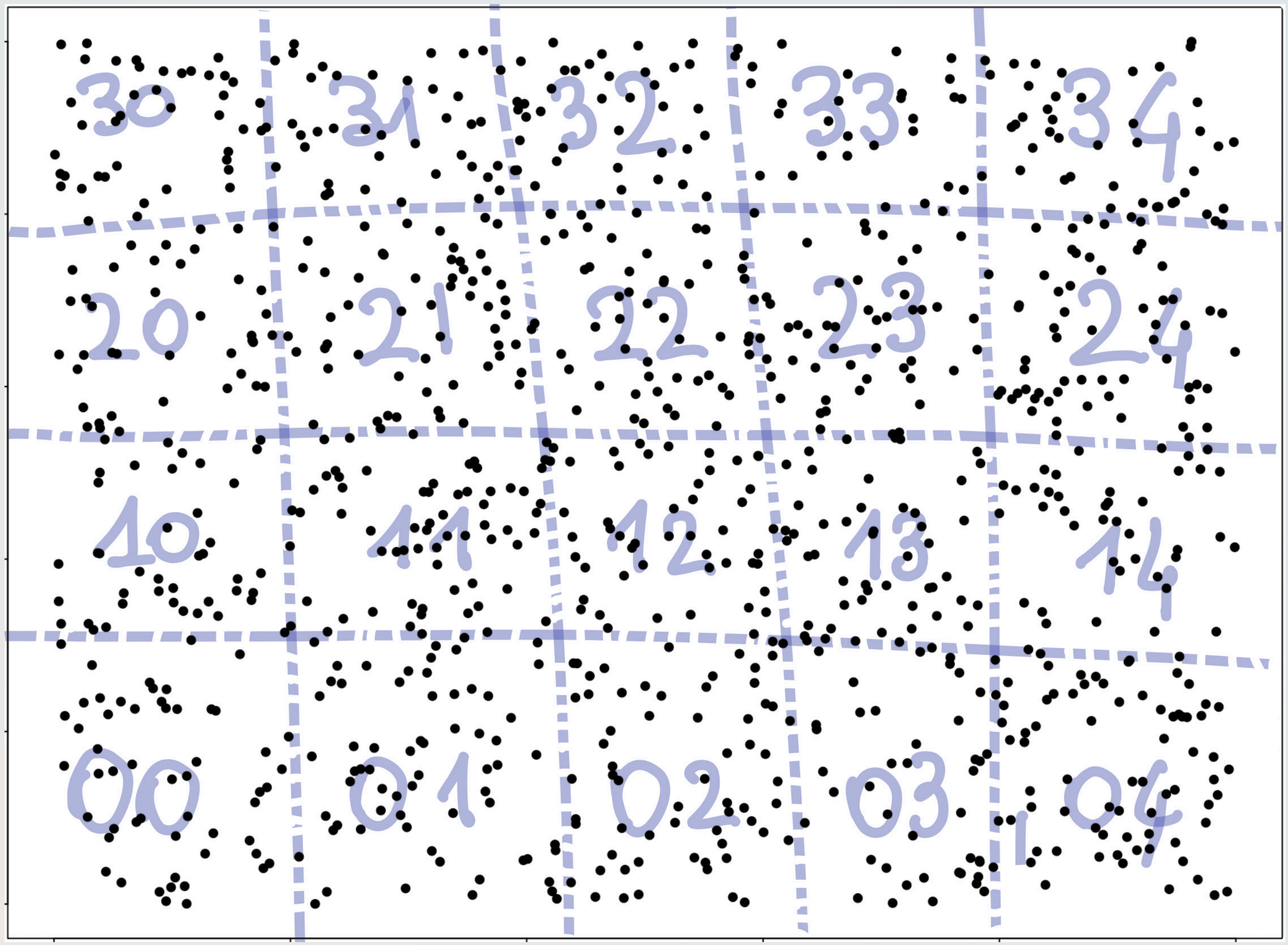
M. Peruzzi, S. Banerjee & A.O. Finley (2022)
Highly Scalable Bayesian Geostatistical Modeling via Meshed Gaussian Processes on Partitoned Domains.
Journal of the American Statistical Association *117(538): 969-982.*
https://www.tandfonline.com/doi/full/10.1080/01621459.2020.1833889



Data locations

Meshed Gaussian processes

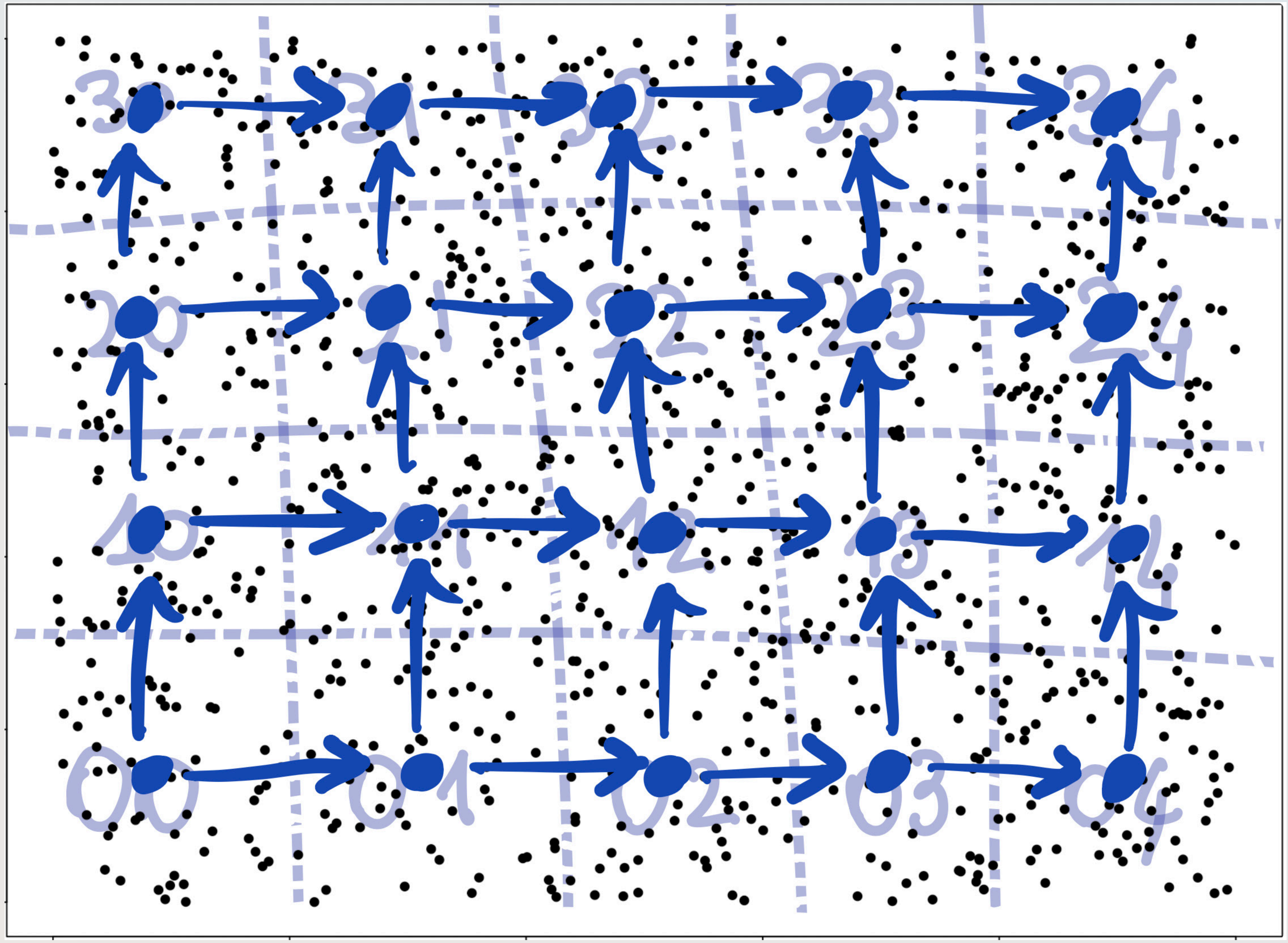
M. Peruzzi, S. Banerjee & A.O. Finley (2022)
Highly Scalable Bayesian Geostatistical Modeling via Meshed Gaussian Processes on Partitoned Domains.
Journal of the American Statistical Association *117(538): 969-982.*
https://www.tandfonline.com/doi/full/10.1080/01621459.2020.1833889



Partition the domain

Meshed Gaussian processes

M. Peruzzi, S. Banerjee & A.O. Finley (2022)
Highly Scalable Bayesian Geostatistical Modeling via Meshed Gaussian Processes on Partitoned Domains.
Journal of the American Statistical Association *117(538): 969-982.*
https://www.tandfonline.com/doi/full/10.1080/01621459.2020.1833889



Link partitions via DAG

Meshed Gaussian processes

M. Peruzzi, S. Banerjee & A.O. Finley (2022)
Highly Scalable Bayesian Geostatistical Modeling via Meshed Gaussian Processes on Partitoned Domains.
Journal of the American Statistical Association *117(538): 969-982.*
https://www.tandfonline.com/doi/full/10.1080/01621459.2020.1833889

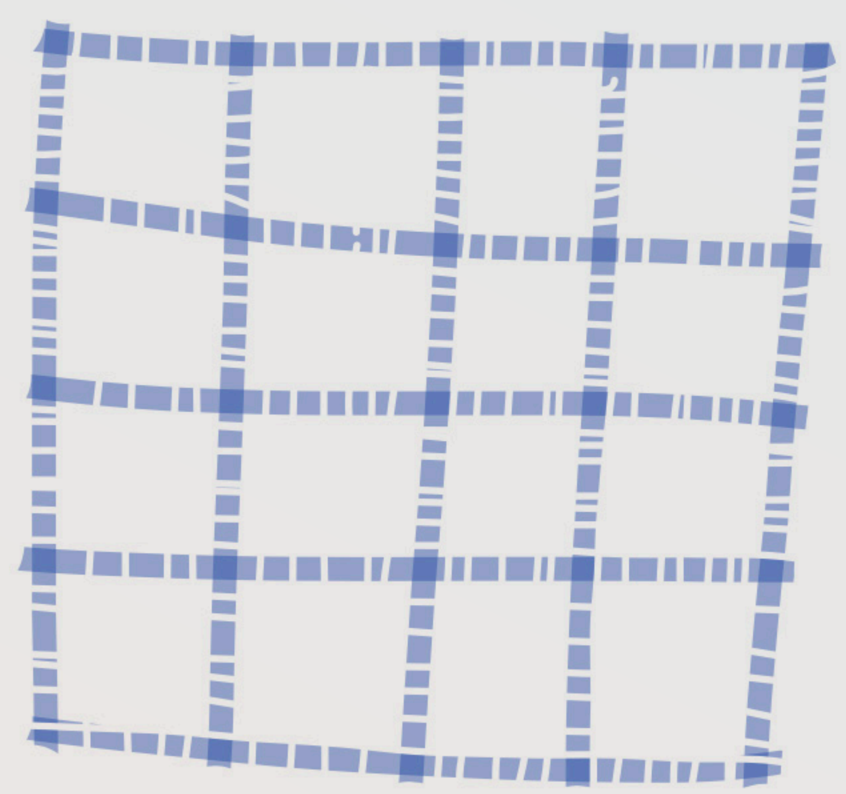
Probability model for outcome *j* $\quad y_j(\boldsymbol{\ell}) \mid \eta_j(\boldsymbol{\ell}), \tau_j \sim P_j(\eta_j(\boldsymbol{\ell}), \tau_j)$

$$\boldsymbol{\ell} \in \mathcal{D} \subset \Re^d \quad j = 1, \dots, q$$

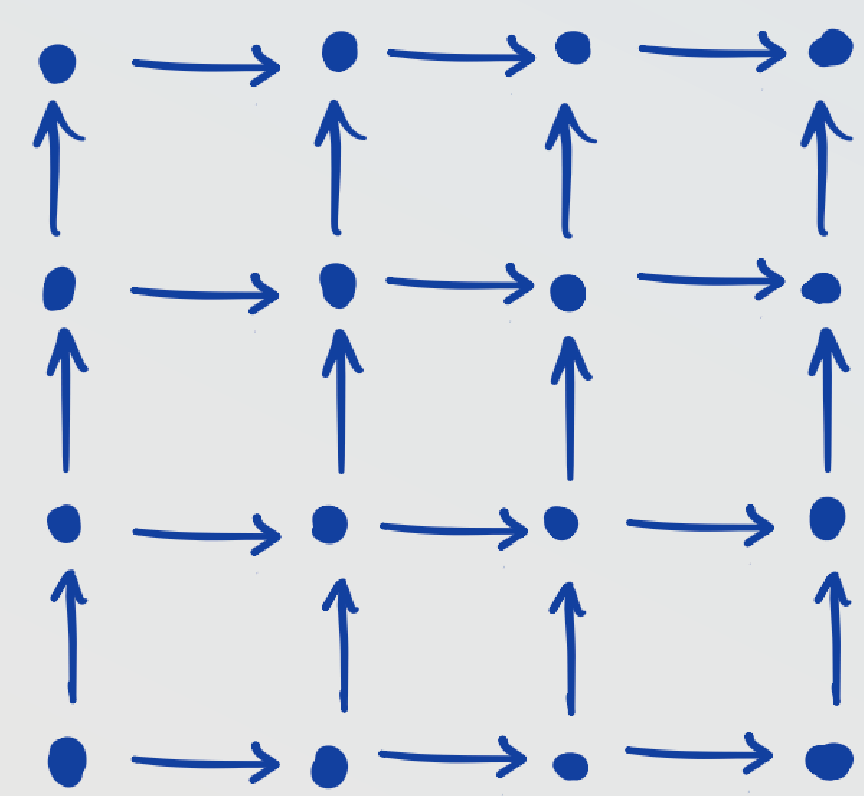
Linear predictor $\quad \eta_j(\boldsymbol{\ell}) = \boldsymbol{x}_j(\boldsymbol{\ell})^\top \boldsymbol{\beta}_j + w_j(\boldsymbol{\ell})$

Prior for spatial random effects $\quad \boldsymbol{w}(\cdot) \sim \text{meshed}GP(\mathbf{0}, \boldsymbol{C}_{\boldsymbol{\theta}}, \mathcal{G}, \boldsymbol{T})$

Domain partitioning **Sparse pattern DAG (mesh)**



+



Suppose we have *M* partitions:

$$p(\boldsymbol{w}) = \prod_{j=1}^{M} p(\boldsymbol{w}_j \mid \boldsymbol{w}_{[j]})$$

Meshed Gaussian processes: Gibbs sampler

M. Peruzzi, S. Banerjee & A.O. Finley (2022)
Highly Scalable Bayesian Geostatistical Modeling via Meshed Gaussian Processes on Partitoned Domains.
Journal of the American Statistical Association *117(538): 969-982.*
https://www.tandfonline.com/doi/full/10.1080/01621459.2020.1833889

Probability model for outcome *j* $y_j(\boldsymbol{\ell}) \mid \eta_j(\boldsymbol{\ell}), \tau_j \sim P_j(\eta_j(\boldsymbol{\ell}), \tau_j)$

$\boldsymbol{\ell} \in \mathcal{D} \subset \Re^d \quad j = 1, \dots, q$

Linear predictor $\eta_j(\boldsymbol{\ell}) = \boldsymbol{x}_j(\boldsymbol{\ell})^\top \boldsymbol{\beta}_j + w_j(\boldsymbol{\ell})$

Prior for spatial random effects $\boldsymbol{w}(\cdot) \sim \text{meshed}GP(\mathbf{0}, \boldsymbol{C}_{\boldsymbol{\theta}}, \mathcal{G}, \boldsymbol{T})$

Gibbs sampler

Cycle through these steps:

- sample $\boldsymbol{w}_j \mid \boldsymbol{w}_{-j}, y, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\tau}$
- sample $\boldsymbol{\theta} \mid \boldsymbol{w}$
- sample $\boldsymbol{\beta} \mid \boldsymbol{w}, \boldsymbol{y}, \boldsymbol{\tau}$
- sample $\boldsymbol{\tau} \mid \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{w}$

Meshed GP:
computationally **cheap** because

$$p(\boldsymbol{\theta} \mid \boldsymbol{w}) \propto \prod_{j=1}^{M} p(\boldsymbol{w}_j \mid \boldsymbol{w}_{[j]}, \boldsymbol{\theta}) p(\boldsymbol{\theta})$$

and

$$p(\boldsymbol{w}_j \mid \boldsymbol{w}_{[j]}, \boldsymbol{\theta}) = N(\boldsymbol{w}_j; \boldsymbol{H}_j \boldsymbol{w}_{[j]}, \boldsymbol{R}_j)$$

where

$$\boldsymbol{H}_j = \boldsymbol{C}_{j,[j]} \boldsymbol{C}_{[j]}^{-1}$$

$$\boldsymbol{R}_j = \boldsymbol{C}_{j,j} - \boldsymbol{H}_j \boldsymbol{C}_{[j],j}$$

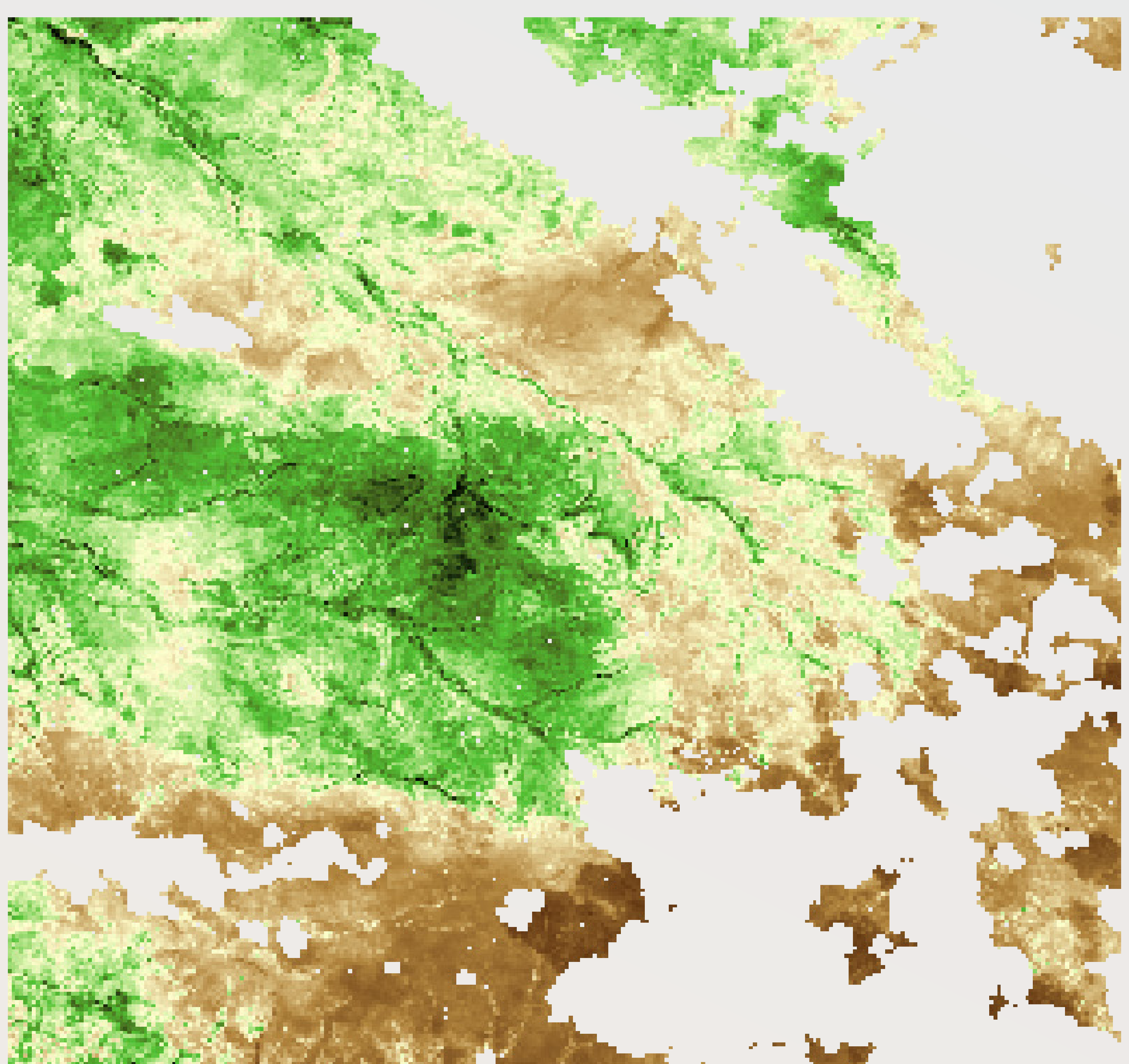
and $\boldsymbol{C}_{[j]}^{-1}$ is ***small for all j.***

Serengeti NDVI data: gap filling via Meshed GP

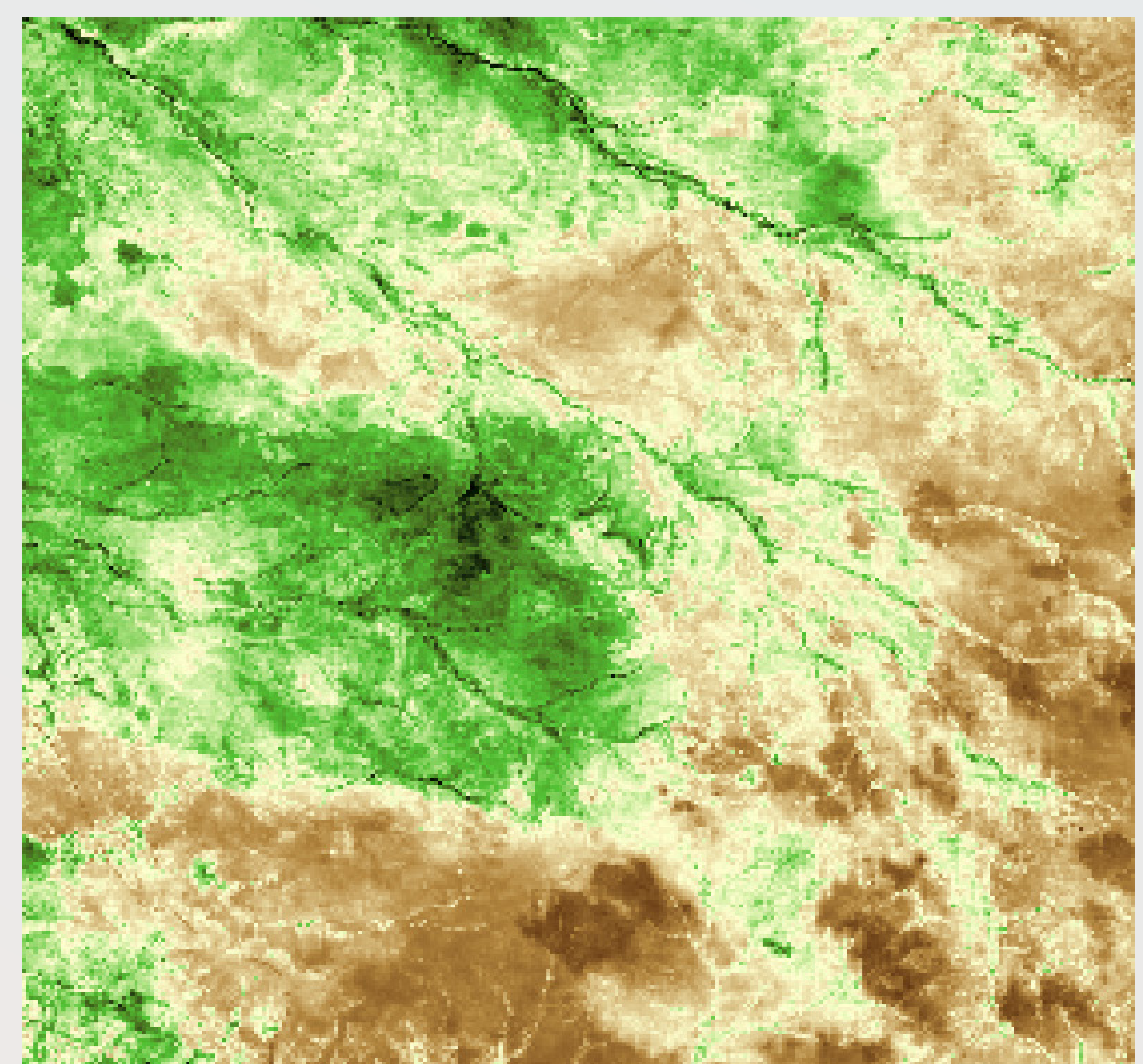
M. Peruzzi, S. Banerjee & A.O. Finley (2022)
Highly Scalable Bayesian Geostatistical Modeling via Meshed Gaussian Processes on Partitoned Domains.
Journal of the American Statistical Association *117(538): 969-982.*
https://www.tandfonline.com/doi/full/10.1080/01621459.2020.1833889

- Normalized Difference Vegetation Index (NDVI) measured by LANDSAT over **Serengeti park area**
- **16M locations** in space & time
- Cloud cover obfuscates images
- Model univariate NDVI outcome using spatiotemporally varying coefficient regression on 2 inputs (intercept, elevation)
- Posterior sampling for a **bivariate latent MGP**
- Results in less than 2 days

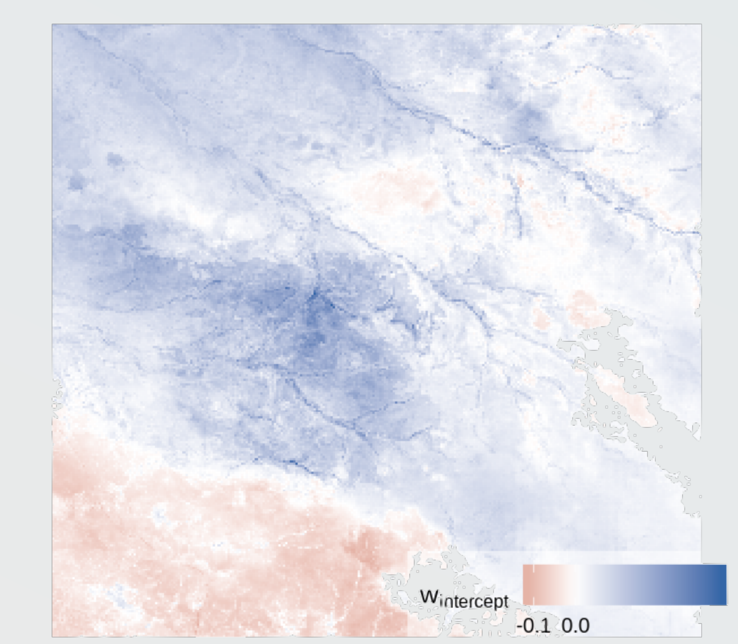
Observed NDVI



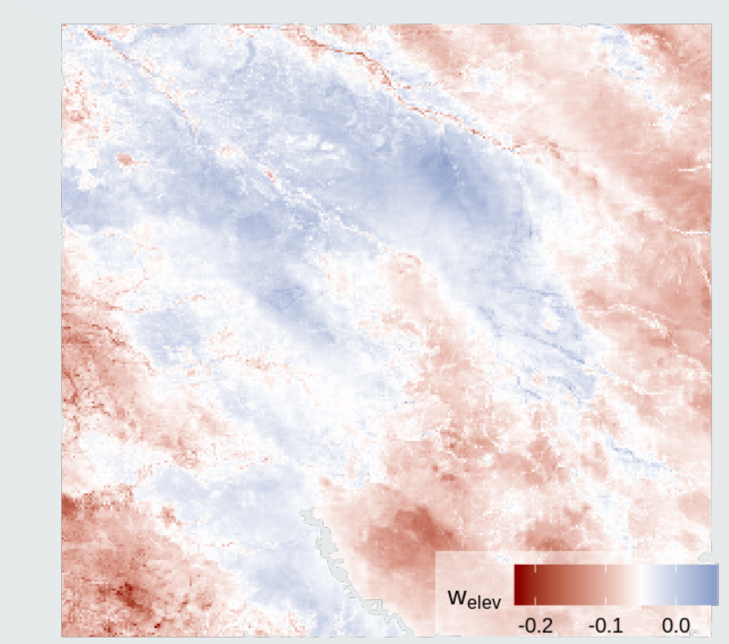
Predicted NDVI



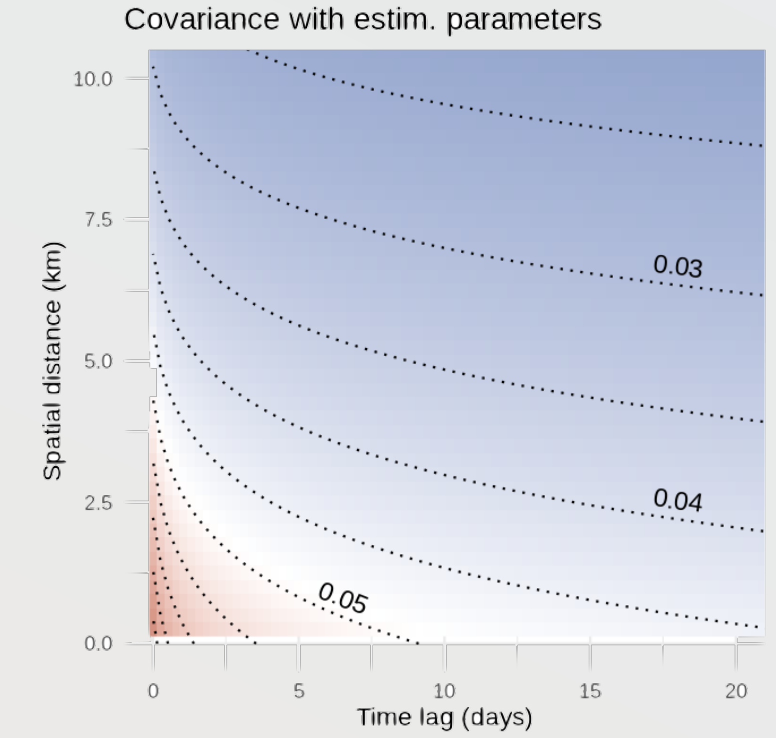
Residual spatial effect



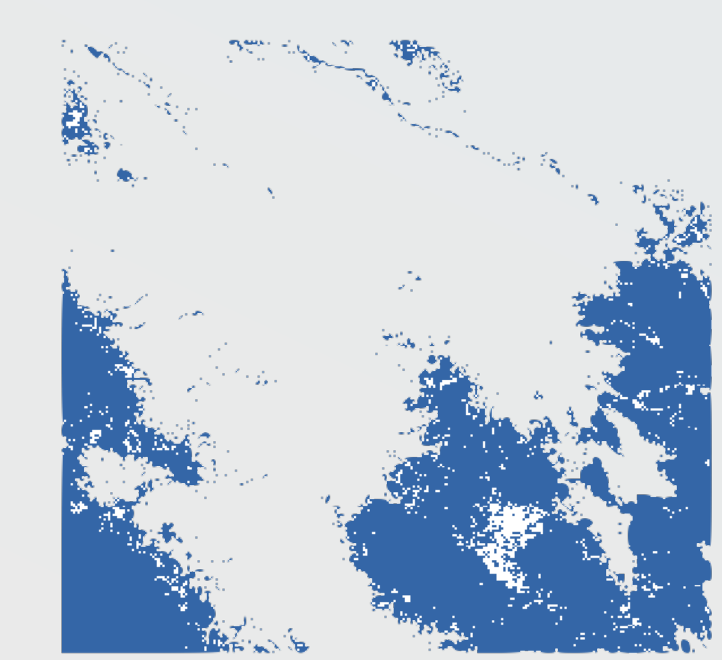
Elevation: effect on NDVI



Covariance with estim. parameters



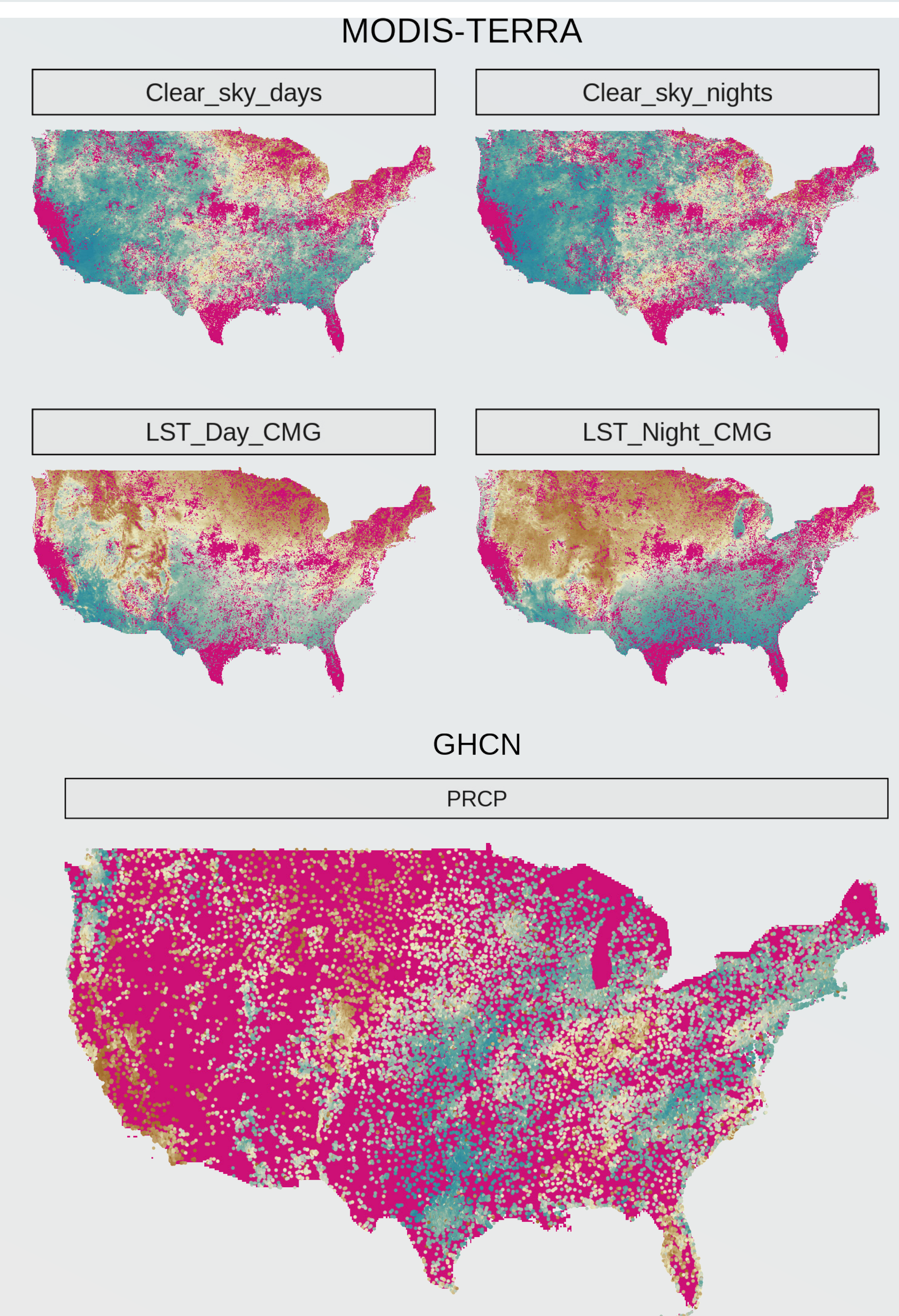
Elevation: nonzero effect locations (95% cred. int.)

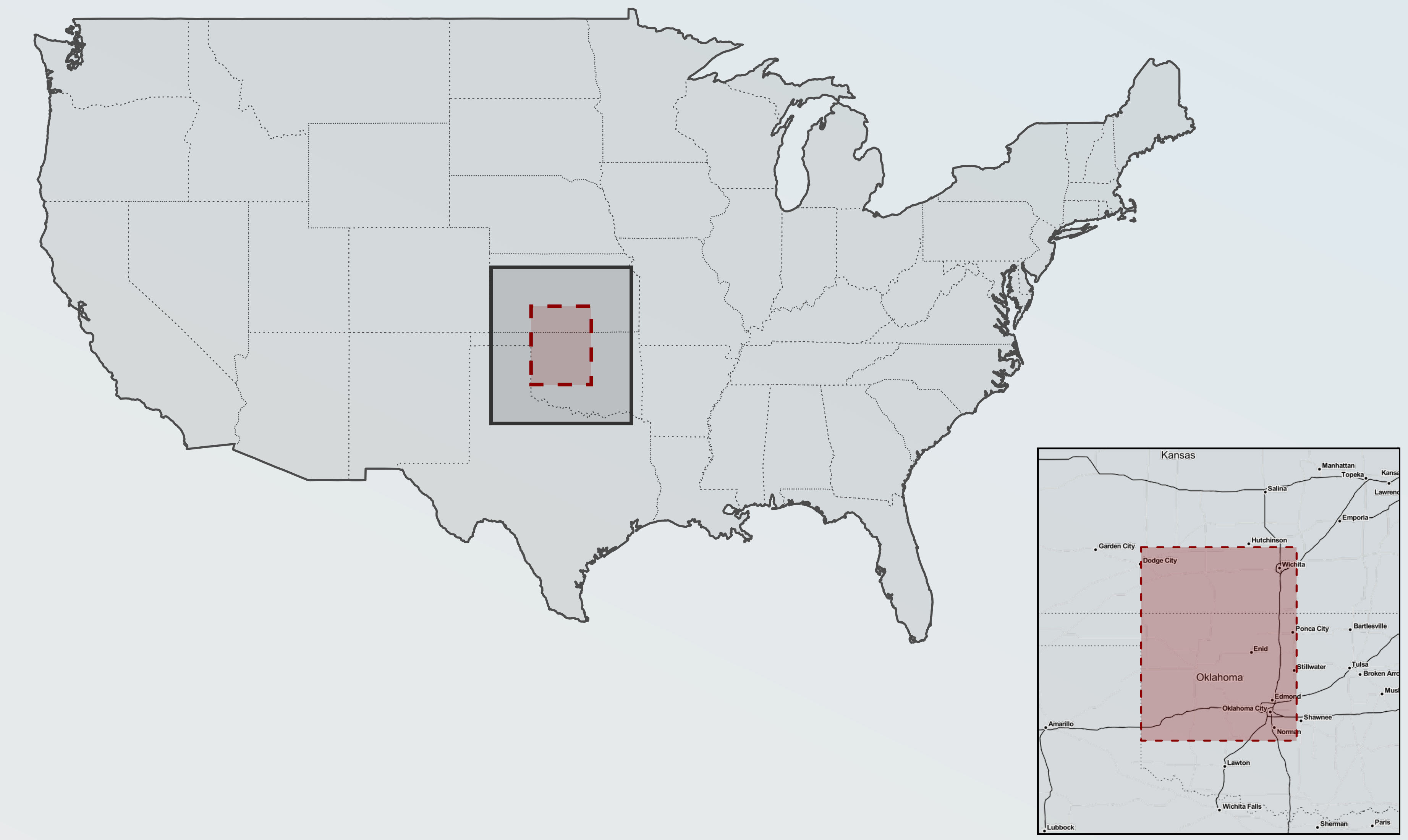


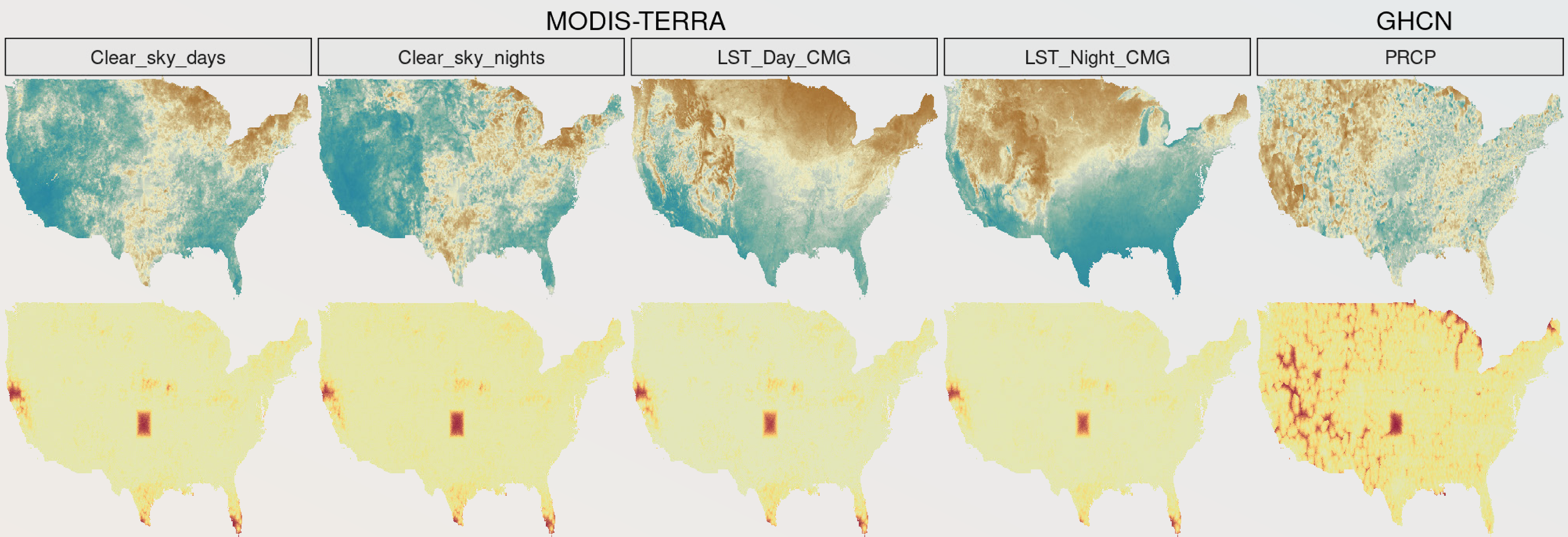
MODIS / GHCN multi-modality data with SpamTrees

M. Peruzzi & D.B. Dunson (2022)
Spatial Multivariate Trees for Big Data Bayesian Regression.
Journal of Machine Learning Research *23(17):1–40.*
https://www.jmlr.org/papers/v23/20-1361.html

- 5 outcomes from 2 different sources
- MODIS **satellite** data
- GHCN **land-based station** data
- PRCP more sparsely observed
- Misalignment
- Data size **~ 1M**
- Compute time 16 hours or less

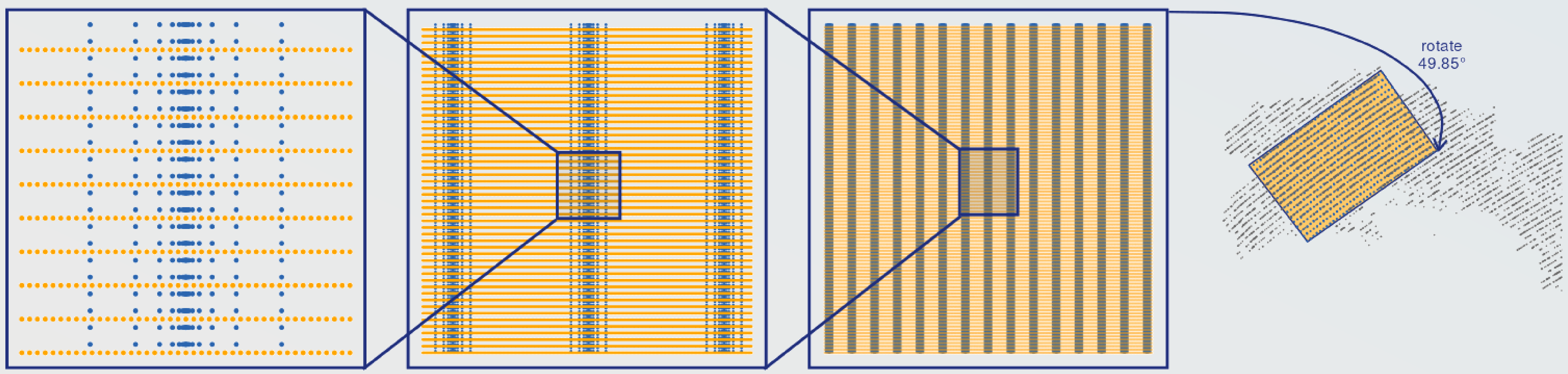


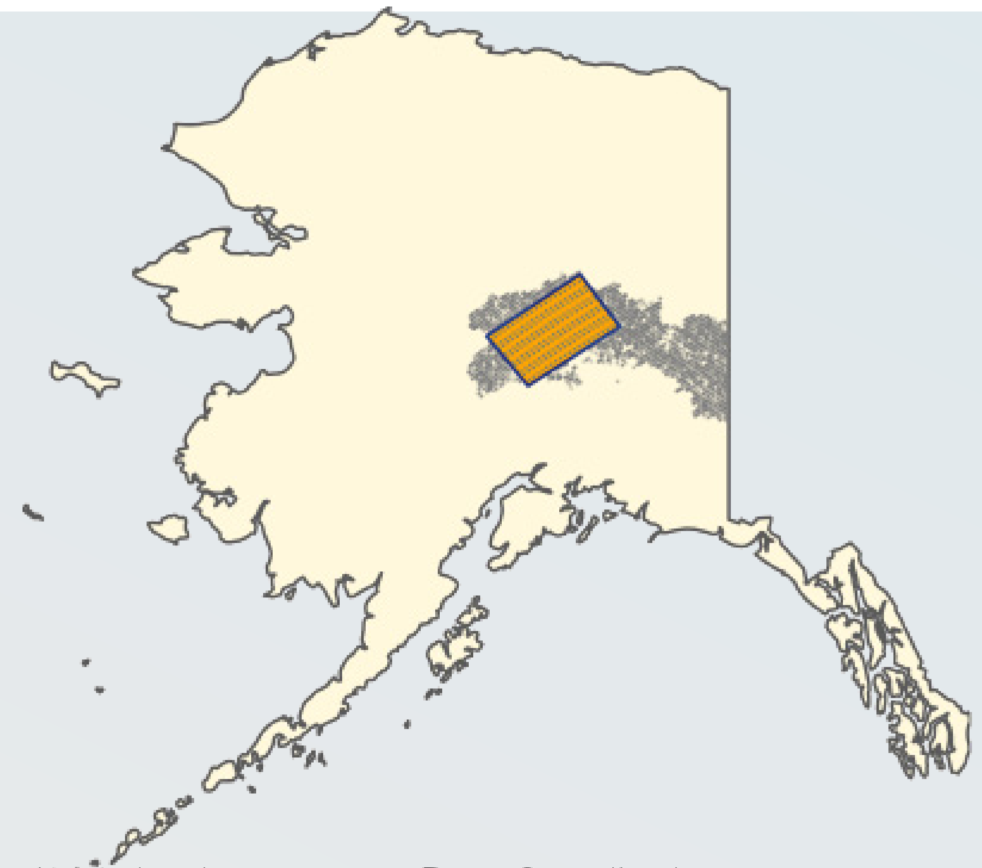


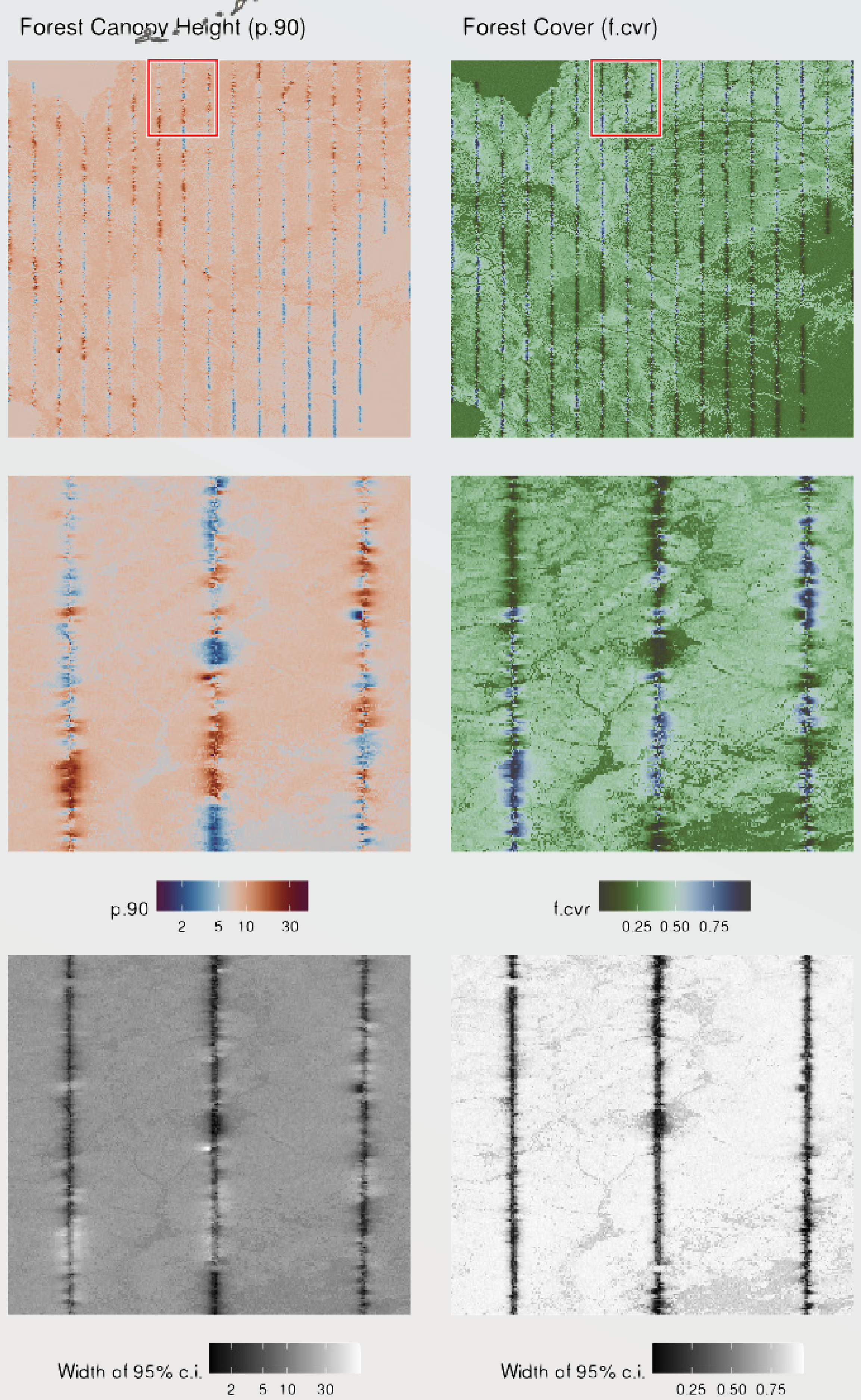


NASA LiDAR data over the Tanana forest in Alaska

M. Peruzzi, S. Banerjee, D.B. Dunson & A.O. Finley (2021)
Grid-Parametrize-Split (GriPS) for Improved Scalable Inference in Spatial Big Data Analysis.
https://arxiv.org/abs/2101.03579







GriPS: a method that involves
knot grid & parameter expansion for
improved efficiency in estimating Matérn models

Application

- Tanana forest, Alaska: 2 outcomes at **2.5M locations**
- Forest cover and canopy height (measured via LiDAR)
- Awkward measurement pattern
- Highly customized setup for MGPs for under 48h compute time

Simplified manifold preconditioning adaptation
for sampling with multivariate non-Gaussian outcomes

M. Peruzzi & D.B. Dunson (2022)
Spatial Meshing for General Bayesian Multivariate Models.
https://arxiv.org/abs/2201.10080

Meshed Gaussian processes: MCMC with **non-Gaussian outcomes**

Probability model for outcome j $\quad y_j(\boldsymbol{\ell}) \mid \eta_j(\boldsymbol{\ell}), \tau_j \sim P_j(\eta_j(\boldsymbol{\ell}), \tau_j)$

$$\boldsymbol{\ell} \in \mathcal{D} \subset \Re^d \quad j = 1, \dots, q$$

Linear predictor $\quad \eta_j(\boldsymbol{\ell}) = \boldsymbol{x}_j(\boldsymbol{\ell})^\top \boldsymbol{\beta}_j + w_j(\boldsymbol{\ell})$

Prior for spatial random effects $\quad \boldsymbol{w}(\cdot) \sim \text{meshed}GP(\mathbf{0}, \boldsymbol{C}_{\boldsymbol{\theta}}, \mathcal{G}, \boldsymbol{T})$

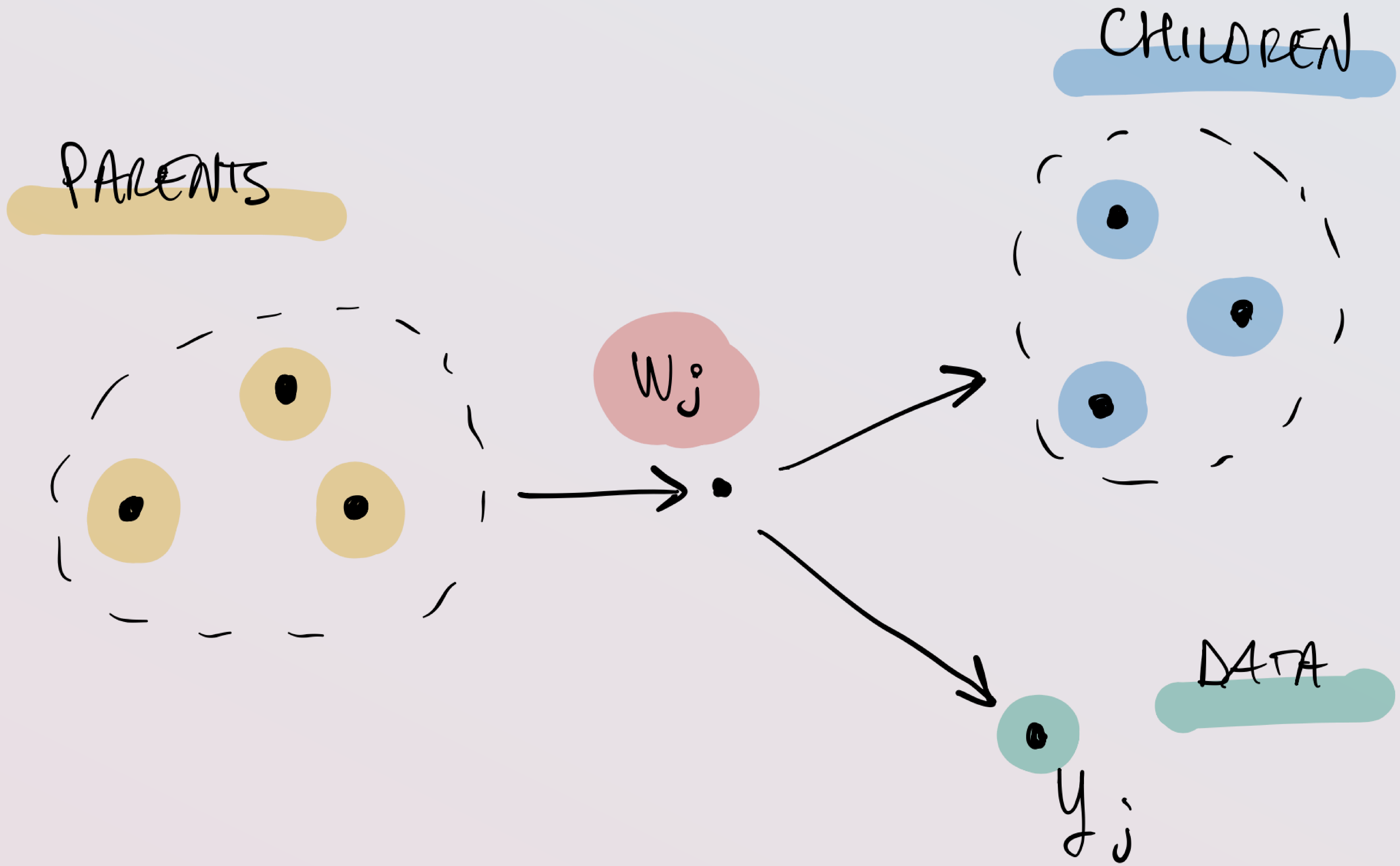
Gibbs sampler

Cycle through these steps:

- sample $\boldsymbol{w}_j \mid \boldsymbol{w}_{-j}, y, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\tau}$
- sample $\boldsymbol{\theta} \mid \boldsymbol{w}$
- sample $\boldsymbol{\beta} \mid \boldsymbol{w}, \boldsymbol{y}, \boldsymbol{\tau}$
- sample $\boldsymbol{\tau} \mid \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{w}$

Lack of conjugacy:
How can we update $\boldsymbol{w}_j$ efficiently?

Meshed GP:
computationally cheap



- Propose move from w_j to w_j^* via the function $q(\cdot \mid w_j)$
- Target density is the full conditional distribution of w_j

$$p(w_j \mid {\scriptstyle everything}) \propto p(y_j \mid w_j, \ldots)\pi(w_j \mid w_{[j]}) \prod_{j \to i} \pi(w_i \mid w_j, w_{[i]\setminus j})$$

- Propose move from w_j to w_j^* via the function $q(\cdot \mid w_j)$
- Target density is the full conditional distribution of w_j

$$p(w_j \mid \textit{everything}) \propto p(y_j \mid w_j, \ldots)\pi(w_j \mid w_{[j]}) \prod_{j \to i} \pi(w_i \mid w_j, w_{[i]\setminus j})$$

- Accept based on the ratio

$$\alpha = \frac{p(w_j^* \mid \textit{everything})q(w_j \mid w_j^*)}{p(w_j \mid \textit{everything})q(w_j^* \mid w_j)}$$

- If we can sample from full conditional itself, then set

$$q(w_j \mid w_j^*) = p(w_j \mid \textit{everything})$$

to get $\alpha = 1$ which leads to a Gibbs sampler

- If we cannot sample from full conditional itself, then **what proposal do we use?**

Preconditioned MALA

$$w_j^* \sim N(w_j + \frac{\varepsilon^2}{2} G g, \varepsilon^2 G)$$

$$g = \nabla p(w_j \mid {}_{everything})$$

gradient of the target density

and G is a preconditioner matrix which **we need to specify**

- Use accepted values to build G based on the Markov chain's past? **Slow to adapt**
- Use a position-dependent G_{w_j} à la Riemann manifold MALA? **Expensive update**

Simplified manifold preconditioner adaptation (SiMPA)

at iteration m of our Metropolis algorithm:

$$w_j^* \sim N(w_j + \frac{\varepsilon^2}{2} G_m g, \varepsilon^2 G_m)$$

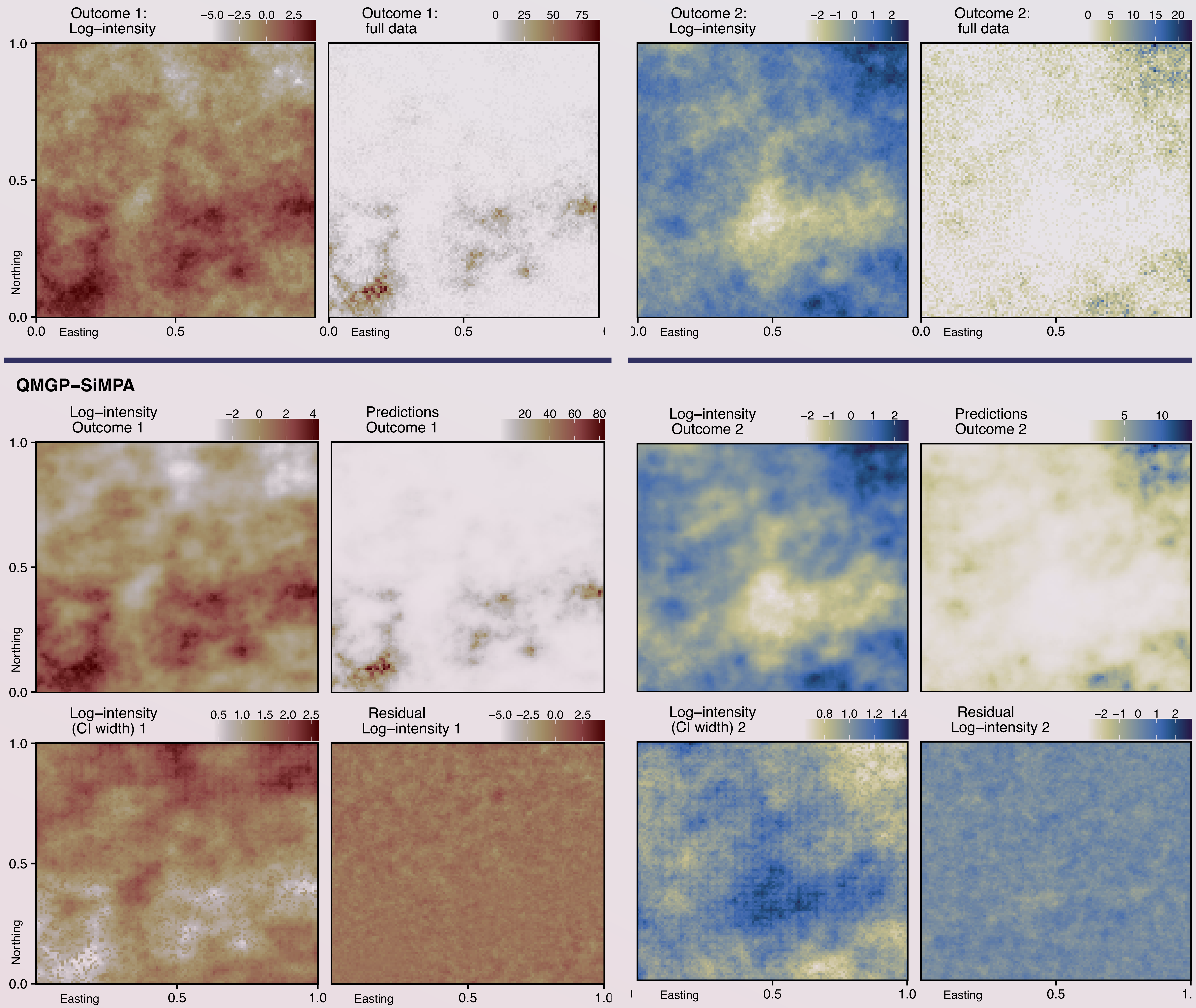
- accept or reject as usual based on this proposal
- afterwards, with probability $\gamma_m \downarrow 0$ set

$$G_m^{-1} = G_{m-1}^{-1} + \kappa(\nabla\nabla p(w_j \mid _{everything})) - G_{m-1}^{-1})$$

- Asymptotically constant preconditioner: cheap update
- Uses target density's second order information, similar to Riemannian manifold MALA: efficient update

SiMPA: Simplified manifold preconditioner adaptation – simulated data

Poisson outcome 1 Poisson outcome 2

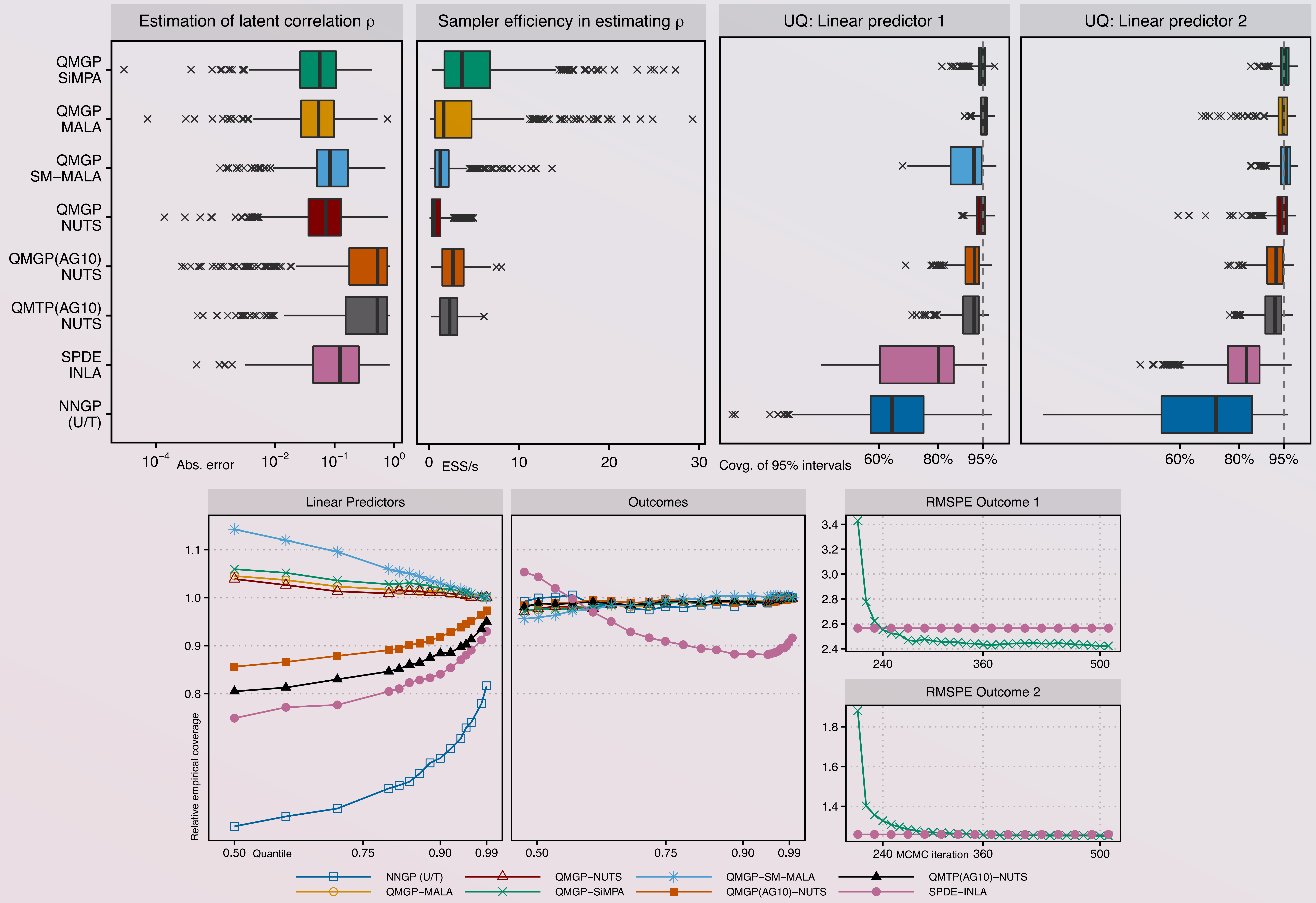


- Two related count outcomes
  - *n=3600*
  - leave out 20% of data
- Goals
  - prediction
  - recovery of latent log-intensity
  - uncertainty quant. about log-intensity

- meshed GPs with SiMPA
  - results in <10s
  - orders of magnitude faster than full GP

- using R package `meshed` 0.2.1
  (it's on CRAN)

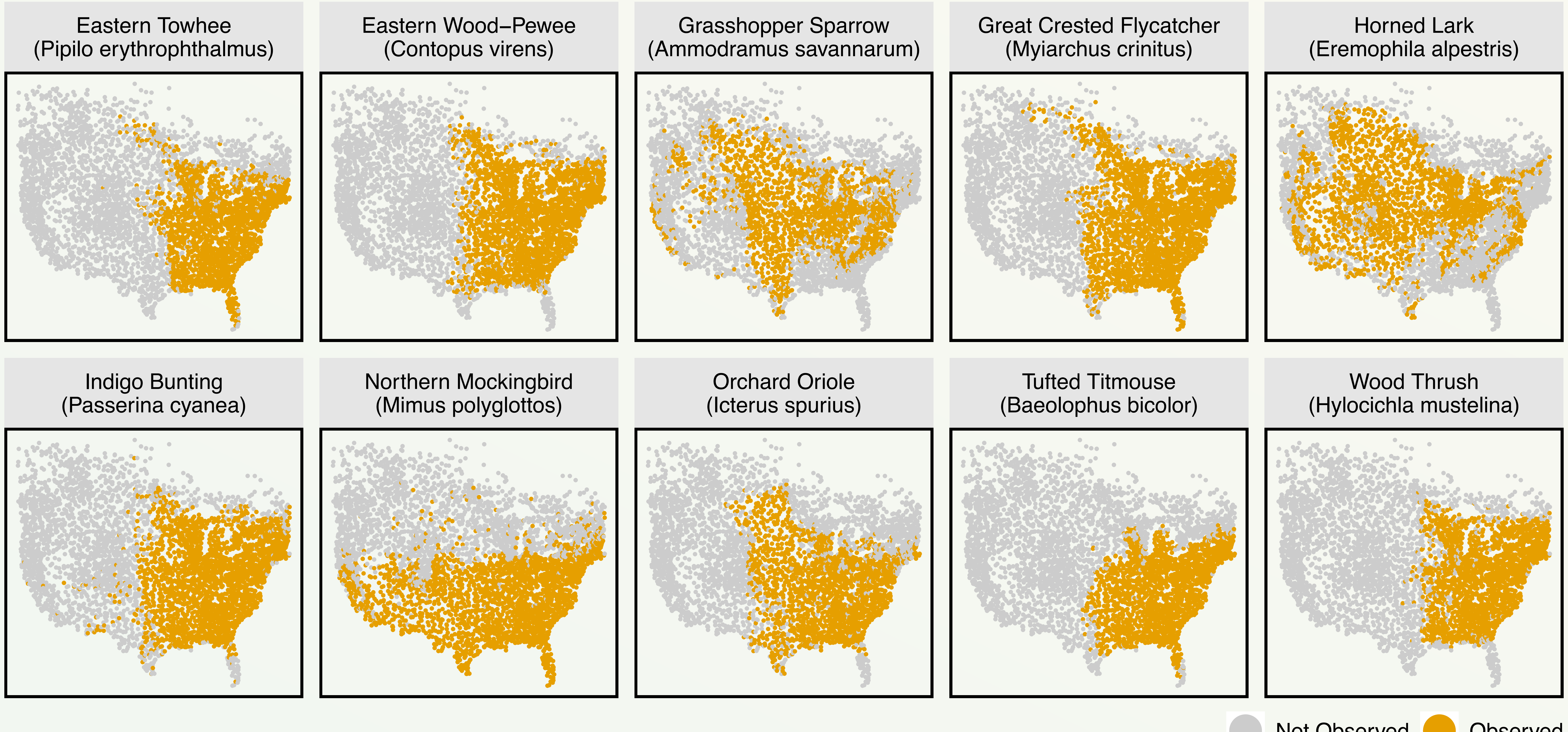
SiMPA: Simplified manifold preconditioner adaptation. Simulated data



R package meshed:
demonstrating MGPs and SiMPA on the
North American Breeding Bird survey

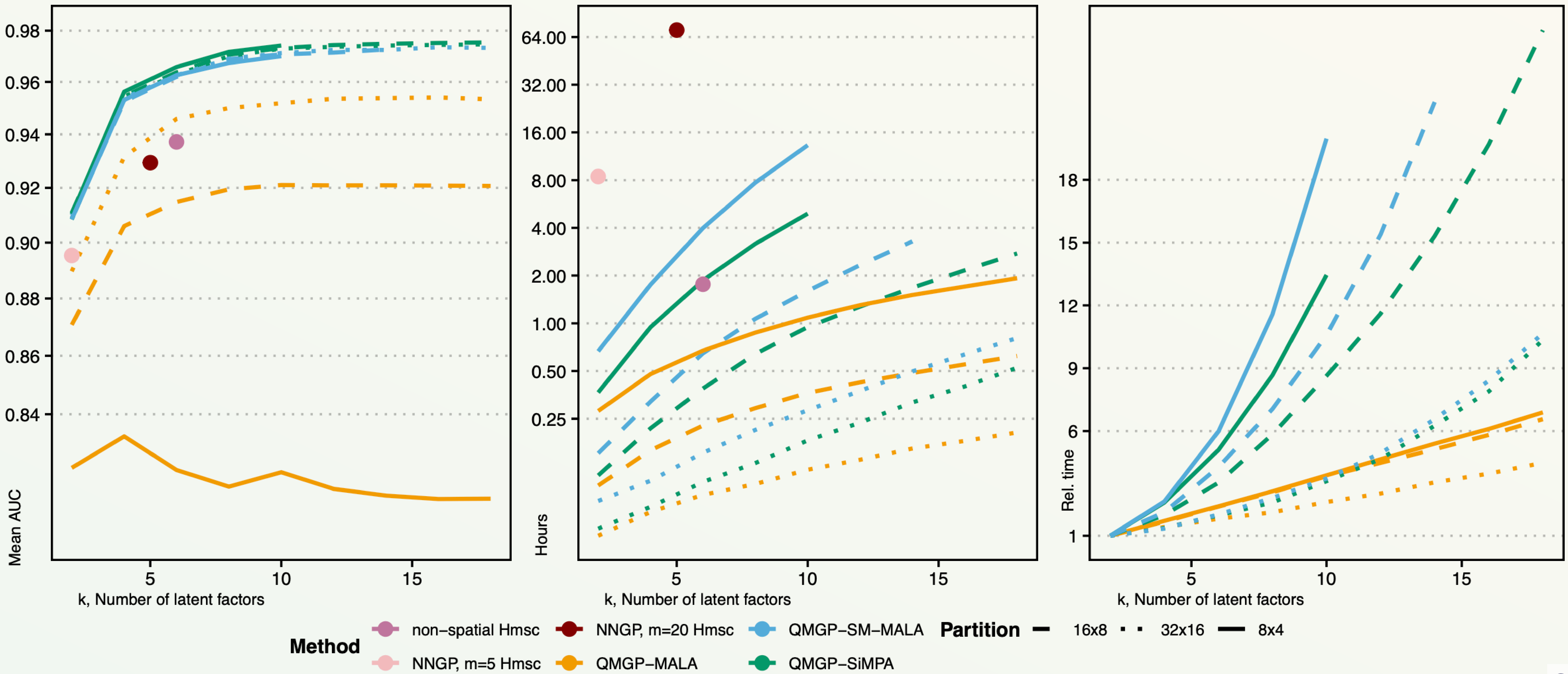
North American Breeding Bird Survey data

- Somewhat small number of spatial locations (~ 4100)
- Large number of binary outcomes, *q=27*
- Effective data size ~100k
- Presence/Absence data about birds of the *Passeriforme* order
- Create a **test set** using 20% of data
- **Target**: classification performance for all bird species measured via area under ROC curve (AUC)
- **Compare** with popular software **`Hmsc`** (non-spatial, spatial w/ NNGP) commonly used by biologists



North American Breeding Bird Survey data

- **Compare** with popular software **Hmsc** (non-spatial, spatial w/ NNGP) commonly used by biologists
- **Average AUC** in out-of-sample classification across 27 bird species



Relevant articles

- Peruzzi M, Dunson DB (2023) Spatial meshing for general multivariate Bayesian models (arXiv: 2201.10080)
- Peruzzi M, Banerjee S, Finley AO (2022) Meshed GPs. JASA 117(538): 969–982. (doi: 10.1080/01621459.2020.1833889)
- Peruzzi M, Dunson DB (2022) Spatial Multivariate Trees. JMLR (https://www.jmlr.org/papers/v23/20-1361.html)
- Peruzzi M, Banerjee S, Dunson DB, Finley AO (2021) Grid-Parametrize-Split (GriPS) (submitted, arXiv: 2101.03579)
- Jin B, Peruzzi M, Dunson DB (2021) Bags of DAGs (submitted, arXiv: 2112.11870)

R packages

- **`meshed`** R package: *https://CRAN.R-project.org/package=meshed*
- **`spamtree`** R package: *https://CRAN.R-project.org/package=spamtree*
- **`spammix`** *R package: https://github.com/mkln/spammix*
  (multi-species N-mixture modeling with latent spatial factors)

peruzzi@umich.edu

Thank you!