DISSERTATION

ADVANCES IN BAYESIAN SPATIAL STATISTICS FOR ECOLOGY AND ENVIRONMENTAL SCIENCE

Submitted by

Wilson J. Wright

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2024

Doctoral Committee:

Advisor: Mevin B. Hooten
Co-Advisor: Daniel S. Cooley

Kayleigh P. Keller
Andee Kaplan
Matthew R.V. Ross

ABSTRACT

ADVANCES IN BAYESIAN SPATIAL STATISTICS FOR ECOLOGY AND ENVIRONMENTAL SCIENCE

In this dissertation, I develop new Bayesian methods for analyzing spatial data from applications in ecology and environmental science. In particular, I focus on methods for mechanistic spatial models and binary spatial processes. I first consider the distribution of heavy metal pollution from a mining road in Cape Krusenstern, Alaska, USA. I develop a mechanistic spatial model that uses the physical process of atmospheric dispersion to characterize the spatial structure in these data. This approach directly incorporates scientific knowledge about how pollutants spread and provides inferences about this process. To assess how the heavy metal pollution impacts the vegetation community in Cape Krusenstern, I also develop a new model that represents plant cover for multiple species using clipped Gaussian processes. This approach is applicable to multiscale and multivariate binary processes that are observed at point locations — including multispecies plant cover data collected using the point intercept method. By directly analyzing the point-level data, instead of aggregating observations to the plot-level, this model allows for inferences about both large-scale and small-scale spatial dependence in plant cover. Additionally, it also incorporates dependence among different species at the small spatial scale. The third model I develop is motivated by ecological studies of wildlife occupancy. Similar to plant cover, species occurrence can be modeled as a binary spatial process. However, occupancy data are inherently measured at areal survey units. I develop a continuous-space occupancy model that accounts for the change of spatial support between the occurrence process and the observed data. All of these models are implemented using Bayesian methods and I present computationally efficient methods for fitting them. This includes a new surrogate data slice sampler for implementing models with latent nearest neighbor Gaussian processes.

ACKNOWLEDGEMENTS

Many thanks to my committee, Kayleigh Keller, Andee Kaplan, and Matthew Ross, for their help and support during my PhD.

I also would like to thank my co-advisor Daniel Cooley for all of the research discussions we had and the helpful advice that he gave me.

I am extremely grateful to my advisor Mevin Hooten. I appreciated having an advisor who always pushed me to do better — Mevin provided motivation that improved my work and helped me grow as a researcher. I have also greatly benefited from Mevin's insights about research and a career in academia.

During my dissertation research, I had the pleasure of working with Peter Neitlich, Alyssa Shiel, and Elisa Di Meglio and am thankful for having such great research collaborators.

I would also like to extend my thanks to the other members of the Hooten Lab: Clint Leach, Lucy Lu, Michael Schwob, and Justin Van Ee. I have appreciated working alongside such hard-working and inspiring individuals.

I never would have started a PhD without the support and encouragement of Megan Higgs and Jenny Green. Thanks to you both for helping me start this journey and for everything you taught me.

Prior to coming to CSU, I was fortunate to be able to work with Kathi Irvine. I learned a lot about research, ecological statistics, and collaborating with other scientists during that time. I am extremely thankful for all of Kathi's mentorship.

I would also like to thank my parents, Jay and Laura, for all of the opportunities that they have provided me and for always encouraging me.

Finally, I could not have undertaken this journey without the love and unwavering support from my wife Jane. She always believed in me, even on the days that I didn't believe in myself.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Motivation

In this dissertation, I develop new statistical methods for modeling spatial dependence in ecological and environmental data. Specifically, these methods focus on mechanistic spatial models and binary spatial processes. The models I develop are motivated by datasets on heavy metal concentrations, vegetation communities, and occurrences of wildlife species — all of which include spatial structure that is vital for understanding these processes. For each dataset, I aim to better incorporate scientific knowledge in analyses and also account for features of the observed data that are often ignored by other approaches. Consequently, inferences from the models I develop can provide insights that are not possible using other statistical methods.

The remainder of this chapter includes brief introductions to species distribution models, spatial statistics, and Bayesian methods. More in-depth overviews of these topics can be found in Kéry and Royle (2016), MacKenzie et al. (2018), Cressie (2015), Schabenberger and Gotway (2017), Gelman et al. (2014), and Carlin and Louis (2008). Each of the models I develop is implemented using Bayesian methods because this framework is useful for fitting models with hierarchical structures and latent spatial processes. Therefore, I also review standard techniques in Bayesian computation that are used throughout Chapters 2–4. This chapter concludes by providing an overview of the remaining chapters and introducing the datasets that motivate this research.

1.2 Species distribution models

Species distribution models are used to learn about the abundance and/or occurrence of a species throughout a study area of interest. These models use environmental variables to predict how species are distributed (Elith and Leathwick, 2009) and often include spatial dependence as well (e.g., Section 1.3; Guisan and Thuiller, 2005). Approaches for modeling species abundance include

N-mixture models (Royle, 2004), the Royle-Nichols model (Royle and Nichols, 2003), and distance sampling (Burnham et al., 1980; Buckland et al., 2016). Common approaches for modeling species occurrence include occupancy models (Hoeting et al., 2000; MacKenzie et al., 2002, 2018) and various of machine learning methods (e.g., Phillips et al., 2006; Valavi et al., 2021). Many of these species distribution models focus on analyzing wildlife data, but similar approaches are available for modeling the distributions of plants (e.g., Hooten et al., 2003; Wright et al., 2017).

Accounting for observation errors is a common theme of many species distribution models because these errors are a typical characteristic of most ecological data. If unaccounted for, observation errors can result in biased inferences for the ecological processes of interest. Consider occupancy models (MacKenzie et al., 2002, 2018) as an example — I introduce this approach in what follows and present a new spatial occupancy model in Chapter 4. For surveyed sites $i = 1, \ldots, n$, standard occupancy models assume

$$z_i \sim \text{Bernoulli}(\psi_i), \tag{1.1}$$

where z_i is an indicator for whether the species of interest is present (1) or absent (0) at site i and ψ_i denotes the corresponding probability of occurrence. If the indicators for species occurrence z_i are observed directly, then standard regression models for binary data can be used to relate the probability of occupancy to predictor variables of interest. However, species often go undetected during surveys of occupied sites and failing to account for imperfect detection can negatively bias occupancy estimates (MacKenzie et al., 2002).

Replicate surveys for some sites are needed to model the observation process of how species are detected (MacKenzie et al., 2002). These replicate surveys can include observations from independent observers or repeated visits to the same site. The observed detection/nondetection data y_{ij} are modeled as

$$y_{ij} \sim \begin{cases} \mathbb{1}(y_{ij} = 0), & z_i = 0, \\ \text{Bernoulli}(p_{ij}), & z_i = 1, \end{cases} \tag{1.2}$$

for $j = 1, \ldots, J_i$ where J_i denotes the total number of visits to site i and p_{ij} denotes the probability of detection at site i during visit j. Visit-level predictor variables can be used to model the probability of detection. The key idea of occupancy modeling is that even when the species is undetected at a site, that is $\sum_{j=1}^{J_i} y_{ij} = 0$, the species may still occur there. Modeling the detection process accounts for this and can provide unbiased inferences for species occurrence.

In some applications more complicated detection models may be needed when analyzing occurrence data. For instance, detection probabilities can depend on the local abundance at a site (Royle, 2006) and some surveys can include false-positive detections as well (e.g., Chambert et al., 2015; Ruiz-Gutierrez et al., 2016). Many models for species abundance, such as N-mixture models and distance sampling, include analogous observation processes to account for imperfect detection. Hefley and Hooten (2016) developed a hierarchical framework that shows how some species distribution models relate to one another. These connections also emphasize that hierarchical models are invaluable for modeling species distributions because they allow complex observation processes to be included in analyses.

I develop new species distribution models in Chapters 3 and 4. First, I consider multispecies plant cover data collected using the point intercept method (Drezner and Drezner, 2021). In Chapter 3, I develop a new model for these data that includes spatial dependence at multiple scales and also allows for dependence among different species. Chapter 4 considers occurrence data for wildlife species and develops a new continuous-space occupancy framework. The models in Chapters 3 and 4 both demonstrate new approaches for modeling the spatial structure in species distributions.

1.3 Spatial statistics

Modeling spatial dependence is a core component of ecological and environmental data analysis because the data from these applications are inherently spatial. Failing to adequately account for spatial dependence can lead to inferences that are overly precise (Legendre and Fortin, 1989). More importantly, however, is the fact that the spatial processes themselves are typically of interest in ecological and environmental applications because they are directly related to the scientific

questions of interest (Legendre and Fortin, 1989; Levin, 1992; Guisan and Thuiller, 2005). For instance, consider mapping the distribution of a species across the landscape (e.g., Latimer et al., 2006; Hefley and Hooten, 2016; Gelfand, 2022), a topic that I discuss in more detail in Chapters 3 and 4. Species distribution models include environmental predictor variables, which are spatially indexed themselves, as well as spatial structure that arises from the dispersal of individuals and/or geographical limitations for where species occur (Legendre and Fortin, 1989; Levin, 1992; Guisan and Thuiller, 2005). In some applications, more specific knowledge about how species spread throughout a region can be incorporated into analyses to characterize the spatial distribution of a species (e.g., Hooten and Wikle, 2008). Consequently, developing spatial models that can better capture these characteristics and incorporate scientific knowledge are imperative for ecological and environmental data analyses.

1.3.1 Gaussian processes

For response variables with a continuous support, a standard approach for modeling the spatial dependence among observations in continuous space is to use a Gaussian process (Neal, 1999; Rasmussen and Williams, 2006; Gelfand and Schliep, 2016). Spatial Gaussian processes are parameterized using a mean function m and a covariance function K that depends on the distance between observations. For example, an exponential covariance function defines the covariance between locations $\mathbf{s}_i$ and $\mathbf{s}_{i'}$ to be

$$K(\mathbf{s}_i, \mathbf{s}_{i'}) = \sigma^2 \text{exp}\left(-\frac{||\mathbf{s}_i - \mathbf{s}_{i'}||}{\rho}\right), \tag{1.3}$$

where $||\mathbf{s}_i - \mathbf{s}_{i'}||$ denotes the Euclidean distance between the locations and σ^2 and ρ are variance and range parameters, respectively. As the distance between locations approaches zero, (1.3) results in the observations at these locations being perfectly correlated. Consequently, if $y(\mathbf{s})$ is assumed to be distributed as a Gaussian process, denoted $y(\mathbf{s}) \sim \text{GP}(m(\mathbf{s}), K)$, then realizations of this process are random continuous functions $\forall \mathbf{s} \in \mathbb{R}^d$. These random functions are constrained so that at any finite collection of point locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$, the random variables $(y(\mathbf{s}_1), \ldots, y(\mathbf{s}_n))$ will follow a

multivariate normal distribution. Specifically, a two-dimensional subset of the Gaussian process $y(\mathbf{s})$ results in

$$\begin{pmatrix} y(\mathbf{s}_i) \\ y(\mathbf{s}_{i'}) \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} m(\mathbf{s}_i) \\ m(\mathbf{s}_{i'}) \end{pmatrix}, \begin{pmatrix} K(\mathbf{s}_i, \mathbf{s}_i) & K(\mathbf{s}_i, \mathbf{s}_{i'}) \\ K(\mathbf{s}_{i'}, \mathbf{s}_i) & K(\mathbf{s}_{i'}, \mathbf{s}_{i'}) \end{pmatrix} \right), \tag{1.4}$$

for any pair of locations $\mathbf{s}_i$ and $\mathbf{s}_{i'}$. This provides a convenient framework for modeling spatial data because the finite-dimensional distributions are normal, allowing us to take advantage of many properties of the normal distribution. Additionally, Gaussian processes with common covariance parameterizations result in models that obey the first law of geography that "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970).

Gaussian processes can still be used for analyzing data that cannot be assumed to follow a normal distribution. In this case, spatial dependence can be included using a latent process that induces dependence in the observed data. For instance, consider a generalized linear model (GLM) framework such that the data y_i for $i = 1, \ldots, n$ are modeled as

$$y_i \sim f(\theta_i, \phi_i), \tag{1.5}$$

where f denotes a distribution from an exponential family, θ_i denotes the natural parameter, and ϕ_i is a scale parameter (Nelder and Wedderburn, 1972). It is assumed that $\mathrm{E}(y_i) = \mu_i$ and that $g(\mu_i) = \eta_i$ for an appropriately specified link function g and linear predictor η_i. Note that choosing g to be the "canonical link function" results in $g(\mu_i) = \theta_i$ (Agresti, 2012) but that other suitable link functions are possible. The linear predictors η_i are modeled as a function of predictor variables $\mathbf{x}_i$ and coefficients $\boldsymbol{\beta}$ such that $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$. Consequently, GLMs allow for linear models to be applied to data that cannot be assumed to follow a normal distribution. To include spatial dependence, this model can be modified by assuming that $\eta(\mathbf{s})$ is a Gaussian process so that

$$\eta(\mathbf{s}) \sim \mathrm{GP}(\mathbf{x}(\mathbf{s})'\boldsymbol{\beta}, K), \tag{1.6}$$

where the observed predictor variables are spatially indexed and each observation y_i has a corresponding spatial location $\mathbf{s}_i$. This is one way to use spatial Gaussian processes when modeling other types of data (e.g, binary, counts, etc.).

There are additional methods that also make use of Gaussian processes for modeling nonnormal spatial data. For example, binary spatial data can be modeled using a clipped Gaussian process (De Oliveira, 2000, 2020). The models developed in Chapters 3 and 4 build upon this approach — I introduce clipped Gaussian processes in what follows and provide additional details on this method in later chapters. First, a spatial GLM for binary data assumes that

$$y(\mathbf{s}_i) \sim \text{Bernoulli}(p(\mathbf{s}_i)), \tag{1.7}$$

$$\Phi^{-1}(p(\mathbf{s}_i)) = \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} + \eta_1(\mathbf{s}_i), \tag{1.8}$$

where $\eta_1(\mathbf{s}) \sim \text{GP}(0, K_1)$ and we have assumed a probit link function with the cumulative distribution function of a standard normal distribution denoted by $\Phi(\cdot)$. To make the connection to clipped Gaussian processes more clear, we can reparameterize the model in (1.7) and (1.8) using an auxiliary variable approach (Albert and Chib, 1993). This parameterization of the spatial GLM is

$$y(\mathbf{s}_i) = \mathbb{1}(\tilde{y}(\mathbf{s}_i) \geq 0), \tag{1.9}$$

$$\tilde{y}(\mathbf{s}_i) \sim \text{N}(\mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} + \eta_1(\mathbf{s}_i), 1), \tag{1.10}$$

where $\mathbb{1}(\cdot)$ denotes an indicator function and $\eta_1(\mathbf{s}) \sim \text{GP}(0, K_1)$ as before. Note that (1.10) could also be rewritten as

$$\tilde{y}(\mathbf{s}_i) = \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} + \eta_1(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), \tag{1.11}$$

where $\epsilon(\mathbf{s}_i) \overset{iid}{\sim} \text{N}(0, 1)$ (Berrett and Calder, 2016; De Oliveira, 2020). The independent error terms $\epsilon(\mathbf{s}_i)$ are analogous to including a "nugget" effect in traditional geostatistical models (Diggle et al., 1998). The model defined by (1.9) and (1.11) shows how the spatial GLM can implemented by clipping a spatial Gaussian process that includes a nugget effect. Alternatively, consider another

clipped Gaussian process model that instead assumes

$$y(\mathbf{s}_i) = \mathbb{1}(\tilde{y}(\mathbf{s}_i) \geq 0), \tag{1.12}$$

$$\tilde{y}(\mathbf{s}_i) = \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} + \eta_2(\mathbf{s}_i), \tag{1.13}$$

where $\eta_2(\mathbf{s}) \sim \text{GP}(0, K_2)$. If this model includes a nugget effect in the spatial covariance function K_2 for $\eta_2(\mathbf{s})$, then there is a spatial GLM (assuming a probit link is used) that is equivalent to it (De Oliveira, 2020), as shown by the parameterization in (1.9) and (1.11). That is, equivalent models exist when

$$K_2(\mathbf{s}_i, \mathbf{s}_{i'}) = K_1(\mathbf{s}_i, \mathbf{s}_{i'}) + \tau^2 \mathbb{1}(\mathbf{s}_i = \mathbf{s}_{i'}) \tag{1.14}$$

and $\tau^2 = 1$, but these models are distinct if $\tau^2 = 0$ (De Oliveira, 2020). I refer to models as clipped Gaussian processes when they do not include a nugget effect to distinguish them from spatial GLMs.

While these models are closely related, the clipped Gaussian process and spatial GLM lead to binary data with different characteristics (Figure 1.1; see also Berrett and Calder, 2016). In particular, the clipped Gaussian process results in well-defined boundaries between regions of the different binary classes. This means that every spatial location $\mathbf{s}$ where $y(\mathbf{s}) = 1$ belongs to a contiguous region $\mathcal{A}$ such that $y(\mathbf{s}') = 1$ for all $\mathbf{s}' \in \mathcal{A}$ (e.g., Figure 1.1a). This is not the case for data generated from a spatial GLM and there are no well-defined boundaries between the binary classes even if the data are collected at a very fine resolution. Distinct regions are expected in the binary maps for many applications. For instance, maps of plant cover (see Chapter 3) have this characteristic due to individual plants defining the occurrence of the species over a study area. Individual plants must correspond to a contiguous area where the species occurs. This same idea applies to species distribution maps (see Chapter 4), home ranges of individuals, or maps of areas burned by wildfires (Yoo and Wikle, 2024). While I focus on clipped Gaussian processes for binary spatial processes in Chapters 3 and 4, other types of data can be modeled with this approach as well (e.g., ordinal or mixed continuous and ordinal; Higgs and Hoeting, 2010; Schliep and Hoeting, 2013).

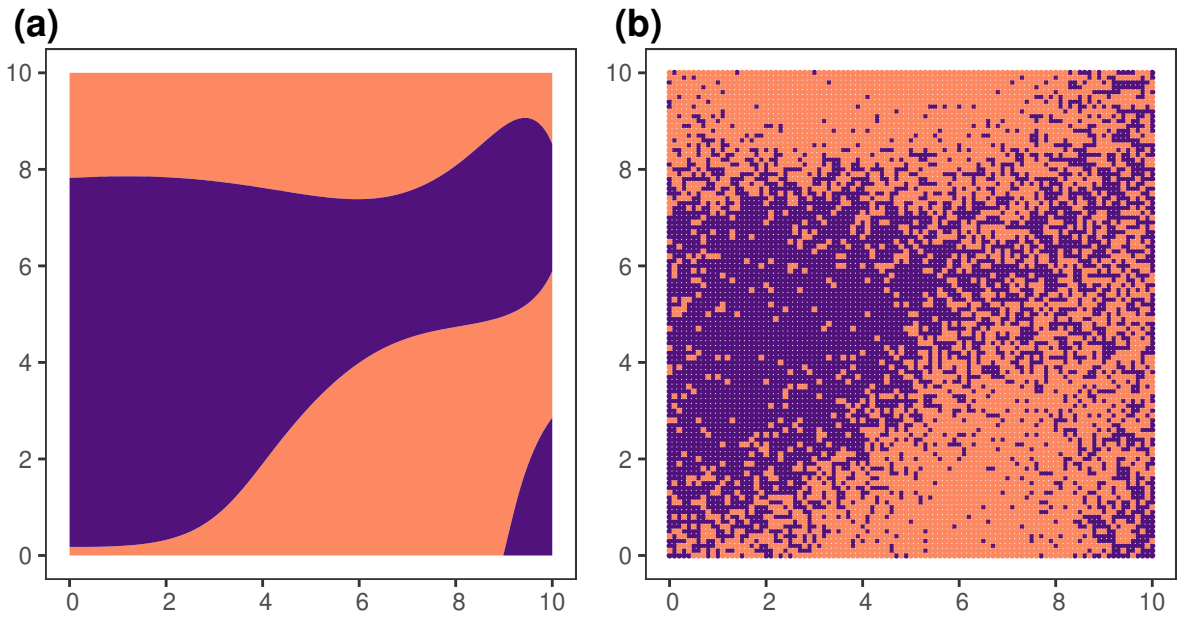

Figure 1.1: Example binary data from a clipped Gaussian process (a) and spatial GLM (b). The clipped Gaussian process results in well-defined boundaries that separate contiguous regions of each binary class (denoted by different colors). This is not the case for the data generated from a spatial GLM. Even if the spatial GLM data are observed at finer resolutions, these data will never result in well-defined boundaries between the binary classes.

Implementing Gaussian process models can be computationally expensive as the sample size increases. For $y(\mathbf{s}) \sim \text{GP}(m(\mathbf{s}), K)$, the finite-dimensional distribution of $\mathbf{y} \equiv (y(\mathbf{s}_1), \ldots, y(\mathbf{s}_n))'$ is multivariate normal and the computation times of the corresponding probability density function are on the order of n^3 because evaluating the density requires inversion of a $n \times n$ matrix. Consequently, there are a variety of methods available to improve the computational efficiency of spatial models (for a review, see Heaton et al., 2019). In general, these methods involve approximating the Gaussian process of interest and/or the associated covariance matrix. For example, these methods include basis function representations (fixed rank kriging, Cressie and Johannesson, 2008), predictive processes (Banerjee et al., 2008), using sparse covariance matrices (e.g., Sang et al., 2011), and nearest neighbor approximations (Datta et al., 2016; Datta, 2022), among others. In Chapter 3 I show how multiscale Gaussian processes can be approximated by introducing sparsity in the

corresponding covariance matrices when modeling plant cover data and in Chapter 4 I use nearest neighbor Gaussian processes when modeling wildlife occurrence data.

1.3.2 Mechanistic models

Spatial models that rely on Gaussian processes are flexible, and thus generally applicable, but typically do not allow for additional scientific knowledge to be included in analyses. Conversely, spatio-temporal models are increasingly incorporating characteristics of known physical processes (Wikle et al., 2001; Wikle and Hooten, 2010) that are ignored by phenomenological approaches. These mechanistic statistical models have proven useful for analyzing spatio-temporal data because the model structure is based on the scientific knowledge about a system (Wikle and Hooten, 2010) and allows for known constraints of physical processes to be reflected in a data analysis (Wikle et al., 2001). Mechanistic statistical models have been used for analyzing data from a wide variety of applications (e.g., Wikle et al., 2001; Wikle, 2003; Hooten and Wikle, 2008; Liu et al., 2016; Hefley et al., 2017c; Lu et al., 2020). The core idea of these mechanistic models is that the spatio-temporal process of interest can be described using a partial differential equation (PDE) for the dynamics of pertinent physical processes. Including this PDE in a statistical model can account for the dependence in the observed data in place of a general spatio-temporal covariance function.

Spatial data can also be viewed as arising from spatio-temporal processes (e.g., Hanks, 2017; Hefley et al., 2017b; Wikle et al., 2022), but mechanistic models are rarely used to analyze spatial data. However, there are connections between traditional spatial models based on Gaussian processes and mechanistic spatio-temporal models. For instance, Hanks (2017) showed that the Matérn covariance function — which is commonly used in spatial statistics (Stein, 1999) — is the covariance function for the stationary distribution of a spatio-temporal stochastic PDE. Similarly, Lindgren et al. (2011) used stochastic PDEs to define spatial covariance functions, including for non-stationary covariance functions, with computationally efficient algorithms. These methods highlight the potential links between commonly used phenomenological approaches that model spatial dependencies using a Gaussian process and mechanistic models that rely on PDEs to describe

spatial-temporal processes. However, Hefley et al. (2017b) point out that using the stationary characteristics of stochastic PDEs can fail to capture some aspects of the modeled processes that could be of scientific interest — such as the initial conditions and time period over which the dynamics occur. In Chapter 2, I develop a mechanistic spatial model that aims to directly incorporate aspects of known spatio-temporal processes to describe spatial patterns in observed heavy metal concentrations (also see Hefley et al., 2017b; Wikle et al., 2022).

1.4 Bayesian statistics

The models described in Chapters 2–4 are implemented using Bayesian methods. Consider a vector of observed data denoted $\mathbf{y}$ and vector of parameters $\boldsymbol{\theta}$. Bayesian statistics is concerned with learning about the conditional distribution of $\boldsymbol{\theta}$ given the observed data $\mathbf{y}$, which is referred to as the posterior distribution. By Bayes Theorem, the posterior distribution is

$$[\boldsymbol{\theta} \mid \mathbf{y}] = \frac{[\mathbf{y} \mid \boldsymbol{\theta}][\boldsymbol{\theta}]}{[\mathbf{y}]}, \tag{1.15}$$

where the notation $[\cdot]$ denotes a probability distribution function (Gelfand and Smith, 1990). The prior distribution $[\boldsymbol{\theta}]$ contains all knowledge about the parameters before the data have been observed. This knowledge is updated using information in the observed data and the assumed data distribution $[\mathbf{y} \mid \boldsymbol{\theta}]$. Consequently, Bayesian statistics relies on the posterior distribution for inferences about the parameters of interest. For instance, a possible point estimate is the posterior mean $\mathrm{E}(\boldsymbol{\theta} \mid \mathbf{y})$ and intervals can be defined using the quantiles of the posterior distribution.

Bayesian statistics provides a natural framework for statistical decision theory (Berger, 2013; Williams and Hooten, 2016) and can also be useful in practice for a variety of reasons. First, the prior distribution provides a natural way to include results from previous studies and/or knowledge about parameters into a statistical analyses. Because Bayesian analyses define a complete joint probability model, it is possible to make probability statements about parameters of interest directly (e.g., the probability that $\boldsymbol{\theta}$ is less than a particular value). Neither of these are possible using frequentist

or pure likelihood-based methods. Additionally, Bayesian statistics can naturally accommodate hierarchical structures and latent processes that are often of interest when modeling data from ecological and environmental applications.

1.4.1 Bayesian computation

A core component of any Bayesian data analysis is characterizing the target posterior distribution because it is required for all inferences about the parameters of interest. Analytical results for the posterior distribution are unavailable in most cases and thus computational implementations are typically required. The most common method in Bayesian statistics is to use Markov chain Monte Carlo (MCMC) to obtain samples from the posterior distribution (Geman and Geman, 1984; Gelfand and Smith, 1990; Tierney, 1994). With a large enough sample, these posterior draws can be used to approximate any summary measure of the posterior distribution that is of interest. For example, the posterior mean can be approximated as $T^{-1}\sum_{t=1}^{T}\boldsymbol{\theta}^{(t)}$ for posterior draws $\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(T)}$ and posterior quantiles can be approximated by the empirical quantiles of the posterior draws. A random sample from the joint posterior distribution also allows for inferences about marginal distributions and transformations of parameters — both of which are often of interest in practice. Inferences about the marginal posterior distribution for a single parameter (i.e., one element of $\boldsymbol{\theta} \equiv (\theta_1,\ldots,\theta_p)'$) can be obtained by only considering the posterior draws for that individual parameter. Similarly, the posterior mean and posterior credible intervals for a transformation of the parameters can be approximated by summarizing the transformed posterior draws. Consequently, inferences for transformed variables can be made without requiring asymptotic assumptions that are typically needed when using other approaches. Both of these properties make MCMC methods useful for implementing Bayesian models.

A primary challenge in Bayesian computation is to construct a valid MCMC algorithm that can sufficiently explore the target posterior distribution with a finite number of iterations. Foundational methods for Bayesian computation include the Gibbs sampler (Geman and Geman, 1984) and Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970). Gibbs sampling is

beneficial because it allows MCMC algorithms for large, hierarchical models to be implemented by sampling each parameter individually. These individual steps can be much simpler than obtaining posterior samples directly from the target posterior distribution. For instance, consider a Bayesian model with the parameters subset into two groups so that $\boldsymbol{\theta} \equiv (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$ and the target posterior distribution of interest is $[\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \mid \mathbf{y}]$. It may difficult to construct a valid MCMC algorithm that samples both parameter vectors jointly and has the correct stationary distribution. Alternatively, a Gibbs sampler considers updating each parameter vector individually from their respective full-conditional distributions. That is, each iteration first updates $\boldsymbol{\theta}_1$ by sampling from $[\boldsymbol{\theta}_1 \mid \mathbf{y}, \boldsymbol{\theta}_2]$ and then updates $\boldsymbol{\theta}_2$ by sampling from $[\boldsymbol{\theta}_2 \mid \mathbf{y}, \boldsymbol{\theta}_1]$. The parameter vectors are updated individually by conditioning on the observed data and the current value of the other parameter vector. These full-conditional distributions are often known analytically for specific prior distributions and this is referred to as "conjugate" updates. The Gibbs sampler naturally extends to more complicated models with more parameters and defines a valid MCMC algorithm assuming that the full-conditional distributions can be sampled from.

Other approaches are necessary when the posterior distribution or a full-conditional distribution cannot be sampled from directly. The Metropolis-Hastings algorithm is commonly used in these cases. Consider the example described above and updating $\boldsymbol{\theta}_1$ from its full-conditional distribution. At iteration $t+1$, the Metropolis-Hastings algorithm first proposes a possible new value for $\boldsymbol{\theta}_1$ from a user-defined proposal distribution $[\boldsymbol{\theta}_1^* \mid \boldsymbol{\theta}_1^{(t)}]$ where $\boldsymbol{\theta}_1^{(t)}$ is the current value for the parameter. This proposed value is accepted with probability $r = \min(1, mh)$ where

$$mh = \frac{[\mathbf{y} \mid \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^{(t)}][\boldsymbol{\theta}_1^*][\boldsymbol{\theta}_1^{(t)} \mid \boldsymbol{\theta}_1^*]}{[\mathbf{y} \mid \boldsymbol{\theta}_1^{(t)}, \boldsymbol{\theta}_2^{(t)}][\boldsymbol{\theta}_1^{(t)}][\boldsymbol{\theta}_1^* \mid \boldsymbol{\theta}_1^{(t)}]}. \tag{1.16}$$

If accepted, then $\boldsymbol{\theta}_1^{(t+1)}$ is set to the proposed value $\boldsymbol{\theta}_1^*$ and otherwise the chain remains at the current value $\boldsymbol{\theta}_1^{(t)}$. Different choices for the proposal distribution can simplify the resulting Metropolis-Hastings ratio (mh). The key idea is that appropriately specifying the proposal distribution can lead to highly efficient MCMC algorithms that explore the posterior distribution in relatively few

iterations. We make extensive use of both Gibbs sampling and the Metropolis-Hastings algorithm to implement the Bayesian models developed in Chapters 2–4.

Another common theme in Bayesian computation is that introducing additional parameters (or latent variables) into a model can lead to more efficient algorithms. The parameter expanded model must be constructed in such a way that the marginal prior distribution for the original parameter vector does not change. Consider slice sampling (Neal, 2003) as an example. For a generic Bayesian model (1.15), slice sampling introduces the auxiliary parameter u with the conditional distribution $u \mid \boldsymbol{\theta}, \mathbf{y} \sim \text{Uniform}(0, [\boldsymbol{\theta} \mid \mathbf{y}])$. The parameter expanded posterior distribution $[\boldsymbol{\theta}, u \mid \mathbf{y}]$ can be sampled from using a Gibbs sampler. First, u can be updated by sampling from the full-conditional distribution $[u \mid \boldsymbol{\theta}, \mathbf{y}]$ which is just a uniform distribution. Then $\boldsymbol{\theta}$ is updated from the full-conditional distribution

$$
\begin{aligned}
[\boldsymbol{\theta} \mid u, \mathbf{y}] &\propto [\mathbf{y} \mid \boldsymbol{\theta}][\boldsymbol{\theta}][u \mid \boldsymbol{\theta}, \mathbf{y}] && (1.17)\\
&\propto [\boldsymbol{\theta} \mid \mathbf{y}] \left(\frac{1}{[\boldsymbol{\theta} \mid \mathbf{y}]} \right) \mathbb{1}(u < [\boldsymbol{\theta} \mid \mathbf{y}]) && (1.18)\\
&\propto \mathbb{1}(u < [\boldsymbol{\theta} \mid \mathbf{y}]), && (1.19)
\end{aligned}
$$

which, as a function of $\boldsymbol{\theta}$, is a uniform distribution over the "slice" where the posterior density of interest $[\boldsymbol{\theta} \mid \mathbf{y}]$ is larger than u. Therefore, slice sampling allows for the posterior distribution to be sampled from using two Gibbs steps that both involve sampling from uniform distributions. While determining the slice in step 2 can be challenging, Neal (2003) provided algorithms to implement this update without explicitly identifying the slice. This approach can be particularly efficient when the vector $\boldsymbol{\theta}$ contains only a few parameters. However, this step is much more difficult to implement in high dimensions.

The Metropolis-Hastings algorithm can also be inefficient in high dimensions because small steps through the parameter space are required to have sufficiently high acceptance rates. This can lead to highly correlated chains that require many iterations to explore the target posterior distribution. For models using a normal prior distribution, elliptical slice sampling (Murray et al.,

2010) provides an approach for addressing this limitation. It can be particularly effective for spatial Gaussian processes because the prior distribution for any finite collection of the spatial terms is a multivariate normal distribution. To introduce elliptical slice sampling, consider the model

$$\boldsymbol{\theta} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{1.20}$$

$$\mathbf{y} \mid \boldsymbol{\theta} \sim [\mathbf{y} \mid \boldsymbol{\theta}], \tag{1.21}$$

where the data model $[\mathbf{y} \mid \boldsymbol{\theta}]$ can be any distribution and we are interested in inferences about $\boldsymbol{\theta}$. Here we are considering $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ fixed, but Chapter 4 illustrates a different approach that relaxes this assumption. The model defined by (1.20) and (1.21) provides inferences for $\boldsymbol{\theta}$ that are equivalent to those from the model

$$\alpha \sim \mathrm{Uniform}(0, 2\pi), \tag{1.22}$$

$$\boldsymbol{\nu}_1 \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{1.23}$$

$$\boldsymbol{\nu}_2 \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{1.24}$$

$$\boldsymbol{\theta} = (\boldsymbol{\nu}_1 - \boldsymbol{\mu})\mathrm{sin}(\alpha) + (\boldsymbol{\nu}_2 - \boldsymbol{\mu})\mathrm{cos}(\alpha) + \boldsymbol{\mu}, \tag{1.25}$$

$$\mathbf{y} \mid \boldsymbol{\theta} \sim [\mathbf{y} \mid \boldsymbol{\theta}], \tag{1.26}$$

where $\alpha, \boldsymbol{\nu}_0, \boldsymbol{\nu}_1$ are not identifiable because there are infinite combinations of these parameters that lead to the same data model. This parameter expanded version of the model is beneficial because it can be sampled from using a convenient blocked Gibbs sampler while still preserving the target marginal posterior distribution for $\boldsymbol{\theta}$. The marginal posterior distribution remains unchanged because for any fixed α, the marginal prior distribution for $\boldsymbol{\theta}$ is the same as (1.20). This means we can still use the parameter expanded model for making inferences about $\boldsymbol{\theta}$. While it may not be immediately obvious that this is helpful, the following blocked Gibbs sampler can be used to implement this model in two efficient steps. The first step is to update $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$ from the full-conditional distribution $[\boldsymbol{\nu}_1, \boldsymbol{\nu}_2 \mid \mathbf{y}, \alpha, \boldsymbol{\theta}] \stackrel{d}{=} [\boldsymbol{\nu}_1, \boldsymbol{\nu}_2 \mid \alpha, \boldsymbol{\theta}]$. Because $\boldsymbol{\theta}$ is a linear combination of

two normally distributed random variables, $[\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \boldsymbol{\theta} \mid \alpha]$ is jointly normal and the full-conditional of interest is also normal. Thus, the first step can be completed by directly sampling from the appropriate normal distribution. The second Gibbs update for elliptical slice sampling is to update α and $\boldsymbol{\theta}$ from the full-conditional distribution $[\alpha, \boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2]$ which is equal in distribution to $[\alpha \mid \mathbf{y}, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2][\boldsymbol{\theta} \mid \alpha, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2]$. Noting that $\boldsymbol{\theta}$ follows a degenerate distribution conditional on the other parameters, the second step can be completed by updating α from $[\alpha \mid \mathbf{y}, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2]$. This update can be performed efficiently using a univariate slice sampler (Neal, 2003).

Elliptical slice sampling can be implemented with a more concise algorithm (see Algorithm 2 in Murray et al., 2010) but the presentation above emphasizes that it is a blocked Gibbs sampler for a parameter expanded model. Nishihara et al. (2014) and Fagan et al. (2016) describe a generalization of elliptical slice sampling that allows for nonnormal prior distributions to be used. These approaches can require extensive tuning and the mixing of MCMC can be sensitive to this tuning (Nishihara et al., 2014; Fagan et al., 2016). I demonstrate how elliptical slice sampling can be used to fit spatial models and also consider other parameter expanded models for implementing models that allow the spatial covariance parameters to vary in Chapter 4.

1.5 Overview

In Chapters 2 and 3, I develop two novel spatial models that are motivated by datasets on heavy metal concentrations and vegetation communities from Cape Krusenstern National Monument, Alaska, USA. These data are of interest to the National Park Service because trucks hauling mining ore through the national monument spread heavy metal pollution throughout the region. The first approach characterizes the spatial structure in observed heavy metal concentrations using a spatio-temporal process for atmospheric dispersion. Mathematically, this is modeled using an advection-diffusion partial differential equation that incorporates information about pollutant sources, diffusion, duration of spread, and advection from prevailing winds. My mechanistic spatial model is beneficial because it allows for learning about the physical processes spreading pollutants and provides a way to predict pollutant concentrations under scenarios where the source

rate changes. For instance, I use this model to forecast how heavy metal concentrations throughout Cape Krusenstern could change if the amount of pollution emitted from the road changes in the future. Next, I develop a novel spatial model for multivariate binary data that includes spatial dependence at multiple scales. Plant cover data are commonly collected using the point intercept method which records binary data for whether a species is present or absent at a grid of points overlaying sampled plots. By clipping a multiscale spatial Gaussian process, my model captures spatial dependence both within a plot and among different plots. I also include spatial latent factors to model the dependence among species. I show how a hierarchical formulation of this model is useful for approximating the overall spatial covariance structure and derive the constraints needed for implementation. I apply this model to the plant cover data from Cape Krusenstern to assess the impacts of heavy metal pollution on the vegetation communities there.

Chapter 4 develops a statistical model for the occurrence of wildlife species in continuous space. A primary pursuit in ecology is to understand how species are distributed across the landscape. Occupancy models are often used to address questions related to species distributions and are beneficial because they allow researchers to account for the imperfect detection of species at surveyed sites. While the occurrence of a species is generally believed to be a process defined for continuous space, standard occupancy analyses assume a discrete spatial domain because the observed data are conducted at areal sites. I develop the first occupancy model that accounts for the change of spatial support between the detection/nondetection data and species occurrence process. My approach allows occupancy to be modeled in continuous space even though the observed data are collected in discrete space. This framework assumes that species occurrence can be modeled by clipping a latent Gaussian process, similar to the model for plant cover in Chapter 3. Additionally, this model is able to account for imperfect detection and also model how the detection probability within a site may be related to the proportion of the site that is occupied. To implement my continuous-space occupancy model using Bayesian methods, I also develop a surrogate data slice sampler that improves the computational efficiency of fitting models with latent nearest neighbor

Gaussian processes. I demonstrate this approach using ovenbird (*Seiurus aurocapilla*) data collected at Hubbard Brook Experimental Forest, New Hampshire, USA.

Chapter 2

Mechanistic spatial models for heavy metal pollution

2.1 Introduction

Mining operations can contribute substantial amounts of pollution in the form of atmospheric dust during the extraction and transportation of ore (see Csavina et al., 2012). Heavy metals are often found in mining pollution and can have detrimental effects on both environmental and human health (Briffa et al., 2020; Mishra et al., 2019). Consequently, statistical models predicting the spread of heavy metal pollutants from known sources are useful for evaluating the impacts of mining activities. Even though the spread of pollution in the atmosphere is governed by known physical processes, statistical models for heavy metal pollution data typically ignore these mechanisms when accounting for spatial structure in data. For instance, geostatistical models are often used to analyze data on heavy metal concentrations in the environment (e.g., Hasselbach et al., 2005; Neitlich et al., 2017; Donovan et al., 2016; Ersoy et al., 2004; Reza et al., 2015) but this approach only provides phenomenological descriptions of observed spatial patterns. We explore the benefits of incorporating physical processes in a mechanistic statistical model for analyzing spatial data on heavy metal pollution from mining activity.

We focus our study on heavy metal concentrations in Cape Krusenstern National Monument (CAKR), Alaska, USA. Previous work found elevated concentrations of heavy metals within CAKR resulting from the transportation of ore from the Red Dog Mine (Hasselbach et al., 2005; Neitlich et al., 2017) and there is evidence that the heavy metal pollution is negatively impacting the plant and animal communities within the monument (e.g., Brumbaugh et al., 2010, 2011). Consequently, the National Park Service seeks to monitor heavy metal concentrations over time and understand the mechanisms spreading heavy metal pollution in the region. This information may be useful for predicting how potential changes to mining operations may alter pollution levels within CAKR. Earlier analyses of these data used geostatistical models that included distance from the road, side

of the road (north or south), and an interaction term between distance and road side as fixed effects (Hasselbach et al., 2005; Neitlich et al., 2017). However, Hasselbach et al. (2005) and Neitlich et al. (2017) noted that it may be helpful to use information about mechanisms governing the spread of pollutants in this region, such as prevailing winds and the topography, in future studies. While some of the fixed effects included in their analyses were motivated by these mechanisms (Hasselbach et al., 2005; Neitlich et al., 2017), additional information about the processes that spread heavy metal pollutants throughout CAKR is available. We develop a mechanistic statistical model to better incorporate this scientific knowledge in the analysis of these data.

Our mechanistic model for the CAKR heavy metal data includes aspects of the physical process for dispersing atmospheric pollution. Mathematically, the spatio-temporal dynamics of airborne pollutants can be modeled using an advection-diffusion partial differential equation (PDE) for atmospheric dispersion (Stockie, 2011). We use this spatio-temporal process to characterize spatial structure in our statistical model. This allows for information about pollutant sources, diffusion, duration of spread, and predominant advection forces (e.g., wind) to be incorporated into our analysis. Mechanistic statistical models have been used to describe spatio-temporal dynamics found in a number of applications including the expanding ranges of wildlife populations (e.g., Wikle, 2003; Hooten and Wikle, 2008; Lu et al., 2020), the spread of wildlife diseases (e.g., Hefley et al., 2017c), patterns of tropical ocean winds (Wikle et al., 2001), and ozone pollution in the atmosphere (Liu et al., 2016). However, mechanistic models are rarely used to analyze spatial data even though spatial data can also be viewed as arising from spatio-temporal processes (e.g., Hanks, 2017; Hefley et al., 2017b; Wikle et al., 2022). Our approach is more flexible than previous mechanistic models for spatial data (Hefley et al., 2017b; Wikle et al., 2022), because it includes temporally varying advection and links indirect concentration measurements to the spatio-temporal dynamics of the model. Additionally, our model provides similar predictive inferences when compared to a geostatistical model, but also allows us to make inference on parameters describing the spread of heavy metal pollutants. Such inference is needed for forecasting future concentration levels and evaluating the potential environmental impacts of changes in pollutant emission rates.

We describe the CAKR heavy metal data in Section 2.2. In Section 2.3, we present the atmospheric dispersion model that we use to characterize the spatial structure in our analysis. We describe our statistical model and algorithm for fitting it in Section 2.4. Section 2.5 presents the results from analyzing the CAKR heavy metal data and we end with a discussion of the benefits of this approach and considerations for future work in Section 2.6.

2.2 Data

The development of our model was motivated by data on chemical concentrations of cadmium (Cd), lead (Pb), and zinc (Zn) collected at 118 locations throughout CAKR in northwest Alaska (Figure 2.1). Concentrations of these heavy metals are elevated in this area as a result of pollution from the transportation of ore along the Red Dog Mine haul road through northern CAKR (Hasselbach et al., 2005; Neitlich et al., 2017). The Red Dog Mine is a large zinc mine located approximately 50 kilometers northeast of the CAKR boundary. Trucks transport ore along the haul road to a port on the Chukchi Sea coast where it is stored before being shipped to other locations. To understand the impacts of mining activities on the ecosystem in this region, the National Park Service is interested in estimating concentrations of Cd, Pb, and Zn along the road and monitoring how these chemical concentrations change over time.

Heavy metal pollutants in CAKR are primarily spread through the air in fugitive dust shed from trucks transporting ore along the haul road. Heavy metal particles in these dusts are deposited onto the ground and vegetation over time. The observations we analyzed are chemical concentrations (mg/kg) of feather moss (*Hylocomium splendens*) samples. In general, observed concentrations are highest near the road and decrease as distance to the road increases (Figure 2.2). Because feather moss has no vascular system, any chemical pollutants in these samples are the result of atmospheric deposition and not due to absorption from ground sources (also see Hasselbach et al., 2005; Berg and Steinnes, 1997). The samples included feather moss tissue that was approximately five years old or younger. Therefore, each observation corresponds to the accumulation of heavy metals at a

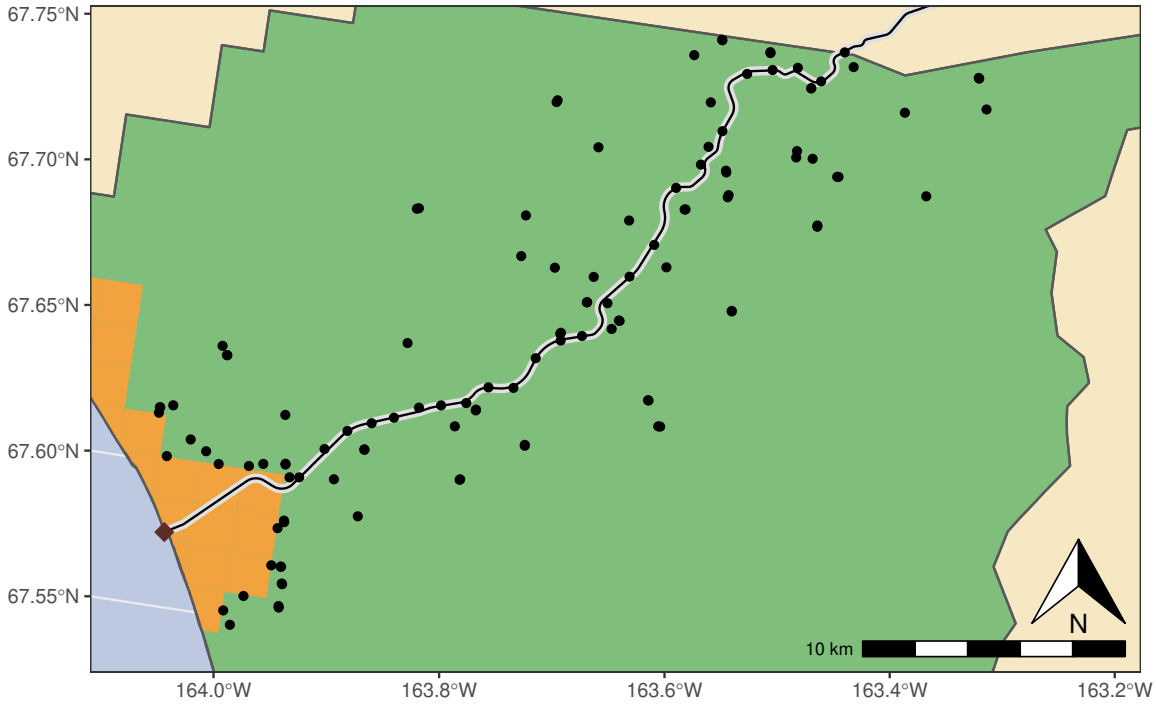

Figure 2.1: Map of the study area in Cape Krusenstern National Monument (green). Points indicate the 118 sampled locations along the Red Dog Mine Haul road (gray line). Ore from the Red Dog Mine (located northeast of the study area, outside the mapped area) is transported along the road to a port (brown diamond) where it is stored before being shipped. Lands around the port are owned by the NANA Alaska Native corporation (orange).

particular location over a five year period. These data were originally analyzed by Hasselbach et al. (2005) and more information about the data collection and sample processing can be found there.

2.3 Atmospheric dispersion model

The atmospheric dispersion of chemical pollutants is often modeled deterministically using an advection-diffusion PDE (Stockie, 2011). We incorporated elements of this physical process in our statistical model to improve our understanding of how heavy metal pollution is spread along the Red Dog Mine haul road in CAKR. Let $\lambda(\mathbf{s}, t)$ denote the chemical concentration at spatial location $\mathbf{s} \equiv (s_1, s_2)'$ and time t. We suppress the notation for different chemicals in this section, but we allowed the parameters describing this PDE to vary across the three chemicals included

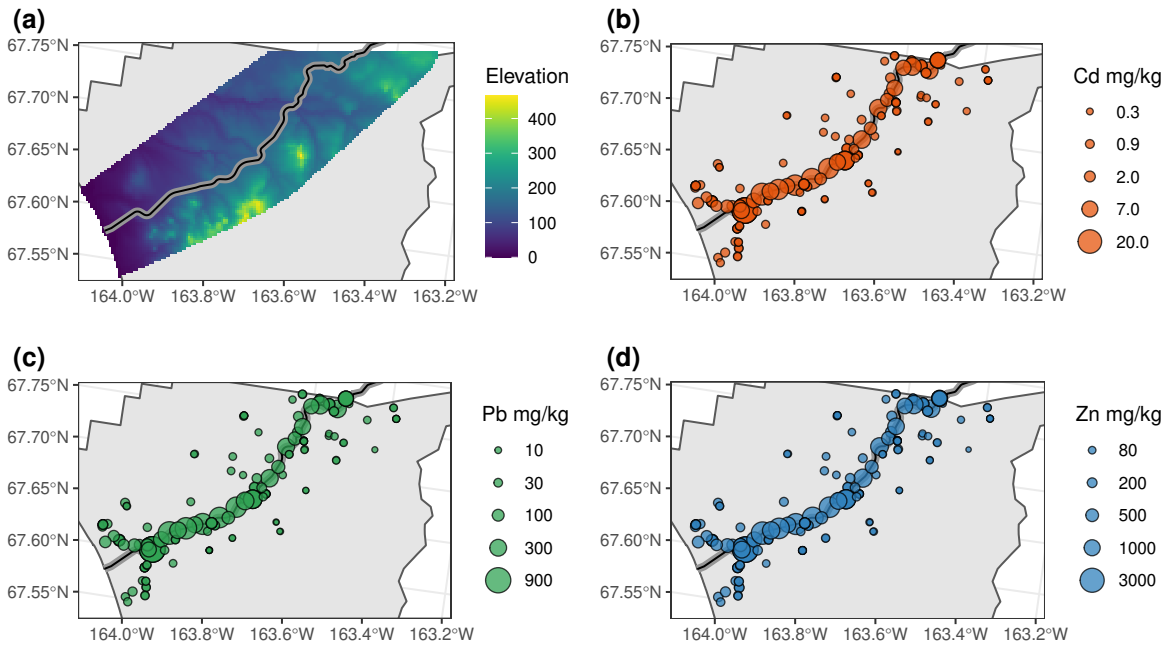

Figure 2.2: Elevation across the study area in CAKR (a). Observed concentrations (mg/kg) of Cadmium (Cd, b), Lead (Pb, c), and Zinc (Zn, d) at sampled locations.

in our analysis. To model the spread of chemical pollutants from the haul road, we used the advection-diffusion equation

$$\frac{\partial \lambda(\mathbf{s},t)}{\partial t} = \eta(\mathbf{s}) + \nabla \cdot [\delta(\mathbf{s}) \nabla \lambda(\mathbf{s},t)] - \nabla \cdot [\lambda(\mathbf{s},t)\mathbf{w}(t)] - \tau\lambda(\mathbf{s},t), \tag{2.1}$$

where $\eta(\mathbf{s})$ is the source term for the pollutant at location $\mathbf{s}$, $\delta(\mathbf{s})$ is the spatially varying diffusion coefficient, $\mathbf{w}(t)$ is the temporally varying velocity corresponding to the strength and direction of advection, $\nabla\cdot$ is the divergence operator, and ∇ is the gradient operator.

We assumed spatial locations along the road were a continuous source of chemical pollutants. See Section 2.4.1 on implementing our statistical models for more details. Because ore is stored at the port location, previous studies suggested that the port site is a larger source of chemical pollutants than other locations along the road (Hasselbach et al., 2005; Neitlich et al., 2017). We accounted for this by allowing the source term to vary spatially. For spatial locations along the road,

we assumed

$$\log(\eta(\mathbf{s})) = \theta_0 + \theta_1 d(\mathbf{s}) + \theta_2 \mathbb{1}_{\text{Port}}(\mathbf{s}), \tag{2.2}$$

where $d(\mathbf{s})$ denotes the road distance (km) to the port site for spatial location $\mathbf{s}$ and $\mathbb{1}_{\text{Port}}(\mathbf{s})$ is an indicator variable for whether spatial location $\mathbf{s}$ is within 2 km of the port or not. The structure in (2.2) allows the road source to change as the distance to port increases and accounts for the possibility that the port could also be a larger source of pollution because ore is stored there. This functional form incorporates knowledge of how pollutant emissions could vary along the road using directly interpretable parameters. Our model assumes that the source component is continuous in time because we have no data to inform potential temporal variability in the source and are ultimately interested in the accumulation of pollution over time (see Section 2.4).

The second component on the right-hand side of (2.1) represents the spatial diffusion of pollutants over time. When the diffusion coefficient $\delta(\mathbf{s})$ varies by spatial location, the PDE in (2.1) is often called 'Fickian' diffusion. We modeled the diffusion coefficient as

$$\log(\delta(\mathbf{s})) = \beta_0 + \beta_1 x(\mathbf{s}), \tag{2.3}$$

where $x(\mathbf{s})$ denotes the elevation at spatial location $\mathbf{s}$. This structure accounts for the variability in diffusion due to the topography at a location.

The next component of the advection-diffusion equation (2.1) describes the change in chemical concentration at a location due to advection (or 'drift'). We modeled the advection velocity as $\mathbf{w}(t) = \gamma \mathbf{a}(t)$ where $\mathbf{a}(t)$ is the wind vector at time t and $\gamma > 0$ is a scaling parameter to be estimated. Wind vector information was obtained from the National Oceanic and Atmospheric Administration website in the form of daily average u and v components from the NCEP–DOE Reanalysis 2 project (Kanamitsu et al., 2002). Using these u and v components, the wind vector at time t is $\mathbf{a}(t) \equiv (u(t), v(t))'$ and the penultimate term of (2.1) can be rewritten as

$$\nabla \cdot [\lambda(\mathbf{s}, t)\mathbf{w}(t)] = \gamma u(t) \frac{\partial \lambda(\mathbf{s}, t)}{\partial s_1} + \gamma v(t) \frac{\partial \lambda(\mathbf{s}, t)}{\partial s_2}. \tag{2.4}$$

Previous studies at CAKR found that the chemical concentrations along the north side of the road were higher than those on the south side (Hasselbach et al., 2005; Neitlich et al., 2017). The prevailing winds to the west during winter months (8 months of the year) and prevailing winds to the east during summer months (4 months of the year) may lead to these observed differences (Neitlich et al., 2017). We accounted for the potential drift in chemical pollutants due to these wind patterns using the advection component of this PDE (2.1). Note that we allowed the wind vector to vary in time but assumed spatial homogeneity in the wind vector at any time t.

The final component of (2.1) represents the deposition of chemical pollutant from the atmosphere onto vegetation (feather moss for these data). We assume that the rate of deposition is related to the current chemical concentration at a location and denoted the corresponding deposition parameter as τ. The advection-diffusion PDE describes the concentration of chemical pollutant in the *atmosphere* and how it varies in space and time. However, the observed data describe the accumulation of chemical pollutant on *vegetation* (feather moss) over a five year time period. The deposition component of the atmospheric dispersion model links the chemical concentration in the atmosphere, where the transportation of pollution occurs, to the concentrations observed in these data.

2.4 Data model

We fit models to analyze the chemical concentrations of Cd, Pb, and Zn collected in CAKR during 2001. Our hierarchical model was jointly specified for these three chemicals to allow some of the parameters describing the advection-diffusion PDE in (2.1) to be related across the different chemicals. Let $j = 1, 2, 3$ index Cd, Pb, and Zn, respectively. We let $y_j(\mathbf{s}_i)$ denote the log-concentration of chemical pollutant j at spatial location $\mathbf{s}_i$ for $i = 1, \ldots, n$. For each heavy

metal, we modeled the log-concentration observations as

$$y_j(\mathbf{s}_i) \sim \text{Normal}(\mu_j(\mathbf{s}_i), \sigma_j^2), \tag{2.5}$$

$$\mu_j(\mathbf{s}_i) = \log(\alpha_j + \tilde{\lambda}_j(\mathbf{s}_i)), \tag{2.6}$$

$$\tilde{\lambda}_j(\mathbf{s}_i) = \int_0^T \tau_j \lambda_j(\mathbf{s}_i, t) dt, \tag{2.7}$$

where $\lambda_j(\mathbf{s}, t)$ for each chemical is described by the advection-diffusion equation in Section 2.3. By assuming a constant variance on the log scale, this model assumes a measurement error structure that is consistent with earlier studies (Hasselbach et al., 2005; Neitlich et al., 2017). The parameter α_j in (2.6) corresponds to a fixed "baseline" concentration for each chemical that is not attributed to pollution from the haul road. These baseline concentrations are assumed to be constant over space and time throughout this region.

The integration over time in (2.7) accounts for the accumulation of chemical pollution over the time interval 0 to T which occurs at deposition rate τ_j. As described in Section 2.3, this allows the mechanistic model for the transportation of pollutants in the atmosphere to be linked to the observed concentrations in the moss tissue samples. The chemical concentrations for these data are the result of approximately five years of deposition, because the collected feather moss tissue was at most five years old, and we therefore assumed $T = 5$ years for this analysis. Quality control checks were in place to confirm moss tissue samples included the desired age range (see Hasselbach et al., 2005, for details). Variability in observed concentrations due to differences in the age distributions of sampled moss at a particular location is accounted for as measurement error variability.

Section 2.3 includes additional parameters for the advection-diffusion PDE (2.1) which describe the pollutant source at the road (2.2), diffusion (2.3), and advection (2.4) due to prevailing winds each season. We allowed these parameters to vary across the three chemicals, but modeled a set of them hierarchically to account for the expected similarity of the atmospheric dispersion process for each. We modeled the parameters describing diffusion (β_{0j}, β_{1j}), advection ($\log(\gamma_j)$), source term coefficient for port distance (θ_{1j}), and the additional source coefficient for the port (θ_{2j}) as each

arising from common population-level distributions. That is, we assumed that

$$\beta_{0j} \sim \text{Normal}(\mu_{\beta_0}, \sigma^2_{\beta_0}), \tag{2.8}$$

$$\beta_{1j} \sim \text{Normal}(\mu_{\beta_1}, \sigma^2_{\beta_1}), \tag{2.9}$$

$$\log(\gamma_j) \sim \text{Normal}(\mu_\gamma, \sigma^2_\gamma), \tag{2.10}$$

$$\theta_{1j} \sim \text{Normal}(\mu_{\theta_1}, \sigma^2_{\theta_1}), \tag{2.11}$$

$$\theta_{2j} \sim \text{Normal}(\mu_{\theta_2}, \sigma^2_{\theta_2}), \tag{2.12}$$

for $j = 1, 2, 3$ and where the corresponding population-level means and variances are additional parameters to be estimated. We assumed the remaining parameters (θ_{0j}, α_j, and σ^2_j) were independent across chemicals because these parameters are related to the overall concentrations that vary substantially for Cd, Pb, and Zn.

2.4.1 Priors and implementation

We used Bayesian methods to fit this model and assumed weakly informative priors for all parameters that represent a wide range of realistic processes (Appendix A.1). The model was fit to data using a Markov chain Monte Carlo (MCMC) algorithm. All parameters required Metropolis-Hastings updates except for the measurement error variance (σ^2_j) and those describing the hierarchical distributions (μ and σ^2 parameters in equations 2.8–2.12). We used multivariate normal proposal distributions when updating the parameters associated with the PDE (2.1) for each chemical because those parameters are correlated and to improve computation time. The variance-covariance matrices for these proposal distributions were tuned during the initial iterations (first 5000) of our MCMC algorithm based on the proposal acceptance rates and empirical variances of the posterior draws for each chemical. We fit the hierarchical model using a total of 10000 MCMC iterations with the last half (5000) used for inferences. Our analysis was conducted in R (version 4.1.0, R Core Team, 2021) using the `Rcpp` (Eddelbuettel and François, 2011; Eddelbuettel

and Balamuta, 2018) and `RcppArmadillo` (Eddelbuettel and Sanderson, 2014; Eddelbuettel et al., 2023) packages.

With only spatial data, the deposition parameters τ_j are not identifiable because they are proportional to the source term $\eta(\mathbf{s})$. For instance, identical accumulated concentrations ($\tilde{\lambda}_j(\mathbf{s})$) for all spatial locations can result from a different τ_j if $\eta(\mathbf{s})$ is scaled appropriately. We included the deposition component in our model by fixing $\tau_j = 1$ for all j. This allows our advection-diffusion process to include deposition, but we interpret the other parameters relative to the assumed deposition rate. An alternative would be to fix the source term for each chemical and then estimate the deposition parameters. We did not use this approach because it is more reasonable to assume the deposition rate is constant over time while changes in the mining activities could result in different source terms for future data.

At each iteration, the MCMC algorithm required solutions to the PDE (2.1) given the model parameters for each chemical. We used finite difference methods to approximate the PDE at each iteration (see Appendix A.2 for additional details). For this approximation, we defined a grid of $N = 7513$ cells (each 250 by 250 meters) and discretized the five year deposition period into 1826 time steps. We let $\boldsymbol{\lambda}_j(t)$ denote a vector of dimension N that contains $\lambda_j(\mathbf{s}, t)$ values corresponding to the spatial locations of the grid cell centroids at time t for chemical j. Similarly, we let $\boldsymbol{\eta}_j$ denote a vector of dimension N that contains $\eta_j(\mathbf{s})$ values for the same spatial locations as those included in $\boldsymbol{\lambda}_j(t)$. Then the finite difference approximation for our model can be represented in matrix form as

$$\boldsymbol{\lambda}_j(t+1) = \mathbf{H}_j(t)\boldsymbol{\lambda}_j(t) + (\Delta t)\boldsymbol{\eta}_j, \tag{2.13}$$

for $j = 1, 2, 3$ and where $\mathbf{H}_j(t)$ is a propagator matrix defined for the advection-diffusion PDE in our model (2.1) using finite difference approximations (see Appendix A.2) and Δt is the discrete time scale used in this approximation. The propagator matrix $\mathbf{H}_j(t)$ and the vector of concentrations $\boldsymbol{\lambda}_j(t)$ for our application are both sparse. We used sparse matrix operations to increase the computational efficiency when solving this PDE.

Solving the PDE also required us to specify initial and boundary conditions for this system. We assumed that $\lambda_j(\mathbf{s}, 0) = 0$ for all locations $\mathbf{s}$ at the initial time point (denoted as time $t = 0$). This initial condition reflects the assumption that newly grown moss tissue has no chemical pollution but that deposition of the chemical pollutants occur over time as the moss grows. We assumed absorbing boundary conditions for all boundary grid cells except for the northern edge where we assumed a no-flux boundary condition. When solving the PDE, we included grid cells within a 5 km buffer of the port to prevent the southwest boundary from being a pollutant sink. The absorbing boundary condition was assumed where it was possible for pollutants from the road to exit the study area, but at locations where it was unlikely that substantial amounts of pollution were also spreading into the study area. However, this is not necessarily the case for the northern border. Because the haul road continues north of the study area (Figure 2.1), concentrations of the chemical pollutants are elevated throughout that region as well. The no-flux northern boundary assumes that the amount of pollutant diffusing out of the study area to the north will be approximately equal to the amount diffusing south into the study area along this border. Therefore, this boundary condition ensures that the northern border of the study area does not behave as a pollutant sink in our model.

2.4.2 Predictions and forecast scenarios

One quantity of interest is the average concentration at locations that are certain distances from the haul road. For instance, Neitlich et al. (2017) predicted concentrations at locations throughout the study area and summarized these predictions for strata that were defined by the distance from each location to the road. These strata included categories for 0–100, 100–2000, 2000-4000, and 4000–5000 meters to the road. We used our model to predict concentrations of each heavy metal at these same locations and then compared the strata summaries to those reported by Neitlich et al. (2017).

We also summarized the posterior predictive distribution for the average Zn concentration by strata and side of road. We then used these summaries to illustrate how our model can provide forecasts for the average Zn concentrations under scenarios when the pollutant source is expected to

change. First, we considered a 50 percent decrease in the pollutant source which could correspond to a scenario where additional pollution mitigation strategies were put in place and expected to reduce fugitive dust from trucks by this amount. We also considered a scenario where the pollutant source increases by 50 percent — this could be expected if there was a proposal to increase the total amount of concentrate shipped from Red Dog mine by 50 percent. The posterior predictive distribution for these scenarios were calculated by appropriately adjusting the intercept term for the pollutant source θ_0. For example, a 50 percent decrease in the pollutant source corresponds to adding $\log(0.5)$ to θ_0. For each MCMC iteration, new $\tilde{\lambda}(\mathbf{s}_i)$ were found based on the adjusted pollutant source to summarize the posterior predictive distributions for these scenarios.

2.5 Results

The overall prevalence of Cd, Pb, and Zn differs in the region — both in terms of naturally occurring concentrations and pollution levels. Baseline concentrations (α_j) and road source levels (θ_{0j}) were estimated to be highest for Zn, followed by Pb next, and lowest for Cd (Table 2.1). The posterior distributions of the measurement error standard deviations (σ_j) were similar for each chemical (Table 2.1). The posterior distributions of the parameters describing diffusion (β_0, β_1), advection ($\log(\gamma)$), the source term coefficient for port distance (θ_1), and source coefficient for the port indicator (θ_2) were also similar across the different chemicals (Figure 2.3). These results suggest that while overall levels of each chemical pollutant differed, the spread of pollution from the road is governed by similar physical processes. By modeling this set of parameters hierarchically across chemicals in (2.8)–(2.12), our approach is able to account for the similarities in these processes.

Previous studies described how the observed differences in chemical concentrations across the study area could be related to the distance from the port, topography, and prevailing winds during different seasons. Our analysis incorporated these features in the advection-diffusion process and provided estimates of the corresponding parameters. For each chemical, we estimated that the pollutant source decreases as distance from the port increases and that the port itself is an additional source of pollutant (posterior distributions for each θ_{1j} and θ_{2j} in Figure 2.3). The

Table 2.1: Posterior means and 95% posterior intervals (equal-tailed) for the baseline concentration parameter α, the intercept coefficient for the source term θ_0, and the standard deviation of the measurement error σ. Results are shown for Cadmium (Cd), Lead (Pb), and Zinc (Zn).

Parameter	Chemical	Posterior mean	95% Posterior interval
α	Cd	0.488	(0.437, 0.558)
	Pb	14.7	(12.6, 19.0)
	Zn	91.9	(78.6, 119.5)
θ_0	Cd	3.60	(3.45, 3.79)
	Pb	7.30	(7.11, 7.457)
	Zn	8.64	(8.43, 8.81)
σ	Cd	0.424	(0.373, 0.512)
	Pb	0.490	(0.422, 0.650)
	Zn	0.430	(0.360, 0.582)

posterior distributions for β_{1j} (Figure 2.3) provide evidence that diffusion decreased as elevation increased.

We also obtained predictions of chemical concentrations across the study area based on this model. For each pollutant, the maps of predicted concentrations show highest concentrations closest to the road that decrease sharply as distance from the road increases (Figures 2.4, 2.5, 2.6). This general pattern was the primary motivation for modeling expected log-concentration as a function of log-distance to the haul road in the original analyses of these data (Hasselbach et al., 2005; Neitlich et al., 2017). However, our results show that this phenomenological pattern can also be explained by an advection-diffusion process like the one included in our analysis. Another feature in the maps of predicted concentrations is higher pollutant levels near the port location in the southwest of our study domain. This result is explained by the higher source rates near the port (due to source coefficients θ_1 and θ_2) and also the higher estimated rates of diffusion at lower elevations. Consequently, higher concentrations of each chemical are found farther from the road in this part of CAKR.

For Zn, the strata summaries from our model are similar to those reported in Neitlich et al. (2017) (Figure 2.7, 'Overall' column) and show a steep decline in concentrations as distance from the road increases. When making comparisons by strata and side of road, our posterior predictive

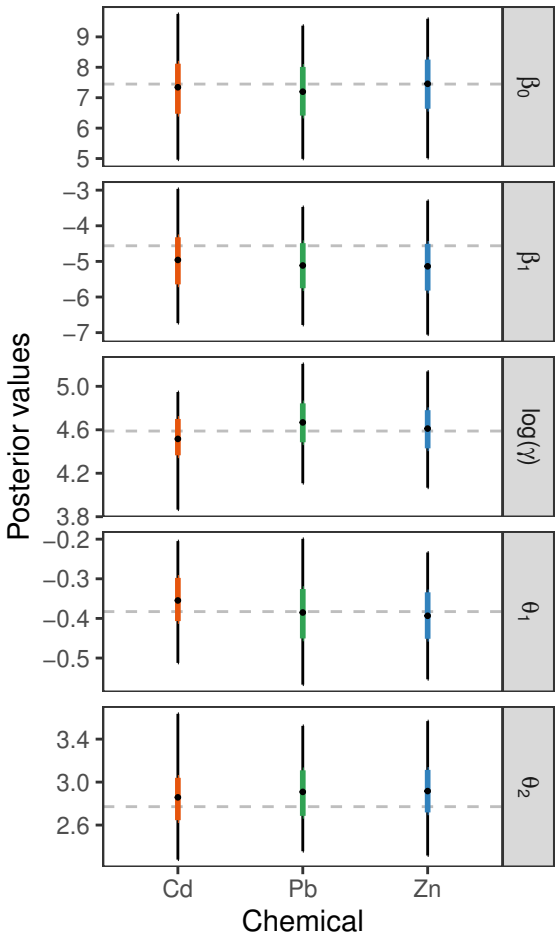

Figure 2.3: Posterior summaries of parameters for Cadmium (Cd), Lead (Pb), and Zinc (Zn) chemical concentrations. The posterior means are shown by points, 50% posterior intervals are shown by thick lines, and the 95% posterior intervals are shown by the thin lines. The posterior mean of the overall means for each set of parameters are indicated by the dashed horizontal lines.

distributions showed higher average concentrations for locations north of the road compared to those south of the road for the 0–100 m and 100–2000 m strata (Figure 2.7). This pattern was also observed in the predictions from the geostatistical model used by Neitlich et al. (2017). For the two strata farther from the road (2000–4000 and 4000–5000 m), our posterior predictive distributions for average Zn concentration do not show a large difference between the north and south sides of the road (Figure 2.7). This is likely due to the assumption that the background concentration does not vary across the study area. Some differences by side of road were reported for these strata by

Neitlich et al. (2017). Comparisons across models for Cd and Pb were similar to those found for Zn (results not shown).

Our mechanistic model also allows us to forecast chemical concentrations under different scenarios where source rates increased or decreased. In these scenarios, we found that the largest predicted changes in Zn concentrations occurred for locations closer to the road (Figure 2.7). For instance, locations in strata 1 (0–100 m) are predicted to have concentrations increase by approximately 50% if the source rate increased by 50%. However, there is a less drastic increase in predicted concentrations for locations in strata 2 (100–2000 m) under this scenario (Figure 2.7). Locations far from the road, on the other hand, show only small changes in predicted concentrations relative to those observed in 2001 if the pollutant source changes.

2.6 Discussion

We designed a mechanistic statistical model to estimate the concentrations of three heavy metal pollutants in CAKR by accounting for spatial structure that arose from a spatio-temporal process. The spatial structure for each chemical was governed by an advection-diffusion PDE for atmospheric dispersion. Consequently, our analysis was able to incorporate aspects of the physical dispersion of pollutants (e.g., road source, temporally varying wind, and topography) that had not been included in previous analyses (Hasselbach et al., 2005; Neitlich et al., 2017). For instance, our model allowed the diffusion coefficients to vary spatially and we inferred that diffusion decreases at higher elevations for all three chemicals. Along with the higher source rates closer to the port, this variation in diffusion rates helps explain why chemical concentrations are higher around the port site. By incorporating an advection-diffusion process in the statistical model, our analysis provides a better understanding of how these processes contribute to the spread of heavy metal pollutants in CAKR.

We estimated concentrations of Cd, Pb, and Zn jointly and modeled the parameters from the advection-diffusion PDE hierarchically across these chemicals. We found that the parameters describing diffusion and advection were similar for these chemicals, suggesting that the physical

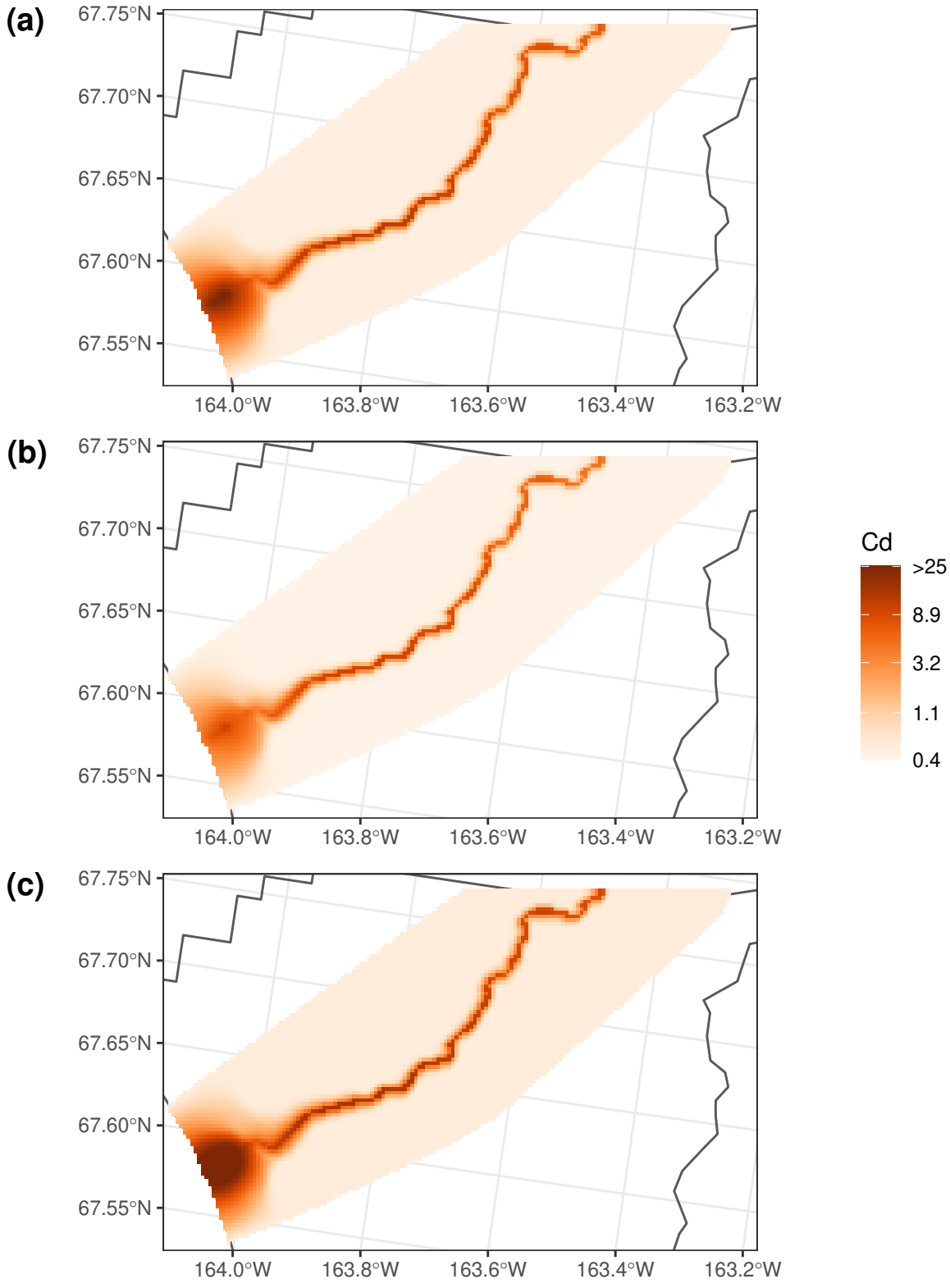

Figure 2.4: Maps summarizing the posterior distribution of Cadmium (Cd) concentrations on the original scale (i.e., $(\alpha_j + \tilde{\lambda}_j(\mathbf{s}))$) from our fitted model. These maps display the posterior means (a), point-wise 2.5% quantiles (b), and point-wise 97.5% quantiles (c) for locations throughout northern CAKR.

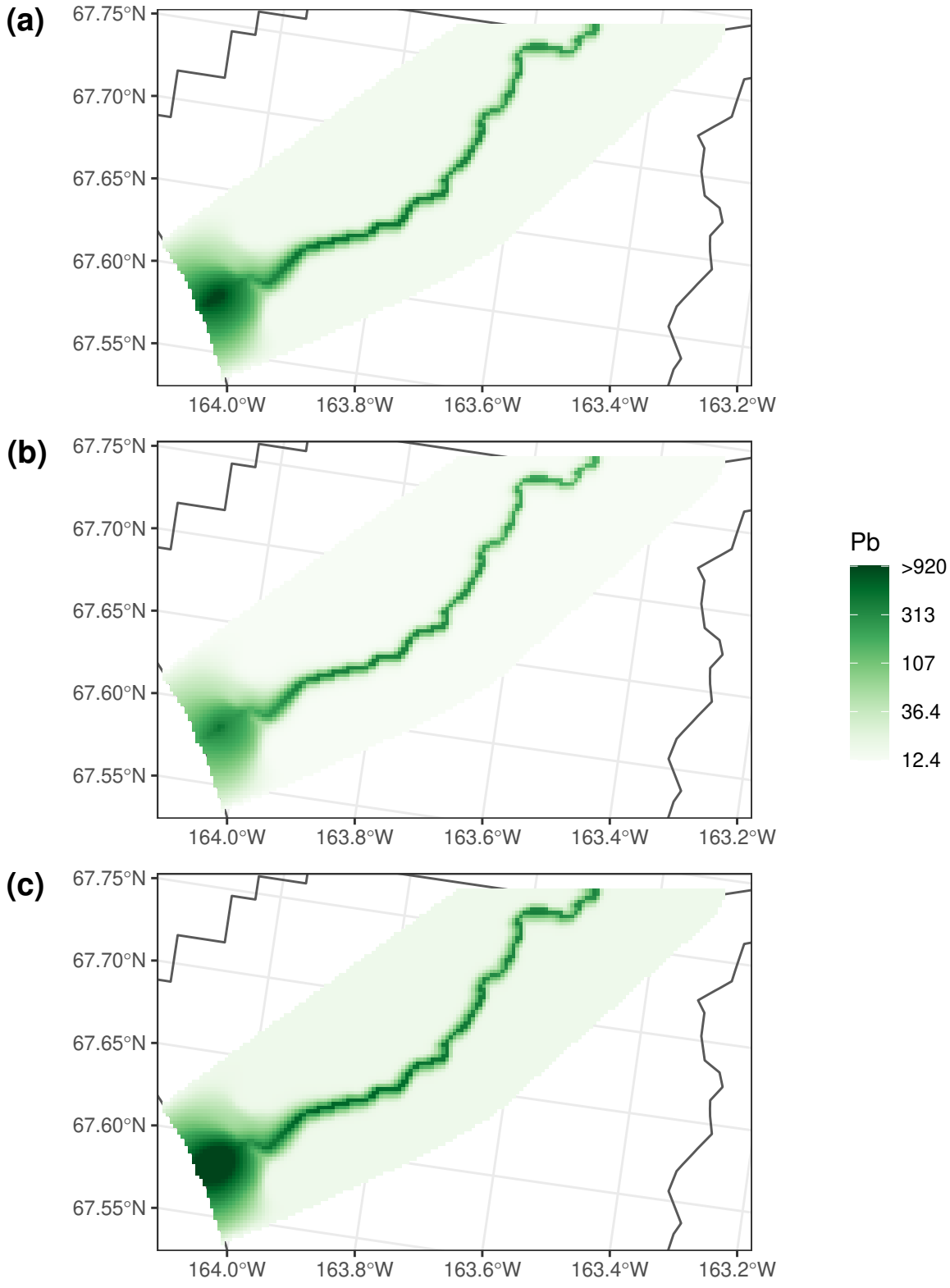

Figure 2.5: Maps summarizing the posterior distribution of Lead (Pb) concentrations on the original scale (i.e., $(\alpha_j + \tilde{\lambda}_j(\mathbf{s}))$) from our fitted model. These maps display the posterior means (a), point-wise 2.5% quantiles (b), and point-wise 97.5% quantiles (c) for locations throughout northern CAKR.

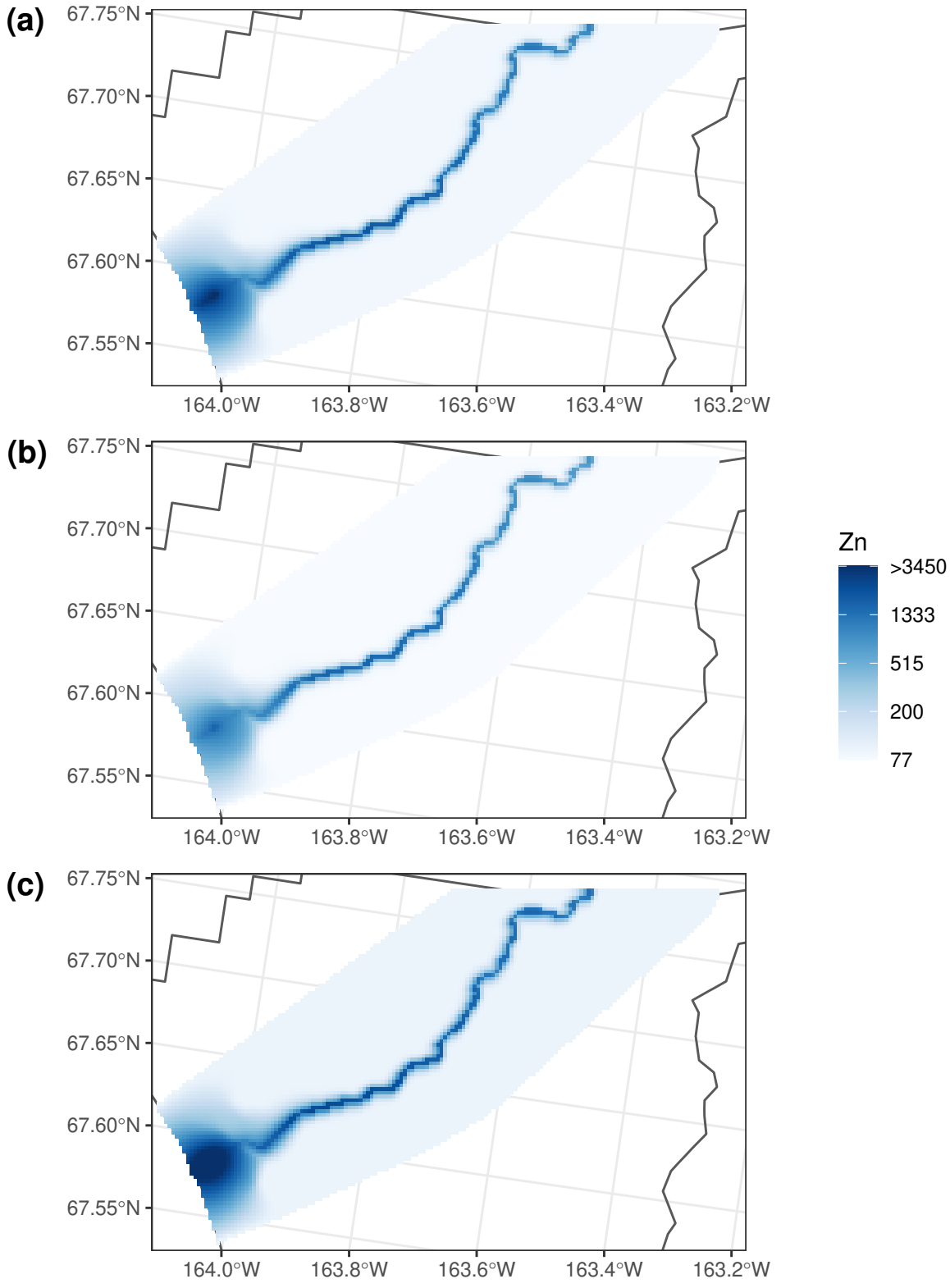

Figure 2.6: Maps summarizing the posterior distribution of Zinc (Zn) concentrations on the original scale (i.e., $(\alpha_j + \tilde{\lambda}_j(\mathbf{s}))$) from our fitted model. These maps display the posterior means (a), point-wise 2.5% quantiles (b), and point-wise 97.5% quantiles (c) for locations throughout northern CAKR.

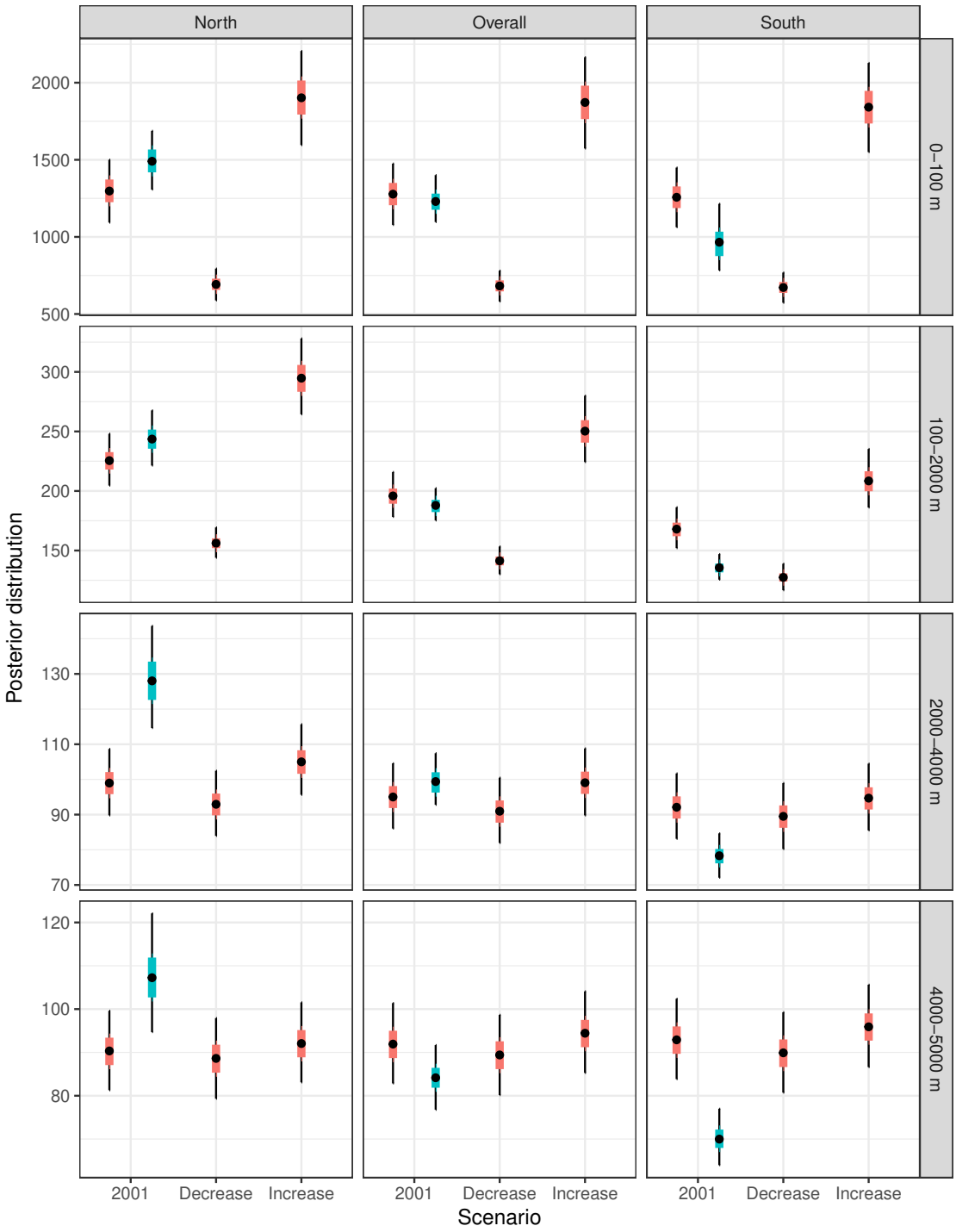

Figure 2.7: Posterior predictive distributions of average concentrations at locations defined by Neitlich et al. (2017). Predictions were summarized distance to road strata (0–100, 100–2000, 2000–4000, 4000–5000 m) and side of road (North, South, or Overall for both combined). The posterior predictive distributions for our model (red) and those from Neitlich et al. (2017) (blue) are shown for 2001. Using our model, we also consider forecast scenarios corresponding to a 50% decrease or increase in pollution from the source.

processes are similar for each. This is also reflected in the maps of estimated concentrations in CAKR (Figures 2.4–2.6). The benefits of partial pooling for these parameters will increase when analyzing datasets with a larger number of chemicals, because it would provide more precise estimates of the PDE parameters.

The latent PDE included in our model is deterministic given values of the associated parameters. Alternatively, Wikle et al. (2022) developed a model for spatial data that is based on a stochastic PDE. They demonstrate how this formulation allows their model to be approximated by a spatial Gaussian process with a covariance matrix determined by the stochastic PDE. For our application, we did not need the computational advantages of reformulating our spatial model in terms of Gaussian processes, but that approach may be necessary when analyzing data over larger spatial extents. However, their approach required the assumption that the advection component was constant over time while we were able to relax this assumption to incorporate daily wind information. A time-varying advection component is beneficial because it closer approximates the known physical process for pollutants dispersing in the atmosphere and allows for spatial structure that may be impossible to capture if advection is assumed to be constant. Our approach also differs from that used by Wikle et al. (2022) because we defined the advection-diffusion PDE (2.1) on the original chemical concentration scale to better reflect the assumed characteristics of the atmospheric dispersion process. Modifying our model to include a stochastic PDE would require additional considerations during implementation to ensure that the modeled chemical concentrations ($\lambda(\mathbf{s}, t)$) were correctly constrained to be positive. This is guaranteed by the PDE in our model (although the numerical implementation must also be considered), but not necessarily by a stochastic PDE.

Compared to the predictions from Neitlich et al. (2017), our model produced similar posterior predictive distributions based on overall summaries by strata and by side of the road for the strata nearest to the road. Some of the differences between our approaches could be due to our assumption that background concentrations were constant over the study region. We considered an expanded model that incorporated additional spatial correlation in the background concentration level but found that this led to computational challenges and negatively affected inferences for the diffusion

parameters. Future work could explore how to expand our model to incorporate additional spatial correlation without negatively impacting inferences for the mechanistic parameters. This may be related to the spatial confounding between fixed and random effects in geostatistical models (Hodges and Reich, 2010) and borrowing ideas from restricted spatial regression may help address these issues (Hanks et al., 2015; Hefley et al., 2017a). Unaccounted for spatial correlation could also reflect unmodeled variability in the advection-diffusion process. We also considered a model that expanded the advection component to incorporate elevation at a location and allow for increased advection downhill. However, in simulations we found the associated parameter difficult to estimate without additional temporal data. Including this additional advection component, or other modifications to the PDE used in this study, could better inform the advection-diffusion process of interest but may require data from additional spatial locations.

In general, mechanistic statistical models are also beneficial because they can allow for improved prediction and forecasting (Hefley et al., 2017b,c; Wikle et al., 2022). Our model can provide insights about how changes in operations at the mine may impact the pollutant concentrations along the haul road. For example, we used our model to predict the heavy metal concentrations for scenarios where fugitive dust emissions from the road are expected to increase or decrease by 50%. Differences in emission levels could result from changes in shipping operations, changes in mining throughput, or modifications to dust mitigation strategies. We illustrated how these hypothetical scenarios could be evaluated to better understand how changes in mining operations may impact heavy metal pollution levels at CAKR. We made posterior predictions under these scenarios by appropriately adjusting the source intercept parameter (θ_0). While such predictions are sensible when using our mechanistic model, it is more challenging to consider forecasts for these scenarios using a geostatistical model that ignores the advection-diffusion process spreading pollutants. This is a known strength of using mechanistic spatio-temporal models to analyze data from environmental processes (Wikle and Hooten, 2010; Cressie and Wikle, 2015). The predictions from our model can help evaluate the possible environmental impacts of potential changes to the mining operations in the area. Such inferences are possible with our mechanistic statistical model

because it estimates the relative rate of pollution generated along the road and how this pollution spreads throughout the study area.

2.7 Data availability

The chemical concentration data are available in the online supporting information of Neitlich et al. (2017). The elevation data were obtained using the R package `elevatr` (Hollister et al., 2023). Daily average wind speed data from the NCEP Reanalysis 2 project was provided by the NOAA/OAR/ESRL PSL, Boulder, Colorado, USA, from their website at http://psl.noaa.gov/.

Chapter 3

Clipped multiscale spatial processes for multivariate binary data

3.1 Introduction

Monitoring the distribution and abundance of plant species is a crucial component of many conservation efforts. The proportion of plot area where a species occurs, called "cover," is a standard measure for monitoring the abundance of plant species (Damgaard and Irvine, 2019; Drezner and Drezner, 2021). Plant cover data are commonly collected using the point intercept method (Drezner and Drezner, 2021) where the presence or absence of a species is recorded at each point of a grid overlaying sampled plots. The resulting binary data are inherently correlated at two spatial scales — small-scale dependence among grid points within a plot and large-scale dependence among different plots. Plant cover data are also multivariate because observations are generally collected for multiple species simultaneously (Godínez-Alvarez et al., 2009; Drezner and Drezner, 2021). Modeling the correlation among different species is essential for many studies because it helps ecologists learn about the community structure and patterns of co-occurrence among species (Ovaskainen et al., 2016; Gelfand, 2022). Together the multiscale and multivariate aspects of plant cover data are challenging to account for and current statistical models for binary spatial data do not address both simultaneously.

Conventional analyses of point intercept data use aggregated observations at the plot level, ignoring subplot-scale information. These analyses rely on the sample proportion of grid points where the species is present to estimate cover within a plot and the plot-level proportions are then modeled. Plot-level observations are also used when analyzing plant cover data that are collected with different methods, such as visually estimated cover proportions (Godínez-Alvarez et al., 2009; Wright et al., 2017; Irvine et al., 2019). For either type of data, some models assume that plant

cover within a plot is a beta-distributed random variable and use beta regression to relate the cover of each species to environmental predictor variables of interest (Wright et al., 2017; Irvine et al., 2019; Damgaard and Irvine, 2019). Other studies described the plot-level community structure in plant cover data using distance-based (e.g., Roberts, 2020; Neitlich et al., 2022) or model-based (e.g., Damgaard et al., 2022) ordination. However, these plot-level analyses ignore the spatial configuration of a species within each plot and do not account for the multiscale spatial structure in plant cover data.

Multiscale spatial models can provide both inferential and computational benefits (Ferreira and Lee, 2007). Ecological processes can occur at multiple spatial scales (Levin, 1992) and statistical models should reflect that. For example, recently Lu et al. (2023) included spatial dependence at multiple levels when modeling the spatio-temporal dynamics of land cover and Kleiven et al. (2023) estimated occupancy at spatially nested sites. In other application areas, multiscale spatial models have been applied to better understand the hierarchical spatial structure in protests from civil unrest (Hoegh et al., 2016), genomic signals (Knijnenburg et al., 2014), mortality ratios, and tornado reports (Fonseca and Ferreira, 2017). While the multiscale spatial structure provides inferential benefits in these examples, it also lends itself to computationally efficient algorithms for analyzing these data (Hoegh et al., 2016). In some cases, multiscale (or "multi-resolution") Gaussian processes can be used to approximate complex spatial models and make analyses of large spatial datasets computationally feasible (Katzfuss, 2017; Katzfuss and Gong, 2020). We develop a Bayesian hierarchical framework that can be applied to plant cover to account for the multiscale spatial structure in point intercept data. Additionally, the multiscale structure in our model leads to a convenient approximation that makes fitting our model more computationally efficient.

Our model explicitly incorporates grid point locations to account for spatial correlation in the distribution of plant species within a plot. By analyzing the binary observations directly, we model the small-scale spatial correlation in addition to the large-scale spatial correlation among different plots. Our approach uses a clipped Gaussian process (De Oliveira, 2000, 2020) to model the binary random field defining the presence and absence of a plant species over a spatial domain of interest.

Clipping discretizes a latent continuous process that is assumed to give rise to certain types of binary data. This approach is similar to latent variable models for univariate (Albert and Chib, 1993) and multivariate (Chib and Greenberg, 1998) binary data and is also useful for analyzing ordinal (Higgs and Hoeting, 2010) or mixed ordinal and continuous multivariate (Schliep and Hoeting, 2013) spatial observations. By linking the binary spatial data to a latent spatial Gaussian process, our approach defines a coherent model for plant occurrence across the entire study region (Gelfand and Shirota, 2019; Gelfand, 2022) and can include spatial structure at multiple scales. Existing spatial models for plant data do not consider the multiscale aspect of plant data because their focus is on modeling occurrence data without directly considering cover (e.g., Hooten et al., 2003; Gelfand and Shirota, 2019; Gelfand, 2022). While our model also uses binary observations at the grid point locations, inferences about cover can be obtained by summarizing the binary random field at the plot level. Overall, our model makes better use of the information available in point intercept data and provides inferences for plant cover that are not possible from analyses of plot-level data.

We also extend our multiscale model to analyze multivariate observations that are typically collected in plant cover studies. The multispecies version of our model includes hierarchical structure in parameters among different species while also allowing for correlation among species. Previous research has shown how including hierarchical structure across species can improve inferences from species distribution models, especially for parameters describing rare or difficult to detect species (Dorazio and Royle, 2005). We allow for interspecies correlations using a latent factor structure that has been used to model ecological communities (Ovaskainen et al., 2016). Another advantage of the multiscale spatial structure of our model is that when analyzing multivariate data, we can use the information about species co-occurrence at individual grid points. This allows for inference about species correlation at the small spatial scale — something that is not possible when analyzing aggregated data at the plot level.

Our model development is motivated by point intercept data collected on lichen, moss, and vascular plant species in Cape Krusenstern National Monument (CAKR), Alaska, USA. Mining activity near CAKR has resulted in elevated concentrations of heavy metal pollutants throughout

the region (Hasselbach et al., 2005; Neitlich et al., 2017; Wright et al., 2022; Neitlich et al., 2024) which negatively affects the plant and animal communities within the monument (Brumbaugh et al., 2010, 2011). The US National Park Service aims to monitor the vegetation within CAKR over time and model how the biological responses, such as cover, are being affected by increased heavy metal concentrations. Previous studies analyzed the plot-level data using nonmetric multidimensional scaling ordination (Neitlich et al., 2022). We explore how our approach can provide additional inference about the impacts of heavy metal pollution by modeling the small scale spatial structure of plant species in CAKR.

In Section 3.2, we describe the point intercept data for lichen, mosses, and vascular plants that were collected in CAKR. Then we detail our general multiscale model and how it can be extended to analyze data from multiple species in Section 3.3. We provide details on the Bayesian methods we use to fit this model in Section 3.4. This includes a hierarchical formulation of our model that is useful for approximating the overall covariance structure, the constraints needed for likelihood identifiability, our assumed prior distributions, and our MCMC algorithm. Section 3.5 summarizes the results of fitting our model to the CAKR data and we end with our conclusions and future work in Section 3.6. We provide additional details about our MCMC algorithm and simulation study in Appendix B.

3.2 CAKR point intercept data

We analyzed point intercept data collected for lichen, mosses, and vascular plants at 104 plots in CAKR during the summer of 2017 (Figure 3.1a). Each plot was a 4×8 meter rectangle and presence/absence observations for each species were collected over a grid of 100 points (Figure 3.1b). The same grid was used at every plot so the spatial configuration of the grid points did not vary across plots. When multiple species overlapped at a single grid point they were all recorded as present, which is common for point intercept data (see also Damgaard and Irvine, 2019). Bryophytes (mosses) were recorded as belonging to one of five possible species groups instead of identifying each observation to an individual species (but we still refer to these taxonomic groups as "species"

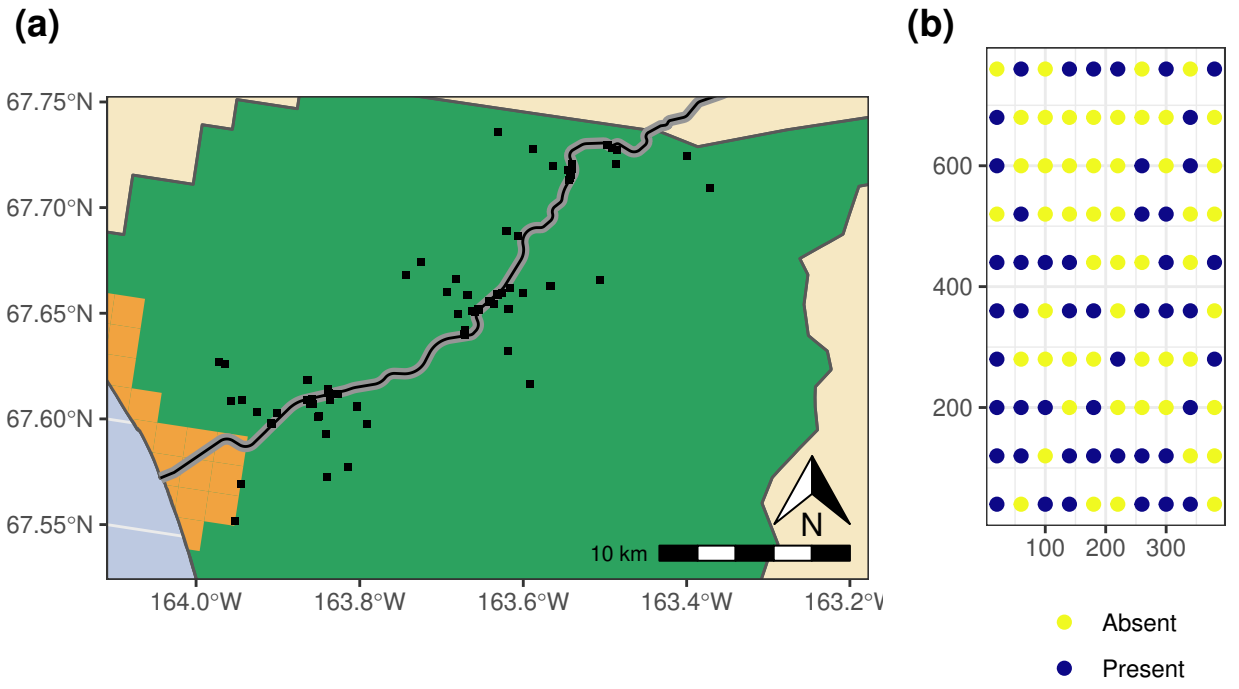

Figure 3.1: Map of northern Cape Krusenstern National Monument (green) with the sampled plots shown by black squares and the Red Dog Mine haul road indicated by the gray line (a). Lands along the coast are owned by the NANA Alaska Native corporation (orange). Each 4 $\times$ 8 m plot includes a grid of 100 points with binary observations of whether a species is present or absent at each point. Data for one plot and one species are shown in (b). The same grid point configuration is used at every plot.

for simplicity). Observations of all vascular plants and lichen were classified to individual species. Additional details on the field methods for these point intercept data can be found in Neitlich et al. (2022).

Mining ore from the Red Dog Mine is hauled along a road through CAKR (Figure 3.1a) and has resulted in elevated concentrations of heavy metals throughout the region (Hasselbach et al., 2005; Neitlich et al., 2017; Wright et al., 2022; Neitlich et al., 2024). We used zinc concentration as a predictor variable in our model because the National Park Service is interested in estimating the impacts of heavy metal pollution on the plant species in CAKR. Zinc concentrations for this analysis were obtained as predictions from a geostatistical model fit to chemical samples collected adjacent to each plot surveyed for the plant cover data (Neitlich et al., 2017, 2024). Model predictions were required for our analysis because zinc concentrations were missing for some plots and we were also interested in predicting plant responses throughout the study area instead of only at the sampled

locations. The concentrations of the other heavy metal pollutants (e.g., cadimum, lead) are all highly correlated with those of zinc (Hasselbach et al., 2005; Neitlich et al., 2017; Wright et al., 2022; Neitlich et al., 2024) and thus we used zinc as a surrogate for heavy metal pollution in our analysis.

3.3 Model

3.3.1 Single plant species

To describe our model for point intercept data, we focus on occurrence over a spatial domain of interest. We let $\{y(\mathbf{s})\}$ denote the binary random field for occurrence of a plant species at spatial locations $\mathbf{s} \in \mathcal{S}$ and assume $\mathcal{S} \subset \mathbb{R}^2$. For any point location $\mathbf{s}$, the random variable $y(\mathbf{s})$ is equal to 1 if the species is present and 0 if absent. Note that the process $\{y(\mathbf{s})\}$ defines species occurrence for an entire study region, as described by Gelfand and Shirota (2019) and Gelfand (2022). Using this binary random field, we can define plot-level summaries that are of interest in studies of plant cover. For instance, the cover within any plot $\mathcal{A} \subseteq \mathcal{S}$ is $|\mathcal{A}|^{-1} \int_{\mathcal{A}} y(\mathbf{s}) d\mathbf{s}$ where $|\mathcal{A}|$ denotes the area of the plot (Gelfand and Shirota, 2019; Gelfand, 2022) and the species is present in the plot if $\int_{\mathcal{A}} y(\mathbf{s}) d\mathbf{s} > 0$.

We model the binary random field $\{y(\mathbf{s})\}$ with a clipped Gaussian process (De Oliveira, 2000, 2020) such that

$$y(\mathbf{s}) = \mathbb{1}(\tilde{y}(\mathbf{s}) \geq c), \tag{3.1}$$

for all $\mathbf{s}$, where $\mathbb{1}(\cdot)$ denotes the indicator function, $\tilde{y}(\mathbf{s})$ is a latent Gaussian process, and c is a constant. The latent continuous process has spatial correlation that induces spatial structure in the binary random field. For point intercept data, we are interested in the small-scale spatial structure in plant cover within a plot as well as the large-scale spatial structure among different plots. We include this multiscale spatial structure by modeling the latent Gaussian process $\tilde{y}(\mathbf{s})$ as

$$\tilde{y}(\mathbf{s}) = \mathbf{x}(\mathbf{s})'\boldsymbol{\beta} + \eta(\mathbf{s}) + \epsilon(\mathbf{s}), \tag{3.2}$$

where $\mathbf{x}(\mathbf{s})$ is a vector of P_1 predictor variables at location $\mathbf{s}$ (including an intercept term), $\boldsymbol{\beta}$ is a vector of regression coefficients, and $\eta(\mathbf{s})$ and $\epsilon(\mathbf{s})$ are spatial Gaussian processes with different effective ranges. That is, we assume

$$\eta(\mathbf{s}) \sim \text{GP}(0, \sigma_\eta^2 K_\eta), \tag{3.3}$$

$$\epsilon(\mathbf{s}) \sim \text{GP}(0, \sigma_\epsilon^2 K_\epsilon), \tag{3.4}$$

where K_η and K_ϵ denote spatial covariance functions with different range parameters, ρ_η and ρ_ϵ. For the point intercept data, we assume that $\eta(\mathbf{s})$ accounts for the large scale spatial dependence among plots while $\epsilon(\mathbf{s})$ accounts for the small scale dependence within each plot. While other spatial covariance functions can include structure at multiple scales, (3.2) allows for a convenient approximation that we use to implement our model (see Section 3.4).

When analyzing point intercept data, we have observations of plant presence at locations $\mathbf{s}_{id}$ for plots $i = 1, \ldots, n$ and grid points $d = 1, \ldots, D$ within each plot. We let $\mathbf{y}$ denote the vector of these nD observations in plot-major ordering so that

$$\mathbf{y} \equiv (y(\mathbf{s}_{11}), \ldots, y(\mathbf{s}_{1D}), \ldots, y(\mathbf{s}_{n1}), \ldots, y(\mathbf{s}_{nD}))'. \tag{3.5}$$

Similarly, we let the vector $\tilde{\mathbf{y}}$ denote the corresponding latent variables associated with these nD observations. For the single species model, (3.2), (3.3), and (3.4) result in the finite-dimensional distribution

$$\tilde{\mathbf{y}} \sim \text{N}(\mathbf{X}\boldsymbol{\beta}, \sigma_\eta^2 \mathbf{R}_\eta + \sigma_\epsilon^2 \mathbf{R}_\epsilon), \tag{3.6}$$

where the matrix $\mathbf{X}$ includes the covariate values from the sampled locations and $\mathbf{R}_\eta$ and $\mathbf{R}_\epsilon$ are correlation matrices defined by the correlation functions K_η and K_ϵ, respectively.

Clipped Gaussian process models require constraints to ensure the likelihood is identifiable (De Oliveira, 2000, 2020). Following convention, we clip the latent process $\tilde{y}(\mathbf{s})$ at zero by letting

$c = 0$ in (3.1) (De Oliveira, 2000, 2020). The likelihood for the single species model is

$$\int_{\mathcal{A}} (2\pi)^{-nD/2} \left|\sigma_\eta^2 \mathbf{R}_\eta + \sigma_\epsilon^2 \mathbf{R}_\epsilon\right|^{-1/2} \times \exp\left(-\frac{1}{2}(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})'(\sigma_\eta^2 \mathbf{R}_\eta + \sigma_\epsilon^2 \mathbf{R}_\epsilon)^{-1}(\tilde{y} - \mathbf{X}\boldsymbol{\beta})\right) d\tilde{\mathbf{y}}, \quad (3.7)$$

where $\mathcal{A}$ defines the area of integration given observations $\mathbf{y}$. When clipping at $c = 0$, $\mathcal{A}$ is the subset of $\mathbb{R}^{nD}$ such that $\tilde{y} \in (-\infty, 0)$ or $\tilde{y} \in [0, \infty)$ when the corresponding element of $\mathbf{y}$ is zero or one, respectively. We assume that the correlation functions K_η and K_ϵ only depend on range parameters ρ_η and ρ_ϵ because smoothness parameters are not identifiable (De Oliveira, 2000). Additionally, the range parameters in the spatial covariance functions of η and ϵ need to differ to avoid symmetry in the likelihood function. Thus, we constrain the range parameters so that $\rho_\eta > \rho_\epsilon$ because the processes account for spatial dependence at different scales.

We rewrite the likelihood as

$$\int_{\mathcal{A}} \left(2\pi\sigma_\epsilon^2\right)^{-nD/2} \left|\frac{\sigma_\eta^2}{\sigma_\epsilon^2}\mathbf{R}_\eta + \mathbf{R}_\epsilon\right|^{-1/2} \times \exp\left(-\frac{1}{2\sigma_\epsilon^2}(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})' \left(\frac{\sigma_\eta^2}{\sigma_\epsilon^2}\mathbf{R}_\eta + \mathbf{R}_\epsilon\right)^{-1} (\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})\right) d\tilde{\mathbf{y}}. \quad (3.8)$$

In this form, parameters in the model do not appear identifiable based on the typical scale invariance property in clipped Gaussian process models (De Oliveira, 2000, 2020). That is, given observations $\mathbf{y}$, the parameter vector $(\boldsymbol{\beta}', \sigma_\eta^2, \rho_\eta, \sigma_\epsilon^2, \rho_\epsilon)'$ has the same likelihood as $(a\boldsymbol{\beta}', a^2\sigma_\eta^2, \rho_\eta, a^2\sigma_\epsilon^2, \rho_\epsilon)'$ for any positive constant a. We fix $\sigma_\epsilon^2 = 1$ to address this (alternatively, we could fix $\sigma_\eta^2 = 1$). The total variance of $\tilde{y}(\mathbf{s})$ is not constrained, however, which is different from clipped Gaussian processes with a single range parameter (De Oliveira, 2000, 2020). In this case, by fixing one of the variances, parameters are identifiable and the second variance parameter will be scaled relative to the fixed variance parameter. While the constraint $\sigma_\eta^2 + \sigma_\epsilon^2 = 1$ would also resolve the identifiability issue, this constraint is computationally inconvenient because it precludes a conjugate update for the variance parameters (see Section 3.4).

Clipped Gaussian processes are closely related to spatial probit models (Hooten et al., 2003; Berrett and Calder, 2016; De Oliveira, 2020). Specifically, if the latent Gaussian process includes a nugget effect (Diggle et al., 1998), then there is a spatial probit model that is equivalent to the clipped Gaussian process model (De Oliveira, 2020). However, these two models can result in data with different characteristics (Berrett and Calder, 2016) — in particular, the spatial probit model does not define distinct boundaries between regions of the binary classes. A spatial logistic model would also fail to define distinct boundaries that delineate where a species occurs. Conversely, by clipping a continuous process (without a nugget effect), our model is well-suited for describing plant cover because it ensures that each point location where a species is present belongs to a contiguous area of locations where the species is present. Individual plants always occur over an area and our model reflects this characteristic of cover data. Thus, we assume the latent spatial process for plant cover does not include a nugget effect and we do not consider spatial probit/logistic models further. Another benefit of our clipped Gaussian process model is that it automatically captures the support of plant cover within a plot. Plot-level cover is a continuous proportion that can also take on values of zero or one exactly. Other approaches model cover as a beta distributed random variable but augment the distribution with a point mass at zero (and potentially one as well; Wright et al., 2017). Summarizing a clipped Gaussian process at the plot level, however, can naturally accommodate the support for plant cover.

3.3.2 Correlation among species

Point intercept data typically include observations for multiple species and thus we expand our clipped Gaussian process model to allow for correlation among species. We model the latent Gaussian processes $\tilde{y}_j(\mathbf{s})$ for species $j = 1, \ldots, J$ as

$$\tilde{y}_j(\mathbf{s}) = \mathbf{x}(\mathbf{s})'\boldsymbol{\beta}_j + \eta_j(\mathbf{s}) + \epsilon_j(\mathbf{s}) + \boldsymbol{\lambda}(\mathbf{s})'\boldsymbol{\theta}_j, \tag{3.9}$$

where $\boldsymbol{\theta}_j$ is a vector of M species-specific coefficients and $\boldsymbol{\lambda}(\mathbf{s})$ is a vector of M independent spatial Gaussian processes that is shared across species. For $m = 1, \ldots, M$, we assume that

$\lambda_m(\mathbf{s}) \sim \text{GP}(0, K_{\lambda_m})$ for the spatial correlation function K_{λ_m} and range parameter ρ_{λ_m}, which can vary across the M Gaussian processes. We fix the variance of each $\lambda_m(\mathbf{s})$ to be one for model identifiability (Huber et al., 2004). This formulation follows the latent factor models developed for joint species distribution models (Ovaskainen et al., 2016; Ovaskainen and Abrego, 2020) where $\boldsymbol{\theta}_j$ denotes the vector of factor loadings for species j and $\boldsymbol{\lambda}(\mathbf{s})$ denotes the vector of latent spatial factors at location $\mathbf{s}$.

The model specified in (3.9) assumes that the correlation among species is spatially homogeneous. For instance, the covariance between the latent processes $\tilde{y}_j(\mathbf{s})$ and $\tilde{y}_{j'}(\mathbf{s})$ for $j \neq j'$ is $\boldsymbol{\theta}_j'\boldsymbol{\theta}_{j'}$ (when the spatial predictor variables are fixed) and this covariance does not depend on the spatial location. Recently, joint species distribution models that allow for variability in the correlation among species have been developed (Tikhonov et al., 2017). In CAKR, for example, elevated heavy metal concentrations could be associated with changes in the correlation among species.

To allow the correlation among species to vary, we include an interaction between each of the latent factors $\lambda_m(\mathbf{s})$ and a vector of spatial covariates $\mathbf{z}(\mathbf{s})$. Then the model for the latent Gaussian process $\tilde{y}_j(\mathbf{s})$ associated with species j is

$$\tilde{y}_j(\mathbf{s}) = \mathbf{x}(\mathbf{s})'\boldsymbol{\beta}_j + \eta_j(\mathbf{s}) + \epsilon_j(\mathbf{s}) + \sum_{p=1}^{P_2} z_p(\mathbf{s})\boldsymbol{\lambda}(\mathbf{s})'\boldsymbol{\theta}_{jp}, \tag{3.10}$$

where $\mathbf{z}(\mathbf{s}) \equiv (z_1(\mathbf{s}), \ldots, z_{P_2}(\mathbf{s}))'$ can include the same or different variables from those in $\mathbf{x}(\mathbf{s})$. Note that if only an intercept is included in $\mathbf{z}(\mathbf{s})$ then (3.10) is equivalent to (3.9) and there is no spatial variability in the coefficients $\boldsymbol{\theta}_j$ for a species. However, when $\mathbf{z}(\mathbf{s})$ includes spatially-varying predictor variables the species-specific coefficients will vary and allow for changes in the correlations among species. This model leads to an identical structure on the factor loadings as the model by Tikhonov et al. (2017) except that we assume the latent factors are spatially correlated.

Based on (3.10), the covariance between the latent processes $\tilde{y}_j(\mathbf{s})$ and $\tilde{y}_{j'}(\mathbf{s})$ for $j \neq j'$ is

$$\text{Cov}(\tilde{y}_j(\mathbf{s}), \tilde{y}_{j'}(\mathbf{s})) = \left(\sum_{p=1}^{P_2} z_p(\mathbf{s})\boldsymbol{\theta}_{jp}\right)' \left(\sum_{p=1}^{P_2} z_p(\mathbf{s})\boldsymbol{\theta}_{j'p}\right), \tag{3.11}$$

which now varies by spatial location $\mathbf{s}$. More generally, (3.10) defines a spatially varying cross-covariance function between the latent processes for any two species. For two spatial locations, $\mathbf{s}_1$ and $\mathbf{s}_2$, the cross-covariance for different species is

$$\begin{aligned}\text{Cov}(\tilde{y}_j(\mathbf{s}_1), \tilde{y}_{j'}(\mathbf{s}_2)) &= \text{Cov}\left(\sum_{p=1}^{P_2} z_p(\mathbf{s}_1)\boldsymbol{\lambda}(\mathbf{s}_1)'\boldsymbol{\theta}_{jp}, \sum_{p=1}^{P_2} z_p(\mathbf{s}_2)\boldsymbol{\lambda}(\mathbf{s}_2)'\boldsymbol{\theta}_{j'p}\right) \\ &= \left(\sum_{p=1}^{P_2} z_p(\mathbf{s}_1)\boldsymbol{\theta}_{jp}\right)' \text{Cov}(\boldsymbol{\lambda}(\mathbf{s}_1)', \boldsymbol{\lambda}(\mathbf{s}_2)') \left(\sum_{p=1}^{P_2} z_p(\mathbf{s}_2)\boldsymbol{\theta}_{j'p}\right), \qquad (3.12)\end{aligned}$$

where $\text{Cov}(\boldsymbol{\lambda}(\mathbf{s}_1)', \boldsymbol{\lambda}(\mathbf{s}_2)')$ is a diagonal matrix with elements $K_{\lambda_m}(\mathbf{s}_1, \mathbf{s}_2)$ for $m = 1, \ldots, M$. This function describes how the covariance between the latent processes for two species decreases as the distance between locations increases. In particular, it depends on the range parameters for each spatial latent factor $\lambda_m(\mathbf{s})$ and shows how the covariance among species can include different spatial scales as well.

Additional constraints are needed for likelihood identifiability in latent factor models (Huber et al., 2004). When assuming no variability in the factor loadings, we use the standard constraints on the coefficients $\boldsymbol{\theta}_j$ to ensure the latent factor model is identifiable. Specifically, we let $\boldsymbol{\Theta} \equiv (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J)$ and then constrain $\boldsymbol{\Theta} \in \mathbb{R}^{M \times J}$ to be an upper triangular matrix with positive diagonal elements (Huber et al., 2004). Note that this constraint does not need to be imposed during the MCMC algorithm, but the factor loadings can be rotated and reflected as needed during post-processing (e.g., Papastamoulis and Ntzoufras, 2022). In general, other constraints can be used for spatial latent factor models depending on the spatial range parameters of the latent factors (Ren and Banerjee, 2013; Zhang and Banerjee, 2022). However, the primary goal with imposing these constraints during post-processing is to assess model convergence — we focus our inferences on the identifiable parameters. When there are additional coefficients $\boldsymbol{\theta}_{jp}$ corresponding to the latent factors interacting with predictor variables z_p, we define the matrices of coefficients $\boldsymbol{\Theta}_p \equiv (\boldsymbol{\theta}_{1p}, \ldots, \boldsymbol{\theta}_{Jp})$ for $p = 1, \ldots, P_2$. In our analysis, we only use indicator variables for each z_p, and thus use the same constraints as defined above for each $\boldsymbol{\Theta}_p$ when post-processing the posterior samples.

3.4 Priors and implementation

3.4.1 Approximate covariance matrix

The spatial clustering of grid points within plots facilitates a natural approximation to the $nD \times nD$ covariance matrix of $\tilde{\mathbf{y}}$ that is implied by the spatial process $\eta(\mathbf{s})$ and $\epsilon(\mathbf{s})$. Note that we have dropped the species subscripts in the first part of this section. The covariance approximation assumes that for point intercept data the distance between plots is much larger than the effective range of the small-scale latent process $\epsilon(\mathbf{s})$ and that the distance between grid points is much smaller than the effective range of the large-scale latent process $\eta(\mathbf{s})$. Under these assumptions, we approximate the covariance of $\tilde{\mathbf{y}}$ in terms of one $n \times n$ correlation matrix and one $D \times D$ correlation matrix that capture the large-scale and small-scale spatial dependence, respectively.

To construct this approximation, we decompose spatial coordinates for plot i and grid point d as $\mathbf{s}_{id} = \mathbf{s}_i + \mathbf{v}_d$ where $\mathbf{s}_i$ is the centroid of plot i and $\mathbf{v}_d$ is the position of grid point d relative to the plot centroid. This decomposition is possible because the same grid is used at every plot when collecting point intercept data. Note that while this decomposition implies every plot is oriented in the same direction, if that is not the case our assumptions still lead to the following approximation. The correlation due to η can be approximated by

$$K_\eta(\mathbf{s}_{id}, \mathbf{s}_{i'd'}) = K_\eta(\mathbf{s}_i + \mathbf{v}_d, \mathbf{s}_{i'} + \mathbf{v}_{d'}) \approx K_\eta(\mathbf{s}_i, \mathbf{s}_{i'}), \tag{3.13}$$

for all grid points d, d' and plots i, i' because the grid point positions $\mathbf{v}_d$ are negligible for the large-scale process. For grid points within the same plot (i.e., $i = i'$), the correlation from the large scale process is approximately one. Similarly, we can rewrite the correlation associated with ϵ as $K_\epsilon(\mathbf{s}_{id}, \mathbf{s}_{i'd'}) = K_\epsilon(\mathbf{s}_i + \mathbf{v}_d, \mathbf{s}_{i'} + \mathbf{v}_{d'})$ for all plots i, i' and grid points d, d'. For grid points within the same plot, the small-scale correlation is

$$K_\epsilon(\mathbf{s}_{id}, \mathbf{s}_{id'}) = K_\epsilon(\mathbf{s}_i + \mathbf{v}_d, \mathbf{s}_i + \mathbf{v}_{d'}) = K_\epsilon(\mathbf{v}_d, \mathbf{v}_{d'}), \tag{3.14}$$

and if the grid points are in different plots this correlation is

$$K_\epsilon(\mathbf{s}_{id}, \mathbf{s}_{i'd'}) = K_\epsilon(\mathbf{s}_i + \mathbf{v}_d, \mathbf{s}_{i'} + \mathbf{v}_{d'}) \approx K_\epsilon(\mathbf{s}_i, \mathbf{s}_{i'}) \approx 0, \tag{3.15}$$

because the distances between plots are much larger than the effective range of the small-scale process. Now we can approximate the covariance of $\tilde{\mathbf{y}}$ using (3.13), (3.14), and (3.15) as follows. We let $\tilde{\mathbf{R}}_\eta$ denote the $n \times n$ correlation matrix defined by K_η using plot centroids $\mathbf{s}_i$ and $\tilde{\mathbf{R}}_\epsilon$ denote the $D \times D$ correlation matrix defined by K_ϵ using point coordinates $\mathbf{v}_d$. Then

$$\sigma^2_\eta \mathbf{R}_\eta + \sigma^2_\epsilon \mathbf{R}_\epsilon \approx (\sigma^2_\eta \tilde{\mathbf{R}}_\eta \otimes \mathbf{1}_D \mathbf{1}'_D) + (\mathbf{I}_n \otimes \sigma^2_\epsilon \tilde{\mathbf{R}}_\epsilon), \tag{3.16}$$

where $\mathbf{1}_D$ is a D-dimensional vector of ones and $\mathbf{I}_n$ is a $n \times n$ identity matrix.

In the original specification of $\tilde{y}(\mathbf{s})$ in (3.2), we included spatial covariates measured at individual location $\mathbf{s}$. However, typically point intercept data include plot-level covariates but not covariates measured at individual grid points. That is, $\mathbf{x}(\mathbf{s}_{id}) = \mathbf{x}(\mathbf{s}_i)$ for all d within the same plot i. Using this and the covariance approximation in (3.16), we approximate the model for the latent process for each species j at the nD observed locations with the hierarchical model

$$\boldsymbol{\mu}_j \sim \mathrm{N}(\tilde{\mathbf{X}}\boldsymbol{\beta}_j, \sigma^2_{\eta,j}\tilde{\mathbf{R}}_{\eta,j}), \tag{3.17}$$

$$\tilde{\mathbf{y}}_{ij} \sim \mathrm{N}(\mu_{ij}\mathbf{1}_D + \boldsymbol{\Lambda}_i\boldsymbol{\theta}_j, \sigma^2_{\epsilon,j}\tilde{\mathbf{R}}_{\epsilon,j}), \tag{3.18}$$

where $\boldsymbol{\mu}_j$ is a n-dimensional vector capturing the large-scale variation across plots for species j, $\boldsymbol{\Lambda}_i$ to be the $D \times M$ matrix with elements $\boldsymbol{\Lambda}_i(m, d) = \lambda_m(\mathbf{s}_{id})$, $\tilde{\mathbf{X}}$ is a matrix of covariates measured at the plot centroids, and $\tilde{\mathbf{y}}_{ij} \equiv (\tilde{y}_j(\mathbf{s}_{i1}), \ldots, \tilde{y}_j(\mathbf{s}_{iD}))'$. By including the latent factor structure at the small scale in (3.18), we assume that the effective range of the spatial correlation functions K_{λ_m} are small relative to the distances between plots. This means the latent factors from different plots are approximately independent and that the correlation among species accounts for dependence at small spatial scales within a plot. Without the correlation among species, this hierarchical form leads to

the same overall covariance structure as the approximation in (3.16) when $\boldsymbol{\mu}_j$ is marginalized out of the model.

3.4.2 MCMC algorithm

With appropriate choices for the prior distributions, most of the parameters can be updated with conditionally conjugate steps. We assumed hierarchical priors for the regression coefficients so that

$$\boldsymbol{\beta}_j \sim \mathrm{N}(\boldsymbol{\mu}_{\beta_g}, \boldsymbol{\Sigma}_{\beta_g}), \tag{3.19}$$

for groups $g = 1, 2$. For the CAKR data analysis, species were grouped depending on whether they were a vascular plant species or not because vascular plants are known to be less sensitive to heavy metal pollution than moss and lichen species. We assumed the mean vectors $\boldsymbol{\mu}_{\beta_g} \sim \mathrm{N}(\nu_\beta \mathbf{1}, \tau_\beta^2 \mathbf{I})$ and that the covariance matrices $\boldsymbol{\Sigma}_{\beta_g}$ are diagonal matrices with elements $(\sigma^2_{\beta_{g1}}, \ldots, \sigma^2_{\beta_{gP}})$ where $\sigma^2_{\beta_{gp}} \sim \text{InverseGamma}(a_{\beta_p}, b_{\beta_p})$. Each factor loading θ_{jpm} was assigned an independent t prior distribution. This prior allows for a conditionally conjugate updates using parameter expansion that can improve mixing of MCMC chains (Ghosh and Dunson, 2009). We assumed that the variance parameters for the large-scale spatial processes $\sigma^2_{\eta,j}$ were independent across species with inverse gamma prior distributions. These prior distributions result in conditionally conjugate updates for parameters $(\boldsymbol{\beta}_j, \boldsymbol{\theta}_j, \sigma^2_{\eta,j})$ and the corresponding hyperparameters $(\boldsymbol{\mu}_{\beta_g}, \boldsymbol{\Sigma}_{\beta_g})$. Similarly, the plot-level means for each species $\boldsymbol{\mu}_j$ and the matrix of latent factors at each plot $\boldsymbol{\Lambda}_i$ can be updated with conditionally conjugate steps. We provide the details and derivations of the full-conditional distributions in Appendix B.1.

The remaining unknown quantities in the model are the spatial range parameters associated with the large-scale and small-scale processes for each species as well as those for the latent factors. We assumed an exponential kernel for all of the spatial correlation functions in our model. We fixed the range parameters for the latent factors λ_m and let $\rho_{\lambda_m} = \rho_\lambda$ for $m = 1, \ldots, M$. With single species models, we found that it may be difficult to obtain reliable inference for range parameters ρ_η and ρ_ϵ when species were less abundant. These challenges are due to limited information about

the range of the latent processes with only binary data and the fact that conditionally conjugate updates for these parameters are not available. We found that discretizing the support for the range parameters (Diggle and Ribeiro Jr, 2002) and also including hierarchical structure across species yields an effective strategy to address these challenges.

Discretizing the support for the range parameters (ρ) is common for geostatistical models (Diggle and Ribeiro Jr, 2002) and, for instance, is standard in the R package `geoR` (Ribeiro Jr and Diggle, 2007). This is computationally advantageous because for each value in the defined support, the correlation matrices only need to be inverted once and can then be accessed as necessary during the MCMC algorithm. Generally the prior is assumed to be a discrete uniform distribution over the specified support, but this does not allow for information about the range parameter to be shared across species. For simplicity in what follows, we omit the subscripts for η and ϵ but the different range parameters were modeled in the same way. We let $\boldsymbol{\rho}_{\text{grid}} \equiv (\rho_{\text{grid},1}, \ldots, \rho_{\text{grid},K})'$ define the discrete support of ρ_j where the values $\rho_{\text{grid},k}$ for $k = 0, \ldots, K$ create an evenly spaced grid. We assumed the discrete prior distribution for ρ_j was

$$\Pr(\rho_j = \rho_{\text{grid},k}) = \begin{cases} \binom{K}{k} \frac{B(\alpha_1+k, \alpha_2+K-k)}{B(\alpha_1,\alpha_2)}, & k = 0, \ldots, K, \\ 0, & \text{otherwise}, \end{cases} \tag{3.20}$$

where $B(\cdot, \cdot)$ denotes the beta function and α_1 and α_2 are hyperparameters. This PMF corresponds to a beta-binomial distribution which we use to map probabilities to the support of ρ_j and allows for pooling across species.

This discretization implies that the full-conditional distribution for each ρ_j can be sampled from directly and no tuning is required for updating these parameters. To update the hyperparameters α_1 and α_2, we used the reparameterization $\nu_\rho = \alpha_1/(\alpha_1 + \alpha_2)$ and $\phi_\rho = \alpha_1 + \alpha_2$. We then specified hyperprior distributions $\nu_\rho \sim \text{Beta}(\delta_1, \delta_2)$ and $\phi_\rho \sim \text{Gamma}(\gamma_1, \gamma_2)$ where $\delta_1, \delta_2, \gamma_1, \gamma_2$ are user-specified. The hyperparameters ν_ρ and ϕ_ρ are updated using Metropolis-Hastings steps in the MCMC algorithm. We chose the discrete support for ρ_η and ρ_ϵ to be non-overlapping to impose the necessary identifiability constraint. Using a simulated example, we found that discretizing the

support of the range parameters provides similar inferences to a model assuming a continuous support (see Appendix B.2). Compared to the model with continuous support, however, the discrete prior distribution increased the effective sample size per computation time for all of the range parameters (Appendix B.2).

3.5 Results of CAKR data analysis

We fit our clipped Gaussian process model to the CAKR data using $J = 30$ species. We included individual species in our analysis when they were observed at least 30 times at individual grid points which corresponds to an overall prevalence of at least $\approx 0.3\%$. The remaining rare species were grouped based on whether they were a vascular plant, moss, or lichen species and these three rare species groups were also included in the analysis. We specified $\tilde{\mathbf{X}}$ to be a B-spline basis matrix of the zinc concentrations (log scale) using a polynomial degree of two and two knots. The spline basis matrix was created using the `splines2` package (Wang and Yan, 2021). This specification allows for non-linear relationships between zinc concentration and the expected value of the latent process $\tilde{y}$ for each species. We assumed $M = 8$ latent factors with all range parameters fixed at 40 cm. We allowed the covariance among species to vary across low ($<$60 mg/kg), medium (60–152 mg/kg), and high ($>$152 mg/kg) zinc concentrations. These levels were defined by the 33% and 66% quantiles of zinc concentrations at the sampled plots and our model included a unique interspecies covariance matrix for each level. Posterior predictive checks based on the cross-covariance of model residuals showed that this latent factor structure was sufficient for modeling the dependence among different species in these data (see Section 3.6 and Appendix B.3). However, if fewer latent factors were included, the model was unable to adequately account for the interspecies correlations in the observed data.

We fit this model using 70,000 total iterations. The first 10,000 iterations were discarded as burn-in and the last 60,000 iterations were used for inference. We examined trace plots for all parameters to assess the convergence of the MCMC algorithm. Our analysis was conducted in R (version 4.2.2, R Core Team, 2021) and our MCMC algorithm was coded in C++ using the

`Rcpp` (Eddelbuettel and François, 2011; Eddelbuettel and Balamuta, 2018) and `RcppArmadillo` (Eddelbuettel and Sanderson, 2014; Eddelbuettel et al., 2023) packages.

The effect of zinc concentration on the expected value of the latent process $\tilde{y}$ and on expected cover varied across species (Figure 3.2). The expected value of the latent process declines as zinc concentration increases for many species (primarily mosses and lichen), but for other species (vascular plants) the expected value of $\tilde{y}$ appears to be constant across the range of sampled zinc concentrations (Figure 3.2a and b). There is some evidence that the relationship with zinc is non-linear on the latent scale. For instance some species appear to be initially tolerant of increases in zinc concentration but then rapidly decline when concentrations surpass approximately 100 mg/kg. The relationships between expected cover and zinc concentration are more varied across species (Figure 3.2c and d). Note that while the expected value of $\tilde{y}$ depends only on the regression

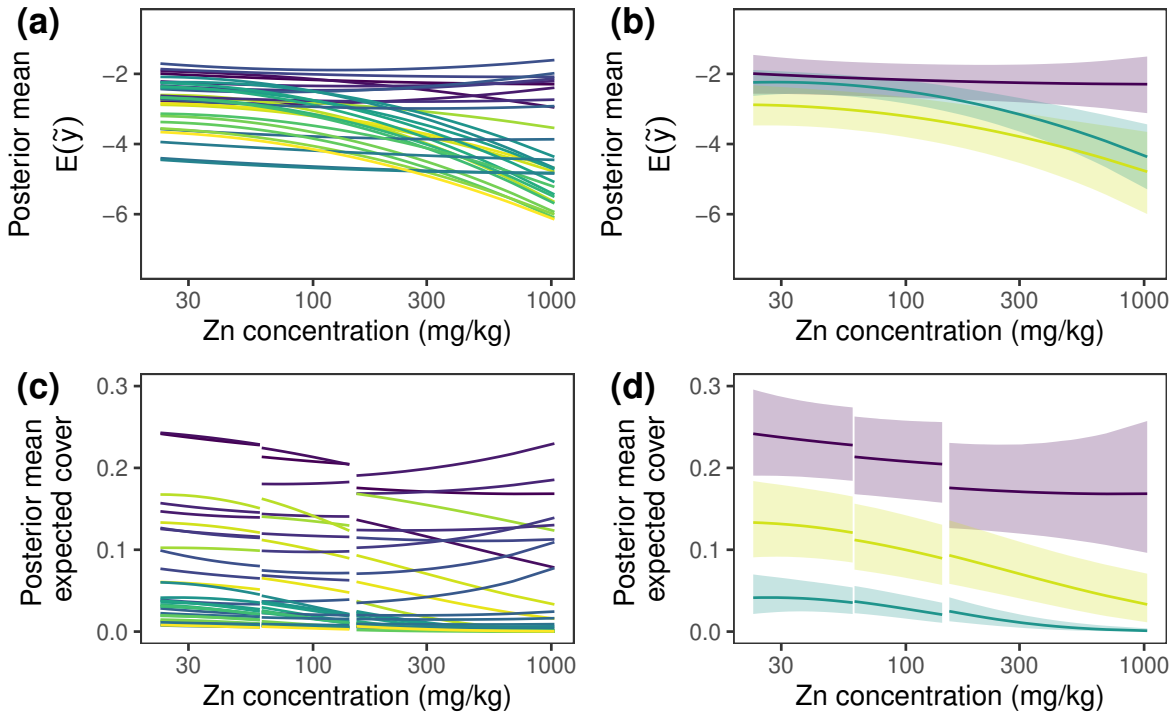

Figure 3.2: For all species in our CAKR analysis, the posterior mean of the expected value of $\tilde{y}$ versus zinc concentration (a). Different species are indicated by the different colors. For a subset of three species, shaded bands show pointwise 95% credible intervals (b). Posterior mean expected cover versus zinc concentration for all species (c) and corresponding pointwise 95% credible intervals for three species (d).

coefficients β, the expected cover is a function of the regression coefficients as well as the factor loadings θ for each species. Because we allowed θ to vary by zinc level, the plots for expected cover have discontinuities at some zinc concentrations. Accounting for the correlation among species appears to change the relationship between expected cover and zinc concentration in some cases, thus it was important to include these interactions.

The large-scale spatial patterns in individual species can also be visualized in maps of the study region. We highlight *Sphagnum* mosses (Figure 3.3a and b) and smooth cup lichen (*Cladonia gracilis*; Figure 3.3c and d) as examples. For both species, the posterior predictive mean of expected cover is lower for locations close to the haul road where zinc concentrations are highest. Based on

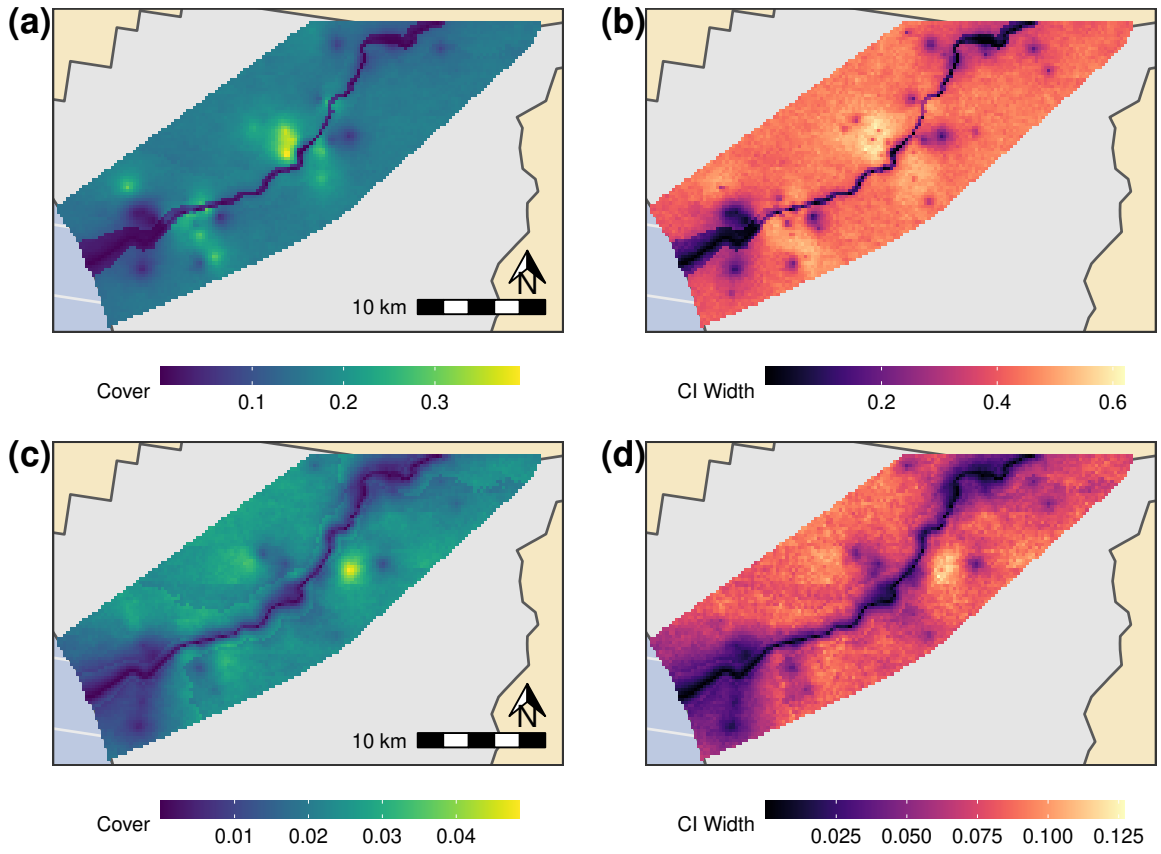

Figure 3.3: For *Sphagnum* moss species, the posterior predictive mean of the expected cover (a) and the width of the associated 95% credible intervals (b) throughout the study region. For smooth cup lichen (*Cladonia gracilis*), the posterior predictive mean of the expected cover (c) and the width of the associated 95% credible intervals (d) throughout the study region.

the model, neither species is expected to occur at locations immediately next to the road. There is also evidence of large-scale spatial variability for these species, but predictive uncertainty is high given the widths of the 95% credible intervals for the expected cover.

In addition to assessing the large-scale relationship between expected cover and zinc concentration for each species, our model allowed us to characterize how the covariance among species varies with zinc concentration. First, consider the overall matrix of cross-covariances for every pair of species in our analysis. We compared these covariances for plots based on the zinc concentration level. We found evidence that the posterior mean covariance differed for some pairs of species based on these two concentrations (Figure 3.4), but there was no evidence of different covariances for most pairs of species. In general, we found negative covariances were larger in magnitude than positive covariances among species, suggesting competitive exclusion.

To illustrate how the cross-covariance for two species can affect the patterns of cover within a plot, we highlight the dependence between Acrocarpous moss species (label 3 in Figure 3.4) and crowberry (label 12 in Figure 3.4, *Empetrum nigrum hermaphroditum*). The fitted model provided evidence that the covariance for these species varied with zinc concentration. Specifically, we found evidence of slight negative covariance for moderate zinc concentrations and positive covariance

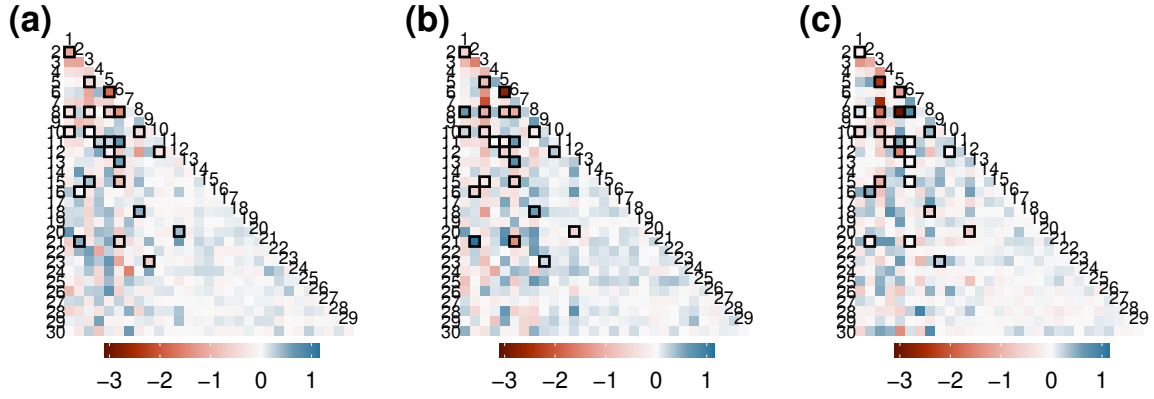

Figure 3.4: Comparison of the covariance among species when zinc concentration is <60 mg/kg (a), 60–152 mg/kg (b), and >152 mg/kg (c). These plots show the posterior mean covariance for every pair of species in our analysis. Outlined cells indicate the pairs of species where 95% credible intervals for the covariances do not overlap for at least two levels of zinc concentration.

for higher zinc concentrations. Posterior predictive realizations of cover within individual plots highlight how these different covariances could lead to different patterns in co-occurrence for these two species. For instance, in a plot where the zinc concentration is 53 mg/kg, and these two species are negatively correlated, they are not expected to occur at the same locations with high probability (Figure 3.5a). At a plot with a higher zinc concentration (237 mg/kg), these species have a positive covariance and co-occur more frequently (Figure 3.5b). This example shows how the correlation among species could change with zinc concentration — in addition to changes in individual species expected cover.

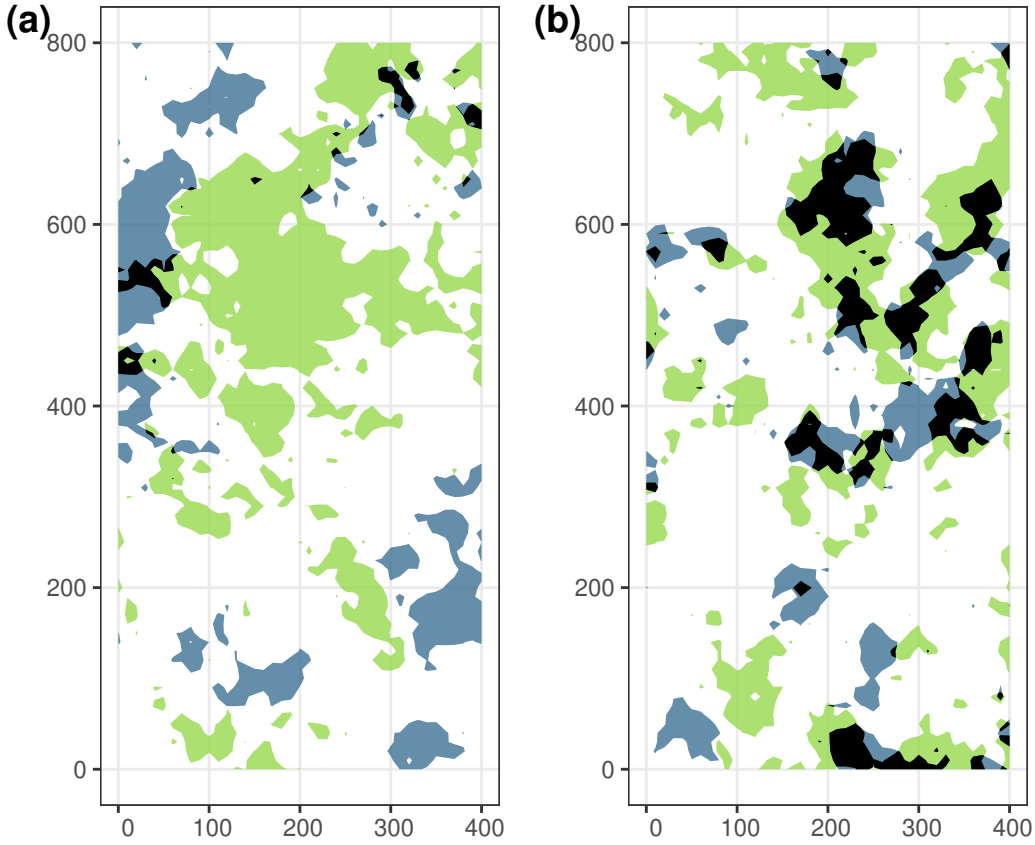

Figure 3.5: For crowberry (*Empetrum nigrum hermaphroditum*, blue) and the Acrocarpous moss species (green), realizations of cover from the posterior predictive distribution for a plot where the zinc concentration is 76 mg/kg (a) and another plot where the zinc concentration is 237 mg/kg (b). The different colors indicate where the different species occur individually within each plot and locations where the species co-occur is shown in black. The posterior mean covariance for this species is -0.18 (95% CI (-0.51, 0.13)) in (a) and 0.45 (95% CI (0.13, 0.79)) in (b).

3.6 Discussion

We developed a multiscale clipped Gaussian process model for analyzing multivariate plant cover data collected with the point intercept method. Our model includes spatial correlation at two scales to account for both the small-scale dependence between observations within each plot as well as large-scale spatial correlation among different plots. To implement our model, we developed an approximation of the multiscale covariance structure and showed the necessary constraints to ensure a multiscale clipped Gaussian process is identifiable in the likelihood. Our model allows us to include subplot scale information in plant cover data that is ignored by other statistical analyses.

We applied our model to analyze cover data for lichen, mosses, and vascular plants in CAKR. Using these data, we modeled how zinc concentration affected the expected cover for various species. While our approach allows for species-specific inferences, community-level inferences (e.g., lichen species richness) are often of interest as well. These can be learned about using multispecies models by appropriately aggregating the species-level inferences (Dorazio and Royle, 2005; Kéry et al., 2009). We used our model to assess how lichen species richness and overall lichen cover was affected by zinc concentration (Appendix B.4). Our model also allowed the correlations among different species to vary with zinc concentration. Because the species correlations in our model can explain associations among species within a plot, our model provides deeper insights into how heavy metal pollution is affecting the vegetation community in CAKR. These inferences are not possible from models that analyze plot-level data and ignore the spatial configuration of cover within a plot. They instead require a model that accounts for both the multiscale and multivariate aspects of plant cover data.

How species interactions are affected by environmental conditions has recently been of interest in community ecology (e.g., Tikhonov et al., 2017; Ovaskainen and Abrego, 2020). Some ecological theories, such as the stress gradient hypothesis (Bertness and Callaway, 1994), are directly concerned with how species interactions can depend on the environmental conditions. Our model provides a way to incorporate and evaluate variability in species interactions with plant cover data collected

using the point intercept method. It could also provide evidence for new and/or different species interactions compared to those observed when analyzing plot-level data.

To assess how well the latent factor structure accounted for the correlations among species in the CAKR data, we conducted posterior predictive checks based on the cross-covariance of model residuals. In simulated datasets, this assessment was able to identify lack of fit in misspecified models relative to data generated with correlated species (Appendix B.3). We first applied these posterior predictive checks after fitting a simple model with no latent factor structure (i.e., $\boldsymbol{\theta}_j = \mathbf{0}$ for all j) and found strong evidence that species are correlated and that for some pairs of species this correlation varied with zinc concentration (Appendix B.3). We then conducted the posterior predictive checks for a model that included the latent factor structure. These assessments showed that accounting for the correlation among species improved the model fit to the CAKR data (Appendix B.3). However, the model required a large number of latent factors relative to the number of species and alternative approaches to model the covariance structure may be more efficient. For instance, the predominance of negative correlations in the CAKR data suggests many species are excluding each other at small scales. This required our model to include a large number of latent factors, but alternative approaches could directly model competition among species. Future work could consider modifications that would allow more flexibility in the covariances among species at small scales.

Measurement errors, such as imperfect detection, are important to consider in studies of plant cover (Wright et al., 2017). Our model assumes that detection of a species at an individual grid point is perfect, which is reasonable considering detection of plant species is generally high even over larger areas (Wright et al., 2017). Even if a plant species is not detected at any of the observed grid points, our model still allows for the species to be present in the plot. This can happen when the corresponding latent process $\tilde{y}(\mathbf{s})$ is greater than or equal to zero at a location within the plot that was not sampled. Therefore, our model still accounts for imperfect detection of a plant species at the plot level. To extend our model to describe the distribution of animal species, we would need to incorporate imperfect detection at the point locations. This would likely require multiple observations per point location, which is standard in occupancy models for animals. Current

occupancy models assume a discretized spatial domain and that a species is present or not within each areal grid cell (Johnson et al., 2013). Conversely, a clipped Gaussian process model for animal occurrence is appealing because it would not require an arbitrary discretization of the study area and would allow for flexible, potentially multiscale, specifications of the spatial dependence in a species' distribution.

Chapter 4

Continuous-space occupancy models

4.1 Introduction

Mapping the distributions of wildlife species is a fundamental component of many ecological studies and wildlife monitoring programs. Occupancy models (Hoeting et al., 2000; MacKenzie et al., 2002, 2018) have become an invaluable approach for modeling species distributions because they account for detection errors that are prevalent in many ecological surveys. This is imperative for obtaining unbiased inferences about species occurrence and how it relates to predictor variables of interest. Additionally, occupancy models are widely applied because they can be used to analyze data from multiple types of surveys, are applicable to a variety of different taxa, and are particularly useful when monitoring species over large spatial extents (MacKenzie et al., 2002; Noon et al., 2012).

Species distributions are the result of inherently spatial processes and there are multiple approaches for modeling spatial dependence in occupancy data (Latimer et al., 2006; Hefley and Hooten, 2016; Gelfand and Shirota, 2019). For instance, spatial dependence in site-level occupancy probabilities can be modeled using conditionally autoregressive terms (in discrete space; Johnson et al., 2013; Broms et al., 2014) or with spatial terms (in continuous space) modeled as a Gaussian process (Ovaskainen et al., 2016; Wright et al., 2021; Doser et al., 2023). These alternative approaches make different assumptions about the spatial support of the process of interest — highlighting a challenge for modeling the spatial structure in occupancy data. While species occupancy is typically viewed as arising from a continuous spatial process (Hooten et al., 2003; Efford and Dawson, 2012), the observed data are collected during surveys of areal units (MacKenzie et al., 2002, 2018). Current spatial occupancy models are unable to account for the change of spatial support between the occupancy and observation processes.

Change of support methods provide a way to make inferences for spatial units that have a support which differs from that of the observed data (Cressie, 1996; Gelfand et al., 2001; Gotway and Young, 2002). These methods are currently available for continuous data (Cressie, 1996; Gelfand et al., 2001), count data (Bradley et al., 2016), and binary data that have been aggregated to areal units (Walker et al., 2020, 2021). Ignoring a change in spatial support can result in biased predictions (Cressie, 1996) and biased inferences for regression coefficients (Walker et al., 2020). We develop a new framework that treats occupancy as a binary process in continuous space and learn about this process using data observed at areal survey units. This provides an approach for modeling a change in spatial support for a new type of data compared to previous spatial models.

Our continuous-space occupancy framework is beneficial because it allows for more realistic spatial models of species occurrence while still properly accounting for the discrete spatial support of the observed data. This provides multiple advantages compared to previously developed spatial occupancy models. For example, our approach allows inferences from areal data to be downscaled, accounting for the fact that areal survey units may only be partially occupied by a species. The observation component of our model can relate these within-unit occupancy proportions to the probability of detecting a species during surveys. This is not possible using standard occupancy models because occupancy is defined as a binary random variable at the level of the survey unit. Another benefit of our model is that it allows for improved inferences about the proportion of area occupied by a species. While this quantity is often of interest, the phrase "proportion of area occupied" can be misleading because modeling species occurrence at the site-level, as done in standard approaches, only permits inferences about the proportion of *sites* occupied (Efford and Dawson, 2012). Our continuous-space occupancy framework directly addresses this limitation and provides inferences for the proportion of area occupied in continuous space.

We implement our model using Bayesian methods and develop a computationally efficient Markov chain Monte Carlo (MCMC) algorithm for fitting our model. We assume that the binary spatial process for occupancy arises from clipping a latent continuous field that is modeled using a Gaussian process (De Oliveira, 2000, 2020). To improve the computational efficiency of fitting

this model over potentially large spatial extents and at many point locations, we approximate the latent Gaussian process using a nearest neighbor Gaussian process (NNGP; Datta et al., 2016). The NNGP approximation makes implementation of spatial models for extremely large datasets possible and helps facilitate Bayesian computation because the necessary calculations associated with the spatial terms are much faster (Datta et al., 2016; Finley et al., 2019). However, MCMC can still be slow to mix for the spatial terms (Finley et al., 2019) and spatial covariance parameters (Murray and Adams, 2010). To address these challenges, we develop an elliptical slice sampler (Murray et al., 2010) to update the spatial terms after marginalizing over the regression coefficients. The conventional elliptical slice sampler assumes that the spatial covariance parameters are fixed. We relax this assumption by modifying the surrogate data slice sampler proposed by Murray and Adams (2010) to be more compatible with NNGPs. The surrogate data slice sampler updates the spatial terms and spatial covariance parameters jointly. This approach is sufficiently general that it could be applied to other spatial models with latent NNGPs.

The remainder is organized as follows. In Section 4.2 we describe occupancy data and the standard models used to analyze these data. Our new model for occupancy in continuous space is presented in Section 4.3 and the details of the MCMC algorithm we use to fit this model appear in Section 4.4. In Section 4.5 we illustrate our approach using a simulated example and perform a simulation study to compare our continuous-space occupancy model to alternative spatial occupancy models. In Section 4.6, we analyze ovenbird occurrence data collected in the Hubbard Brook Experimental Forest, New Hampshire, USA. Section 4.7 discusses future directions that build upon this research. Additional details on our MCMC algorithm, NNGP calculations, and alternative spatial occupancy models are provided in the Appendices.

4.2 Occupancy data and standard analyses

We begin with an overview of the typical data available for conventional occupancy analyses (see also MacKenzie et al., 2002, 2018). Data on species occupancy are collected at areal survey units called "sites." We let $\mathcal{A}_i \subseteq \mathcal{S}$ for $\mathcal{S} \subset \mathbb{R}^2$ denote the region defining site i for $i = 1, \ldots, n$.

Binary detection/nondetection data are collected at each site i during visits j for $j = 1, \ldots, J_i$ where J_i denotes the total number of visits to site i. In standard occupancy models, the binary data y_{ij} are modeled as

$$y_{ij} \sim \begin{cases} \mathbb{1}(y_{ij} = 0), & z_i = 0, \\ \text{Bernoulli}(p_{ij}), & z_i = 1, \end{cases} \tag{4.1}$$

where z_i denotes a partially observed binary random variable for whether the species is present (1) or not (0) at site i and p_{ij} is the probability of detecting the species during visit j to site i. The detection probabilities can be modeled as a function of predictor variables using a generalized linear model framework. When the species is not present at a site (i.e., $z_i = 0$) we assume that there are no false positive detections and $y_{ij} = 0$ for all j with probability 1. However, this assumption can be relaxed and there are approaches available to model false positive detections as well (e.g., Chambert et al., 2015; Ruiz-Gutierrez et al., 2016).

Standard approaches also model the occupancy process at the site-level and therefore assume a discrete spatial domain. That is, occupancy at each site is modeled as

$$z_i \sim \text{Bernoulli}(\psi_i), \tag{4.2}$$

where ψ_i denotes the probability of occupancy at site i which is modeled as a function of spatial predictor variables. A site is considered occupied if the species occurs anywhere within the site area $\mathcal{A}_i$ and we address this in our continuous-space model in Section 4.3. Spatial occupancy models allow ψ_i to have additional spatial structure. This spatial structure can be included using a conditional autoregressive term (Johnson et al., 2013; Broms et al., 2014) or with spatial effects modeled using a continuous Gaussian process based on the locations of the site centroids (Ovaskainen et al., 2016; Wright et al., 2021; Doser et al., 2023).

The discretization of the spatial domain imposed by defining sites is generally arbitrary and does not necessarily have an ecological interpretation. Additionally, occupancy is a process in continuous space (Efford and Dawson, 2012) and it is possible for a species to occur in only a portion of a site

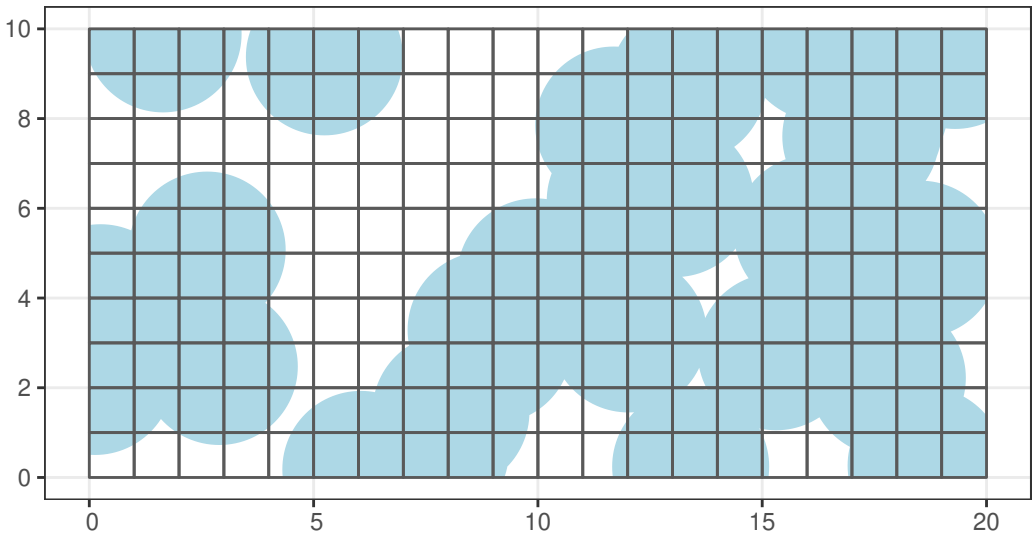

Figure 4.1: Hypothetical example where a regular grid defines sites throughout the study region and species occurrence is shown by the shaded regions. The occupancy process is defined for continuous space even though the detection data are collected at areal sites that discretize the spatial domain. In this example, 72.3% of the study region is occupied but the species occurs in 87.5% of the sites.

(e.g., Figure 4.1). How well standard models can approximate the underlying spatial occupancy process depends on the resolution of the discretization defined by the sites. However, the proportion of *sites* occupied will always be higher than the proportion of the *study area* that is occupied when considering observations over a regular grid (e.g., Figure 4.1; see also Efford and Dawson, 2012). We define an alternative approach for analyzing species occupancy data that models the occupancy process on a continuous spatial domain. The primary challenge for this new approach is that the observation data are still measured at areal sites and we must account for the resulting change of support (e.g., Cressie, 1996; Gelfand et al., 2001; Gotway and Young, 2002) between the occupancy process and the observed data.

4.3 Model

We model species occurrence for spatial locations $\mathbf{s} \in \mathcal{S} \subseteq \mathbb{R}^2$ as a clipped Gaussian process (De Oliveira, 2000, 2020) defined over continuous space. We let $\{z(\mathbf{s})\}$ denote the binary spatial

process for whether the species occurs (1) at location $\mathbf{s}$ or not (0). Note that we define species occurrence $z(\mathbf{s})$ for any point location $\mathbf{s}$ which differs from the site-level occupancy z_i that is used in standard occupancy models (4.1)–(4.2). We assume that this binary spatial process $\{z(\mathbf{s})\}$ arises from clipping a latent continuous process $\{\tilde{z}(\mathbf{s})\}$ such that $z(\mathbf{s}) = \mathbb{1}(\tilde{z}(\mathbf{s}) \geq 0)$ and

$$\tilde{z}(\mathbf{s}) = \mathbf{x}(\mathbf{s})'\boldsymbol{\beta} + \eta(\mathbf{s}), \tag{4.3}$$

where $\mathbf{x}(\mathbf{s})$ is a vector of spatially indexed predictor variables with corresponding coefficients $\boldsymbol{\beta}$ and $\eta(\mathbf{s})$ is a spatial Gaussian process with mean zero and spatial covariance function K_η.

The detection/nondetection data y_{ij} are recorded at sites $i = 1, \ldots, n$ and visits $j = 1, \ldots, J_i$. We still assume there are no false positive detections and thus the species can only be detected if it occupies at least a portion of the site. Therefore, we model the detection data y_{ij} as

$$y_{ij} \sim \begin{cases} \mathbb{1}(y_{ij} = 0), & \max_{\mathbf{s}\in\mathcal{A}_i} z(\mathbf{s}) = 0, \\ \text{Bernoulli}(p_{ij}), & \max_{\mathbf{s}\in\mathcal{A}_i} z(\mathbf{s}) = 1, \end{cases} \tag{4.4}$$

$$\Phi^{-1}(p_{ij}) = \mathbf{w}_{ij}'\boldsymbol{\alpha} + \gamma|\mathcal{A}_i|^{-1}\int_{\mathcal{A}_i} z(\mathbf{s})d\mathbf{s}, \tag{4.5}$$

where the probit link function $\Phi(\cdot)$ denotes the cumulative distribution function for a standard normal random variable, $\mathbf{w}_{ij}$ is a vector of predictor variables related to detection, and the proportion of the site area where the species occurs is $|\mathcal{A}_i|^{-1}\int_{\mathcal{A}_i} z(\mathbf{s})d\mathbf{s}$. The occupancy proportion within a site is related to the probability of detection with corresponding parameter γ. In general, detection probabilities should increase as the proportion of the site that is occupied increases. The observation process of our model is similar to that of the standard occupancy model defined in Section 4.2 except that we explicitly define site-level occupancy as a function of the continuous occupancy process and we allow the detection probability to depend on the proportion of the site that is occupied. This observation process also provides more flexibility than standard models because we do not need to assume that the site regions $\mathcal{A}_i$ are mutually disjoint (e.g., see Section 4.6).

Our model accounts for the change of support between the observation and occupancy processes. This allows for inferences about occupancy to be downscaled to continuous space even though the observed data are measured at the site-level. In other words, the observed detection/nondetection data are recorded at areal sites but we are still able to obtain inferences for occupancy in continuous space — we do not need to assume the entire site is occupied when a species is detected there. Modeling occupancy in continuous space is beneficial because it allows for more realistic inferences about this ecological process. Additionally, it allows us to relate the detection process to the proportion of a site that is occupied. This relationship is intuitive but cannot be incorporated into traditional occupancy models. Our statistical model is the first to incorporate observation errors, in the form of imperfect detection, when modeling binary spatial data using a clipped Gaussian process.

4.4 Priors and implementation

We implement our model using Bayesian methods and assume normal prior distributions for $\boldsymbol{\alpha}$, γ, and $\boldsymbol{\beta}$. Using a Gibbs sampling approach, these prior distributions facilitate conjugate updates for many of the parameters. For instance, because we use a probit link for modeling detection probabilities in (4.5), the detection-level parameters ($\boldsymbol{\alpha}$, γ) can be updated using conditionally conjugate steps by introducing latent variables $\tilde{y}_{ij}$ such that $y_{ij} = \mathbb{1}(\tilde{y}_{ij} \geq 0)$ (Albert and Chib, 1993). The regression coefficients for occupancy $\boldsymbol{\beta}$ can also be updated with a conjugate step conditional on the latent spatial process $\tilde{z}(\mathbf{s})$. However, we found that the convergence and mixing of the MCMC samples could be improved by integrating $\boldsymbol{\beta}$ out of the model and directly updating the latent spatial process $\tilde{z}(\mathbf{s})$. Inferences for $\boldsymbol{\beta}$ can then be recovered in a post-processing step. The main challenges for implementing this model are associated with updating the latent spatial terms $\tilde{z}(\mathbf{s})$ and the spatial covariance parameters $\boldsymbol{\theta}$. We describe how we address these challenges and provide the details for our MCMC algorithm throughout the rest of this section and in Appendix C.1. All analyses for our simulated (Section 4.5) and real data examples (Section 4.6) were conducted

in R (R Core Team, 2022). We also used the `Rcpp` (Eddelbuettel and Balamuta, 2018) and `RcppArmadillo` packages (Eddelbuettel and Sanderson, 2014) to code our MCMC algorithm.

4.4.1 Numerical quadrature

While our occupancy model is defined for continuous space, we consider only a finite number of locations to approximate $\tilde{z}(\mathbf{s})$. We let $\tilde{\mathbf{z}} \equiv (\tilde{z}(\mathbf{s}_1), \ldots, \tilde{z}(\mathbf{s}_D))'$ define the vector used to implement this approximation. The locations $\mathbf{s}_d$ for $d = 1, \ldots, D$ are chosen to cover all sites in the study region and we provide more details on choosing these locations later in this section. Similarly, we define design matrix $\mathbf{X} \equiv (\mathbf{x}(\mathbf{s}_1), \ldots, \mathbf{x}(\mathbf{s}_D))'$ and vector of spatial terms $\boldsymbol{\eta} \equiv (\eta(\mathbf{s}_1), \ldots, \eta(\mathbf{s}_D))'$ at the same point locations. The finite-dimensional occupancy process can now be modeled as $\tilde{\mathbf{z}} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta}$ where $\boldsymbol{\eta} \sim \text{N}(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$ and $\boldsymbol{\Sigma}_\eta$ is defined by the spatial covariance function K_η. Additionally, we use numerical quadrature to approximate the functions of $z(\mathbf{s})$ in (4.4) and (4.5) as

$$\max_{\mathbf{s} \in \mathcal{A}_i} z(\mathbf{s}) \approx \max_{\mathbf{s}_d \in \mathcal{A}_i} \mathbb{1}(\tilde{z}(\mathbf{s}_d) \geq 0) \tag{4.6}$$

and

$$|\mathcal{A}_i|^{-1} \int_{\mathcal{A}_i} z(\mathbf{s}) d\mathbf{s} \approx D_i^{-1} \sum_{\mathbf{s}_d \in \mathcal{A}_i} \mathbb{1}(\tilde{z}(\mathbf{s}_d) \geq 0), \tag{4.7}$$

where D_i is the total number of points $\mathbf{s}_d$ contained in $\mathcal{A}_i$. If the sites define a regular grid over the study region, then we define point locations $\mathbf{s}_d$ such that each site contains the same number of points. The errors associated with both of these approximations can be made arbitrarily small by making D sufficiently large.

4.4.2 Nearest neighbor Gaussian process

As with many spatial models, increasing the number of points D used to approximate the latent spatial process can result in this model becoming computationally challenging to implement. We utilize a nearest neighbor Gaussian process (NNGP) to approximate $\tilde{z}(\mathbf{s})$ which results in the finite-dimensional distribution of $\tilde{\mathbf{z}}$ having a sparse precision matrix (Datta et al., 2016). We

provide a general description of NNGPs in what follows and more details about constructing the resulting sparse precision matrix can be found in Appendix C.2. The NNGP approach relies on a Vecchia approximation (Vecchia, 1988) of the joint distribution of a spatial process. Using standard factorization properties, the joint distribution of the spatial terms from our model can be written as

$$[\tilde{\mathbf{z}}] = [\tilde{z}_1][\tilde{z}_2 \mid \tilde{z}_1] \cdots [\tilde{z}_D \mid \tilde{z}_1, \ldots, \tilde{z}_{D-1}], \tag{4.8}$$

where we use $[\cdot]$ to denote a probability density function (Gelfand and Smith, 1990). The Vecchia approximation defines the conditional distributions in the factorization (4.8) to only depend on a set of nearest neighbors selected from the previous observations. Consequently, we can approximate the joint distribution of the spatial terms as

$$[\tilde{\mathbf{z}}] \approx \prod_{d=1}^{D} [\tilde{z}_d \mid \tilde{\mathbf{z}}_{c(d)}], \tag{4.9}$$

where $c(d) \subseteq \{1, \ldots, d-1\}$ defines a set of nearest neighbors among the previous terms ($c(1)$ is the null set) and $\tilde{\mathbf{z}}_{c(d)}$ denotes a vector containing the spatial terms in that set.

The NNGP approximation requires decisions about how the D points are ordered and how the neighbors $c(d)$ are selected. We use the "maxmin" ordering proposed by Guinness (2018) and select the m nearest neighbors based on the spatial distances to the previous points based on this ordering (Vecchia, 1988; Datta et al., 2016). The maxmin ordering chooses the next location to be the one that maximizes the minimum distance to previous points and can substantially improve the accuracy of NNGP approximations (Guinness, 2018). Additionally, for at least some points, this ordering will include neighbors that have large distances from the point of interest which can provide more information about parameters in the spatial covariance function (Stein et al., 2004).

To approximate the maxmin ordering for a large number of points, we first define a set of regular grid cells that cover the study region. The sites can be used if they form a regular grid. Otherwise, an arbitrary grid can be defined when sites form an irregular grid and/or overlap one another (see Section 4.6). The maxmin ordering is used to first order the grid cells (based on their centroids) and

then the maxmin ordering can be used for the points within each grid cell. This approach results in an overall ordering that will approximate the overall maxmin order for the points (Guinness, 2018).

4.4.3 Updating the spatial terms

The NNGP approximation allows for improved computation of the joint density of the spatial terms that can be utilized in our MCMC algorithm. However, updating the spatial terms $\tilde{\mathbf{z}}$ can still be challenging because they are highly correlated and, consequently, MCMC chains can converge very slowly (Datta et al., 2016; Finley et al., 2019). When the data model is Gaussian, marginalizing over the spatial random terms can improve the mixing of MCMC (Shi et al., 2017; Finley et al., 2019). This approach is not straightforward for our model because the likelihood defined in (4.4) and (4.5) becomes challenging to evaluate. We consider an alternative approach that marginalizes over the coefficients $\boldsymbol{\beta}$ and updates the spatial terms jointly using an elliptical slice sampler (Murray et al., 2010) or a surrogate data slice sampler (Murray and Adams, 2010).

We describe our MCMC algorithm when the spatial covariance parameters are fixed (i.e., $\boldsymbol{\Sigma}_\eta$ is known) and then generalize this algorithm to allow the spatial covariance parameters to be modeled as well. Assuming the prior distribution $\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$, integrating $\boldsymbol{\beta}$ from (4.3) implies

$$\tilde{\mathbf{z}} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\mu}_\beta, \mathbf{X}\boldsymbol{\Sigma}_\beta\mathbf{X}' + \boldsymbol{\Sigma}_\eta), \tag{4.10}$$

for the finite-dimensional locations. We denote the mean and variance in (4.10) by $\boldsymbol{\mu}_{\tilde{z}}$ and $\boldsymbol{\Sigma}_{\tilde{z}}$, respectively. The marginal distribution for $\tilde{\mathbf{z}}$ is approximated using the NNGP approach described in Section 4.4.2 and Appendix C.2. Conditional on the other parameters in the model, these spatial terms can be updated from the full-conditional distribution

$$[\tilde{\mathbf{z}} \mid \mathbf{y}, \cdot] \propto [\tilde{\mathbf{z}}][\mathbf{y} \mid \tilde{\mathbf{z}}, \cdot], \tag{4.11}$$

where $[\mathbf{y} \mid \tilde{\mathbf{z}}, \cdot]$ denotes the observed data likelihood conditional on the spatial terms and all other parameters in the model. The form of (4.11) can be sampled from using an elliptical slice sampler

(Murray et al., 2010) because the prior $[\tilde{\mathbf{z}}]$ is multivariate normal. Elliptical slice sampling is appealing for models using a latent Gaussian process because there are no restrictions on the form of the likelihood, it is easy to implement, and does not require tuning (Murray et al., 2010). Conditional on $\tilde{\mathbf{z}}$, we sample $\boldsymbol{\beta}$ from its full-conditional distribution (see Appendix C.1).

When the spatial covariance parameters $\boldsymbol{\theta}$ are not fixed, we update them and the spatial terms $\tilde{\mathbf{z}}$ jointly using a modified version of the surrogate data slice sampler developed by Murray and Adams (2010). A joint update for these parameters is important because they are highly correlated in the posterior distribution and conditional updates result in poor mixing of the MCMC chains (Murray and Adams, 2010). The surrogate data slice sampler introduces auxiliary parameters into the model that allow for a series of convenient Gibbs updates for $\tilde{\mathbf{z}}$ and $\boldsymbol{\theta}$. Directly applying the approach from Murray and Adams (2010) would introduce surrogate data for every spatial location $\mathbf{s}_d$. We found that this was not conducive to using the NNGP approach for the spatial terms and instead introduce surrogate data for only a subset of the spatial locations. This allows us to retain the computationally efficient calculations facilitated by the NNGP when applying the surrogate data slice sampler.

To define our surrogate data slice sampler, we partition the spatial terms such that

$$\tilde{\mathbf{z}} \equiv \begin{pmatrix} \tilde{\mathbf{z}}_1 \\ \tilde{\mathbf{z}}_2 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right), \tag{4.12}$$

where we have simplified the notation by assuming $\boldsymbol{\mu}_\beta = \mathbf{0}$, which implies that $\boldsymbol{\mu}_{\tilde{z}} = \mathbf{0}$, and omitting the $\tilde{z}$ subscripts from the partitioned covariance matrix. The partitions are chosen such that $\tilde{\mathbf{z}}_1$ has relatively few locations — these will be the locations where we introduce surrogate data. We expand this model with auxiliary variables $\boldsymbol{\nu}_1$, $\boldsymbol{\nu}_2$, and $\mathbf{g}$ that are assumed to be marginally distributed

$$\begin{pmatrix} \boldsymbol{\nu}_1 \\ \boldsymbol{\nu}_2 \\ \mathbf{g} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_g \end{pmatrix} \right), \tag{4.13}$$

where $\boldsymbol{\Sigma}_g$ is a user-specified covariance matrix corresponding to surrogate data $\mathbf{g}$. Conditional on these auxiliary variables, we set

$$\tilde{\mathbf{z}}_1 = \mathbf{m}_1 + \mathbf{L}_1\boldsymbol{\nu}_1, \tag{4.14}$$

$$\tilde{\mathbf{z}}_2 = \mathbf{m}_2 + \mathbf{L}_2\boldsymbol{\nu}_2, \tag{4.15}$$

where

$$\mathbf{m}_1 = \boldsymbol{\Sigma}_{11}\left(\boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_g\right)^{-1}\mathbf{g}, \tag{4.16}$$

$$\mathbf{L}_1\mathbf{L}_1' = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{11}\left(\boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_g\right)^{-1}\boldsymbol{\Sigma}_{11}, \tag{4.17}$$

$$\mathbf{m}_2 = \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\left(\mathbf{m}_1 + \mathbf{L}_1\boldsymbol{\nu}_1\right), \tag{4.18}$$

$$\mathbf{L}_2\mathbf{L}_2' = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}. \tag{4.19}$$

Routine calculations show that constructing the auxiliary variables in this way induces the same marginal distribution for $(\tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2)$ as that in (4.12). This construction also defines a joint normal distribution for $(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \mathbf{g}, \tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2)$. Thus, we obtain a sample from our target posterior distribution using a Gibbs sampler for this parameter expanded model.

Conditional on the detection-level parameters, the posterior distribution of the parameter expanded model is $[\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \mathbf{g}, \tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2, \boldsymbol{\theta} \mid \mathbf{y}]$. The first step of our Gibbs sampler updates the auxiliary parameters from

$$[\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \mathbf{g} \mid \mathbf{y}, \tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2, \boldsymbol{\theta}] = [\mathbf{g} \mid \tilde{\mathbf{z}}_1, \boldsymbol{\theta}][\boldsymbol{\nu}_1, \boldsymbol{\nu}_2 \mid \tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2, \boldsymbol{\theta}, \mathbf{g}], \tag{4.20}$$

where $(\mathbf{g} \mid \tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2, \boldsymbol{\theta}) \stackrel{d}{=} (\mathbf{g} \mid \tilde{\mathbf{z}}_1, \boldsymbol{\theta}) \sim \mathrm{N}(\tilde{\mathbf{z}}_1, \boldsymbol{\Sigma}_g)$ as shown in Appendix C.1. The distribution of $(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2 \mid \tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2, \boldsymbol{\theta}, \mathbf{g})$ is degenerate as implied by (4.14) and (4.15). The second step of the Gibbs sampler is to update the spatial terms and covariance parameters from

$$[\tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2, \boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \mathbf{g}] = [\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \mathbf{g}][\tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2 \mid \boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \mathbf{g}, \boldsymbol{\theta}], \tag{4.21}$$

where $[\tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2 \mid \boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \mathbf{g}, \boldsymbol{\theta}]$ is also degenerate by (4.14) and (4.15). Updating the spatial covariance parameters from $[\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \mathbf{g}]$ can be performed efficiently using a slice sampler because $\boldsymbol{\theta}$ will generally contain only a few parameters (Neal, 2003). As suggested by Murray and Adams (2010), in our full MCMC algorithm we perform the surrogate data slice sampling step every 25 iterations (real data example) or every 10 iterations (simulated data examples) to reduce computation time. For the remaining iterations, we fix $\boldsymbol{\theta}$ and update the spatial terms using elliptical slice sampling.

Our surrogate data slice sampling algorithm uses the NNGP approximation when updating $\boldsymbol{\nu}_2$ and $\tilde{\mathbf{z}}_2$. In general, we can reparameterize our model by introducing a vector of independent standard normal random variables $\boldsymbol{\nu}$ and redefining the spatial terms as $\tilde{\mathbf{z}} = \boldsymbol{L}\boldsymbol{\nu}$, where $\boldsymbol{L}$ is a factorization of $\boldsymbol{\Sigma}_{\tilde{z}}$ (and therefore depends on $\boldsymbol{\theta}$) such that $\boldsymbol{L}\boldsymbol{L}' = \boldsymbol{\Sigma}_{\tilde{z}}$. Using standard NNGP results, $\boldsymbol{L}^{-1}$ is readily available and efficient algorithms exist for calculating $\tilde{\mathbf{z}}$ by solving the sparse system $\boldsymbol{L}^{-1}\tilde{\mathbf{z}} = \boldsymbol{\nu}$ (Saha and Datta, 2018; Datta, 2022). Given $\tilde{\mathbf{z}}_1$, (4.15) defines $\tilde{\mathbf{z}}_2$ using mean $\mathbf{m}_2$ and variance $\mathbf{L}_2\mathbf{L}_2'$ which are equivalent to the conditional mean $\mathrm{E}(\tilde{\mathbf{z}}_2 \mid \tilde{\mathbf{z}}_1)$ and variance $\mathrm{Var}(\tilde{\mathbf{z}}_2 \mid \tilde{\mathbf{z}}_1)$, respectively. Thus,

iterative algorithm for solving $\boldsymbol{L}^{-1}\tilde{\mathbf{z}} = \boldsymbol{\nu}$ when using a NNGP approximation (see Appendix C.2 for more details). Similarly, we can reverse this algorithm to find $\boldsymbol{\nu}_2$ conditional on $\boldsymbol{\nu}_1$, $\tilde{\mathbf{z}}_1$, and $\tilde{\mathbf{z}}_2$ in the first Gibbs step. Introducing the surrogate data for $\tilde{\mathbf{z}}_1$ does not allow for the same NNGP calculations to be used. However, by choosing the dimension of the surrogate data to be sufficiently small, we can perform the requisite calculations in (4.14), (4.16), and (4.17) exactly. That is, we do not need to rely on the sparsity of $\tilde{\boldsymbol{\Sigma}}_{11}^{-1} \approx \boldsymbol{\Sigma}_{11}^{-1}$ when updating $\tilde{\mathbf{z}}_1$ or $\boldsymbol{\nu}_1$.

Tuning is required for the surrogate data slice sampler. First, the number of points to introduce surrogate data must be chosen. This choice is in part limited by the computational resources available. In general, we choose $\tilde{\mathbf{z}}_1$ to include one point per site or grid cell used to order the points as described in Section 4.4.2. The second choice is the specification of the surrogate data covariance $\boldsymbol{\Sigma}_g$. We specify this covariance to be a diagonal matrix with elements $\sigma_{g,d}^2$ tuned to be approximately twice the posterior variance of the corresponding spatial term $\tilde{z}_{1,d}$. Other approaches to specify the surrogate data covariance are possible as well (see Murray and Adams, 2010).

4.5 Simulations

4.5.1 Simulated example

We start by demonstrating our model using a simulated data example. We simulated occupancy in continuous space as a clipped Gaussian process with a single spatial covariate and a spatial term that had a Gaussian covariance function. We considered a 40×20 unit rectangular study area and defined sites using a 1×1 unit regular grid over the region. Occupancy data were simulated at 200 randomly selected sites out of the 800 total sites over the region with three visits per site (Figure 4.2a). We also simulated a visit-level covariate from a Uniform(-1, 1) distribution. Data were generated based on parameters $\boldsymbol{\beta} = (-0.5, 2)'$, $\boldsymbol{\alpha} = (-2, 1)'$, and $\gamma = 3$. Because γ is positive, for this simulated example detection probabilities within a site increase as the proportion of the site that is occupied by the species increases.

We fit our continuous-space occupancy model to these simulated data using 20,000 iterations to tune the surrogate data slice sampler followed by 100,000 iterations for posterior inference. The 95% marginal credible intervals captured values of the parameters used to generate these data. Our focus is on inferences for mapping occupancy and summarize the posterior probability of occupancy for spatial locations throughout the region considered. We calculate the marginal posterior probability of occupancy for spatial location $\mathbf{s}$ as $\text{P}(z(\mathbf{s}) = 1 \mid \mathbf{y}) \approx T^{-1} \sum_{t=1}^{T} z(\mathbf{s})^{(t)}$ where $z(\mathbf{s})^{(t)}$ denotes the sampled value for the binary spatial process z on MCMC iteration t and T is the total number of iterations. Overall, the map of the posterior probabilities of occupancy recovers the underlying occupancy process well for this simulated example (Figure 4.2b and c).

In general, this example illustrates how we can recover an occupancy process in continuous space using data observed at areal sites. The downscaling of inferences is possible because i) the spatial structure in the occupancy process, ii) the spatial predictor variable, and iii) the relationship between detection probabilities and the proportion of a site that is occupied. All of these provide information at a finer resolution than the areal sites.

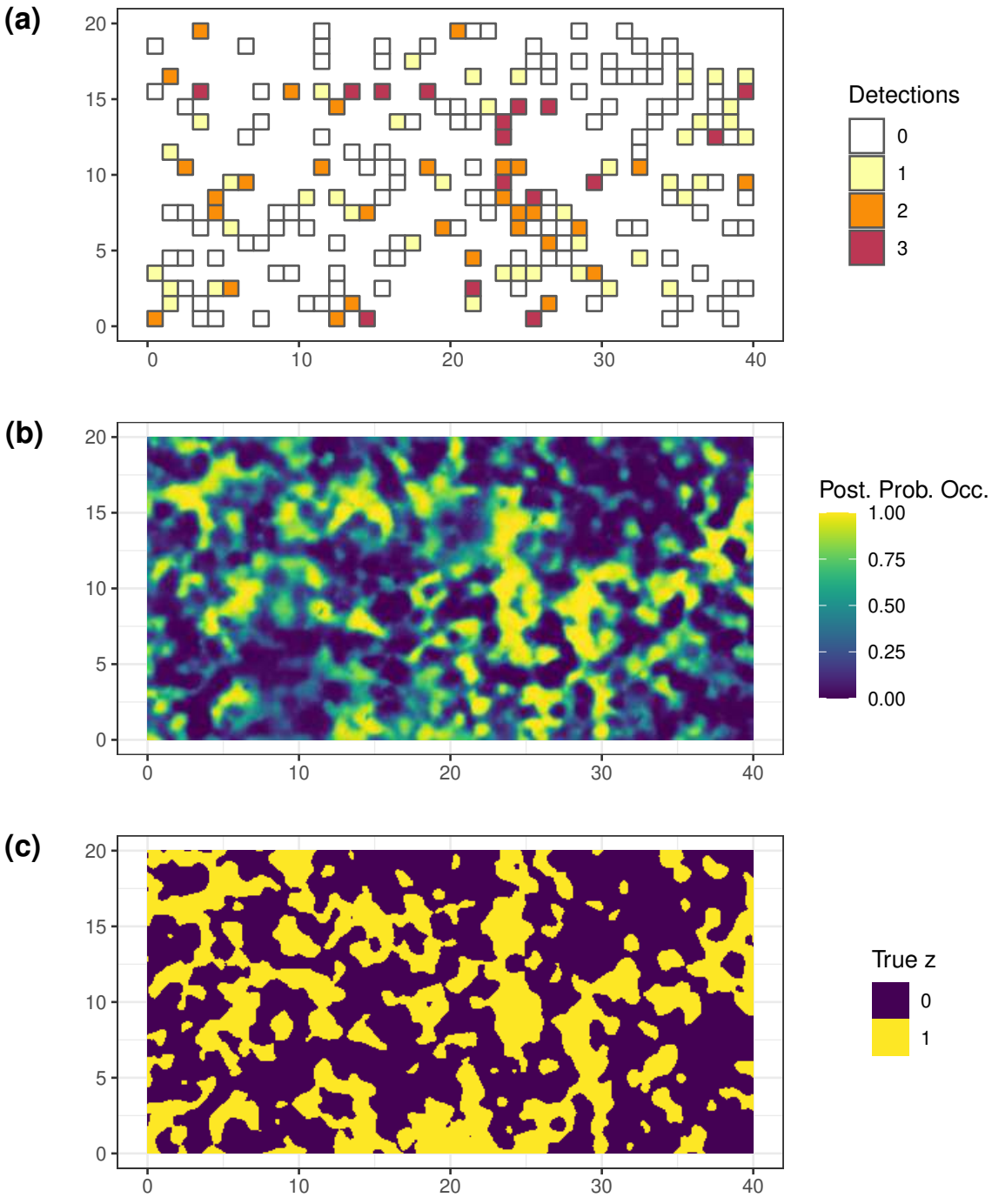

Figure 4.2: Simulated data example with occupancy related to one spatial covariate and an additional covariate related to detection probabilities. The number of observed detections out of three visits is shown in (a) for the sampled sites. After fitting our model, the posterior probability of occupancy (b) is well-aligned with the true underlying occupancy (c).

4.5.2 Comparisons to other models

We also performed simulations to compare our continuous-space occupancy model to other spatial occupancy models. Data were generated using the same parameter values as specified in Section 4.5.1. We defined sites using a 1×1 unit regular grid and considered a study area that was 20×15 units in size. Data were generated for three visits at each of the 300 sites within this region. We considered a smaller region and surveys at all sites to simplify the computation in this simulation study.

We generated 100 different datasets and fit three different spatial occupancy models to each. We first fit our continuous-space occupancy model that matches the data-generating process. The first alternative approach models the occupancy process at the areal sites and includes spatial structure among sites using an intrinsic conditional autoregressive (ICAR) model (Johnson et al., 2013). The second alternative approach ignores the areal support of the survey data and treats each site as a point location at the site centroid. Spatial dependence in occupancy is included in the second alternative model using a clipped Gaussian process to make it comparable to our continuous-space model. Neither of these alternative approaches accounts for the change of spatial support between the occupancy process and the observed data. Consequently, these alternative approaches are unable to model how the within-site occurrence proportions lead to spatial heterogeneity in detection probabilities. Additional details for these alternative models and their implementation can be found in Appendix C.3.

We compared the different models by considering the bias of the posterior means and coverage of the 95% credible intervals (CIs) for both the proportion of area occupied and the proportion of sites occupied. For our continuous-space occupancy model, we calculated the proportion of the area occupied as $D^{-1} \sum_{d=1}^{D} z(\mathbf{s}_d)$ at each MCMC iteration where the locations $z(\mathbf{s}_d)$ are used to approximate the binary spatial process. If the proportion of sites occupied is of interest, we can use our continuous-space model to make inferences at the site-level by summarizing the spatial terms to the areal survey units. We calculated the proportion of sites occupied as $n^{-1} \sum_{i=1}^{n} \max_{\mathbf{s}_d \in \mathcal{A}_i} z(\mathbf{s}_d)$ for our continuous-space model. For the areal occupancy model, the proportion of area occupied

Table 4.1: For the proportion of area occupied and proportion of sites occupied, the empirical bias of posterior means and the coverage of 95% CIs for different spatial occupancy models based on analyses of 100 simulated datasets. Data were simulated under our continuous-space occupancy model. We compared our model to two other approaches that both ignore the change of spatial support. The first approach ("Areal") models the occupancy process at the areal sites used to collect data. The second approach ("Centroid") ignores the defined sites and treats all data as point-level observations corresponding to the site centroids.

	Proportion of Area Occupied		Proportion of Sites Occupied	
Model	Bias	Coverage	Bias	Coverage
Areal	0.143	0%	-0.268	0%
Centroid	0.156	0%	-0.255	3%
Continuous-space	-0.002	94%	-0.010	92%

and proportion of sites occupied are equivalent and calculated as $n^{-1} \sum_{i=1}^{n} z_i$. This same calculation can be used to calculate the proportion of sites occupied under the centroid occupancy model as well. For the proportion of area occupied using the centroid occupancy model, we considered posterior predictions for the same grid of points used to implement our continuous-space model and calculated this quantity as $D^{-1} \sum_{d=1}^{D} z(\mathbf{s}_d)$ (see Appendix C.3 for more details).

Both of the alternative models were biased and had minimal coverage for both the proportion of area occupied and the proportion of sites occupied (Table 4.1). Our continuous-space occupancy model was unbiased and had high coverage for the proportion of area occupied. Our approach also provided unbiased inferences for the proportion of sites occupied and had high coverage (Table 4.1). Both of the alternative approaches were positively biased for the proportion of area occupied because they fail to account for the change of spatial support. However, even considering the proportion of sites occupied resulted in biased inferences from these models — due to unaccounted for heterogeneity in detection probabilities resulting from the species only occurring in a portion of a site.

4.6 Avian data application

We analyzed detection/nondetection data for ovenbirds (*Seiurus aurocapilla*) collected during the summer of 2015 in the Hubbard Brook Experimental Forest, New Hampshire, USA. These data are part of ongoing bird surveys within the experimental forest (Rodenhouse and Sillet, 2019)

and are available in the `spOccupancy` R package (Doser et al., 2022). These data include two visit-level predictor variables — time of day and survey date — to model detection probabilities. We obtained elevation data for the study region using the `elevatr` R package (Hollister et al., 2023) to use as a spatial predictor variable for occupancy. All predictor variables were standardized to have mean 0 and standard deviation of 1 prior to fitting the model.

The Hubbard Brook Experimental Forest is located in a valley in the White Mountains (Figure 4.3a). The detection/nondetection data were obtained from 10 minute point count surveys of circular sites with 100 meter radii (Figure 4.3b). Most sites were visited 3 times, but a few sites only had 1 or 2 visits. Note that some of the site areas overlap (Figure 4.3b) and that the sites do not form a regular grid over the study region. Standard occupancy models cannot account for the overlapping sites in these surveys. Typically, this feature of the data would need to be ignored (by treating sites as point locations) or by excluding data from some of the sites that overlap. Our continuous space occupancy model, on the other hand, is able to include all the observed data and naturally accommodate the overlapping sites when making inferences about species occurrence.

We fit our model to these data using 1 million MCMC iterations. The first half were discarded as burn-in after using them to tune the surrogate data variances and the final 500,000 iterations were saved for inferences. We thinned the iterations by 100 to reduce the amount of memory required to save the results. In this example we assumed a Gaussian spatial covariance function so that

$$K(\mathbf{s}_i, \mathbf{s}_{i'}) = \exp\left(-\frac{||\mathbf{s}_i - \mathbf{s}_{i'}||^2}{2\rho^2}\right), \tag{4.22}$$

where ρ is the spatial range parameter. Note that we fixed the variance of this covariance function to be 1 for identifiability of the clipped Gaussian process (De Oliveira, 2000, 2020). We checked convergence using traceplots for all parameters and summarized posterior inferences using posterior means and 95% CIs.

There was no evidence that detection probabilities varied with the predictor variables date (α_1: posterior mean = -0.03 and 95% CI = (-0.12, 0.06)) or time of day (α_2: posterior mean = -0.03 and 95% CI = (-0.13, 0.06)). There was strong evidence that detection probabilities increased as the

proportion of the site that was occupied increased (γ: posterior mean = 1.81, 95% CI = (1.24, 2.51)). Consequently, the posterior mean detection probability at sites with 15% occurrence was 0.22 (95% CI from 0.07 to 0.37) while that for sites with 85% occurrence was 0.67 (95% CI from 0.61 to 0.73). This suggests there is substantial heterogeneity in the detection probabilities due to the variability of within-site species occurrence.

Our analysis also provided strong evidence that ovenbird occurrence was negatively related to elevation (β_1: posterior mean = -0.62, 95% CI = (-0.94, -0.35)). We obtained posterior inferences on the probability of ovenbird occurrence across the study region, including for areas that were not surveyed. There was evidence of additional spatial variability that was not due to elevation (Figure 4.3c). We also were able to map the distribution of the species in this region based on the posterior probability of ovenbird occupancy (Figure 4.3d).

4.7 Discussion

We developed a new spatial occupancy model for wildlife species. While standard surveys for wildlife species collect detection/nondetection data over areal survey sites, species occupancy is a process that can be defined in continuous space (Efford and Dawson, 2012). Our approach is the first to treat species occurrence in continuous space while accounting for the change of support required for analyzing the areal survey data. Additionally, our model accounts for imperfect detection and allows for heterogeneity in detection probabilities related to the proportion of the site that is occupied. This detection process is similar to that in other occupancy models that incorporate a detection-abundance relationship (Royle and Nichols, 2003) or heterogeneity in detection probabilities using a mixture model (Royle, 2006). Our real data analysis provided strong evidence that detection probabilities of ovenbirds increased as the proportion of the site area that is occupied increased. Failing to account for this variability in detection probabilities can lead to biased inferences even when considering the proportion of sites occupied, as shown in our simulation study.

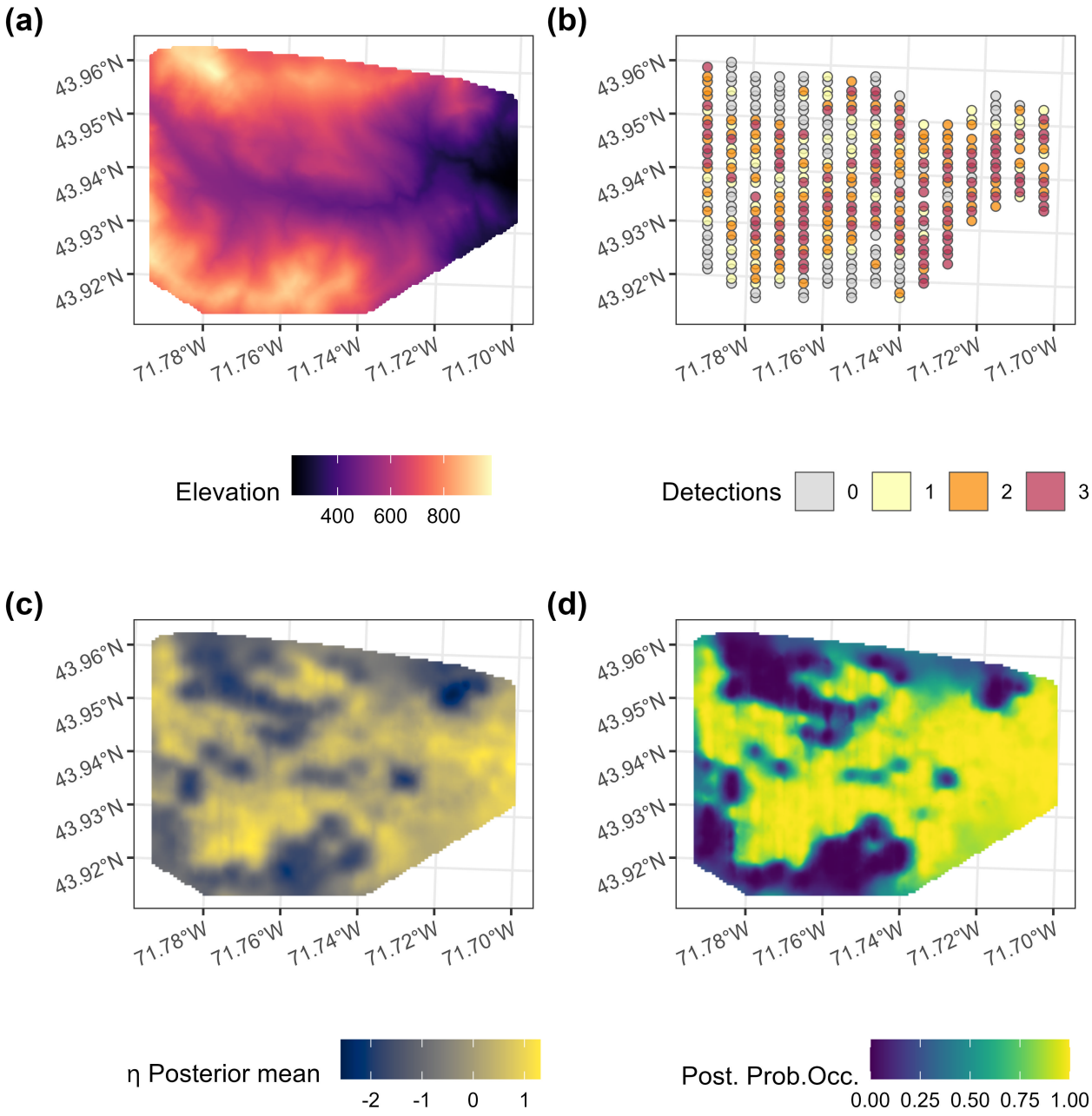

Figure 4.3: Elevation (m) in the Hubbard Brook Experimental Forest (a) and the observed detection/nondetection data for ovenbirds (*Seiurus aurocapilla*) at 100 m radius plots (b). Most sites had 3 visits, but some had only 1 or 2 total visits. After fitting our continuous-space occupancy model, the posterior mean of the spatial effects (c) and the posterior probability of ovenbird occupancy (d) within the study region.

The biggest limitation of our model is the increased computational burden compared to standard occupancy models. We used a nearest neighbor approximation of the clipped Gaussian process that models species occurrence. This increased the computational efficiency of our spatial model and made Bayesian inferences using MCMC feasible. Additionally, we modified the surrogate data slice sampler developed by Murray and Adams (2010) to better accommodate NNGPs. Our version of the sampler introduces surrogate data for only a portion of the spatial locations and is able harness the computational advantages provided by NNGPs for the remaining spatial locations. While our approach was motivated by our continuous-space occupancy model, it should be noted that the surrogate data slice sampler is applicable to any model that includes a latent spatial Gaussian process. This is because all steps update parameters from multivariate normal distributions, resulting from the Gaussian process prior, except when updating the covariance parameters using slice sampling. The slice sampling step is general and can be applied to any assumed data model. Overall, the surrogate data slice sampler provides an efficient approach for jointly updating spatial terms and spatial covariance parameters when modeling data using a latent NNGP.

We defined the spatial terms in (4.3) using a generic covariance function. In some applications it will be useful to assume occupancy is a multiscale process and model the spatial terms as

$$\eta(\mathbf{s}) = \sum_{m=1}^{M} \eta_m(\mathbf{s}), \tag{4.23}$$

$$\eta_m(\mathbf{s}) \sim \text{GP}(0, K_{\eta_m}), \tag{4.24}$$

where the η_m for $m = 1, \ldots, M$ are assumed to be independent of one another and their corresponding spatial covariance functions K_{η_m} have different parameters, including different range parameters. The independence assumption implies that the overall covariance function K_η is equal to $\sum_{m=1}^{M} K_{\eta_m}$. This allows the spatial dependence in the species occupancy process to vary across different spatial scales. Such a multiscale model in continuous space differs from current multiscale occupancy models (e.g., Nichols et al., 2008) and future research could compare inferences from these different approaches.

Other variations to standard occupancy models exist that account for multiple species, multiple seasons, and false positive detections (Bailey et al., 2014). These ideas could also be incorporated into our continuous-space occupancy model. For instance, it would be straightforward to construct a model for multiple seasons by assuming each season is a discrete-time snapshot of species occupancy in continuous space. An alternative would be to consider using a clipped Gaussian process for modeling species occupancy in continuous space-time. Care would be needed to ensure detection probabilities are identifiable given the available survey data. This could require concurrent visits to a site or visits close together in time relative to the effective range of temporal covariance function (analogous to the closure assumptions of standard models). Modeling occupancy in continuous space-time could allow for improved insights into how species occurrence changes over time due to changing environments.

Chapter 5

Conclusion

I developed three new models for analyzing spatial data from ecological and environmental applications. I begin this chapter with an overview of the previous chapters. The discussion sections of Chapters 2–4 each included possible directions for future research, and the final section of this chapter includes additional ideas that could build upon the models I developed.

5.1 Overview

In this dissertation, I developed new Bayesian spatial models for applications in ecology and environmental science. Chapter 2 showed how a PDE for atmospheric dispersion can be used to account for the spatial dependence in heavy metal concentrations. This allowed mechanistic knowledge about the spread of heavy metal pollutants to be included in my statistical analysis and provided improved forecasts for how the spatial distribution of heavy metals may change for scenarios of interest. In Chapter 3, I developed a model for plant cover that accounted for multiscale spatial dependence as well as associations among different species. This model used a clipped Gaussian process to model the binary spatial process of plant cover for each species. I also used a clipped Gaussian process to analyze species occurrence data in Chapter 4. My continuous-space occupancy framework accounts for the change of support between the occurrence process and the observed data. It also accounts for imperfect detection and relates the probability of detecting the species to the within-site occupancy proportions. In Chapter 4, I also developed a surrogate data slice sampler (Murray and Adams, 2010) for jointly updating spatial covariance parameters and spatial terms. This MCMC algorithm is beneficial because it can be applied to nearest neighbor Gaussian processes for big spatial data.

Ecological and environmental research is relied upon for making critical conservation decisions. These decisions should be informed by statistical analyses that are able to make the most of the available data. The models I developed in this dissertation place an emphasis on accounting for the

spatial dependence that is a fundamental component of ecological data. Advancing the statistical methods available for these applications helps to improve inferences for conservation research. In some cases, new statistical models even allow new applied research questions to be explored.

5.2 Future directions

5.2.1 Joint model for heavy metals and plant cover

The mechanistic spatial model for heavy metal pollutants from Chapter 2 could be combined with the plant cover model from Chapter 3. A single hierarchical model would appropriately propagate the uncertainty in heavy metal concentrations when modeling its impacts on the vegetation community. Additionally, the plant cover data provide some information about the heavy metal concentrations at a location based on the relationships between concentration and cover for each species. Thus, the plant cover data could be used to improve monitoring of pollution levels.

A joint model for these data could be most beneficial if future data are collected using a different sampling design. Currently, the heavy metal data and point intercept data are collected at the same locations (Neitlich et al., 2017). It may be more efficient to only collect one type of data at certain spatial locations. For instance, if the point intercept data are easier and/or cheaper to collect than the heavy metal data, sampling additional locations to collect only plant cover data could be beneficial. These additional data would be able to inform inferences about the relationship between cover and pollution concentration even if heavy metal concentrations are not measured at that location. Another aspect of the sampling design that could be considered is the grid configuration for the point intercept data. One recommendation for point intercept data is to space grid points far apart from one another so that they are approximately independent (Drezner and Drezner, 2021). However, having observations close to one another can improve learning about the spatial covariance parameters associated with Gaussian process models (Zhu and Stein, 2006; Zimmerman, 2006). Changing the sampling design to collect point intercept data at points closer together and over an irregular grid may improve inferences for plant cover when using the model I developed in Chapter 3.

Future work could evaluate different sampling designs based on a joint model for the heavy metals and plant cover. Designing ecological monitoring programs based on preliminary spatio-temporal data can help improve inferences (e.g., Wikle and Royle, 2005; Hooten et al., 2009). Previous studies have often focused on Gaussian process models and there has been less research on sampling designs for mechanistic models. However, Williams et al. (2018) and Leach et al. (2022) illustrate how an adaptive optimal design framework can be used to select a sampling protocol assuming a mechanistic model based on ecological diffusion will be used to analyze the data. For the heavy metal study, the atmospheric dispersion model could be used to better inform the sampling design compared to assuming a phenomenological model.

A challenge with implementing a joint model for these data would be the increased computational costs. In particular, solving the atmospheric dispersion PDE at each iteration of the MCMC algorithm is computationally expensive. In Chapter 2 I used the finite difference method to solve the PDE, but other approaches could be used as well. For instance, the finite element method may improve the computation times for fitting this model. Another strategy for implementing mechanistic models is to use an emulator (or surrogate) model (e.g., Liu and West, 2009). This approach fits a statistical model, such as a Gaussian process, to inputs (parameters) and outputs (solutions) to a PDE model. The statistical model can then be used to approximate the mechanistic model given parameter values of interest. Exploring these alternative computational strategies could be beneficial for implementing a joint model for the heavy metal pollutants and plant cover.

5.2.2 Bayesian computation for Gaussian process models

Chapter 4 showed how surrogate data slice sampling (Murray and Adams, 2010) could be used when implementing Gaussian process models. In particular, I developed a modification of this MCMC algorithm that was compatible with nearest neighbor Gaussian processes. Additional considerations that I did not thoroughly explore include how to optimally select the number of locations to introduce surrogate data for and how to tune the surrogate data variance. Future research

could explore different strategies to provide more guidance on how to optimize the efficiency of this algorithm for big spatial data.

I also used elliptical slice sampling to update the spatial terms in my continuous-space occupancy model. One limitation of this MCMC algorithm is that it can be slow to mix when the posterior distribution is very different from the prior distribution (Nishihara et al., 2014; Fagan et al., 2016). To avoid this limitation, Nishihara et al. (2014) and Fagan et al. (2016) developed a generalized version of elliptical slice sampling that can improve MCMC mixing. However, this approach can be difficult to tune and its performance is sensitive to the tuning parameters. It also can increase the computing time of each iteration. An alternative strategy to modify elliptical slice sampling is to add an additional level to the parameter expanded model that allows the ellipse center to vary at each iteration. This modified model is

$$\alpha \sim \text{Uniform}(0, 2\pi), \tag{5.1}$$

$$\boldsymbol{\nu}_0 \sim \text{N}(\boldsymbol{\mu}, a\boldsymbol{\Sigma}), \tag{5.2}$$

$$\boldsymbol{\nu}_1 \sim \text{N}(\boldsymbol{\nu}_0, (1-a)\boldsymbol{\Sigma}), \tag{5.3}$$

$$\boldsymbol{\nu}_2 \sim \text{N}(\boldsymbol{\nu}_0, (1-a)\boldsymbol{\Sigma}), \tag{5.4}$$

$$\boldsymbol{\theta} = (\boldsymbol{\nu}_1 - \boldsymbol{\nu}_0)\text{sin}(\alpha) + (\boldsymbol{\nu}_2 - \boldsymbol{\nu}_0)\text{cos}(\alpha) + \boldsymbol{\nu}_0, \tag{5.5}$$

$$\mathbf{y} \mid \boldsymbol{\theta} \sim [\mathbf{y} \mid \boldsymbol{\theta}], \tag{5.6}$$

where $a \in (0, 1)$ is a tuning parameter. The marginal prior distribution for $\boldsymbol{\theta}$ is still $\text{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and there are no restrictions on the form of the data model $[\mathbf{y} \mid \boldsymbol{\theta}]$. Note that if $a = 0$, then this model is equivalent to the model defined by (1.22) – (1.26) and therefore it is a generalization of standard elliptical slice sampling. However, now $\boldsymbol{\theta}$ lies on an ellipse centered at the parameter $\boldsymbol{\nu}_0$ instead of the prior mean $\boldsymbol{\mu}$. Compared to the standard algorithm, implementing this approach only requires one additional step that updates the ellipse center $\boldsymbol{\nu}_0$ from the distribution $[\boldsymbol{\nu}_0 \mid \alpha, \boldsymbol{\theta}, \mathbf{y}]$.

This modification to elliptical slice sampling could be beneficial for two reasons. First, the ellipses that define the slices used to update the parameter $\boldsymbol{\theta}$ can be smaller and thus include

relatively larger segments in regions of high posterior density (Figure 5.1 shows a toy example). This means that fewer evaluations of the likelihood would be needed in the shrinkage algorithm that is typically used to implement slice sampling (Neal, 2003). Second, variability in the ellipse centers results in more variability in the slices which could improve MCMC mixing. Future research is needed to explore if and when this approach is beneficial in practice. Additionally, guidance on how to tune the parameter a for different models would be useful as well. Other variations of elliptical slice sampling that build upon this idea could also be possible.

5.2.3 Modeling correlation among different species

Multispecies models are increasingly of interest in ecology (e.g., Ovaskainen et al., 2016; Ovaskainen and Abrego, 2020). In Chapter 3, I used a latent factor approach to account for correlations among species when modeling plant cover data. This approach is commonly used to model multispecies data because it can be computationally efficient when fitting models to a

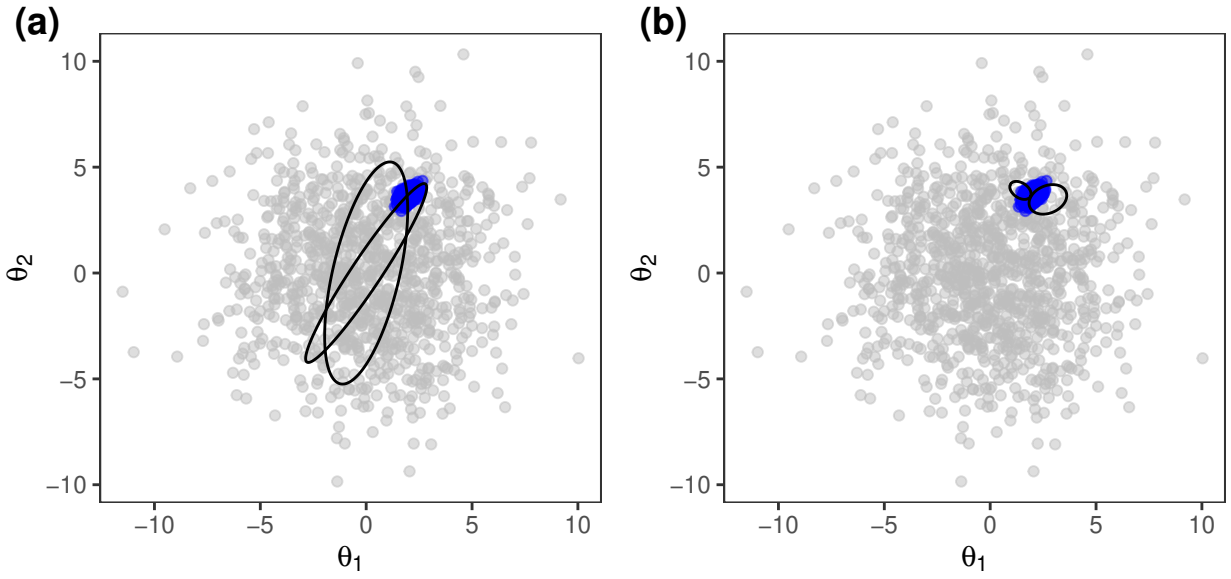

Figure 5.1: Hypothetical example of standard elliptical slice sampling (a) and a recentered variation (b) for updating a two-dimensional parameter $\boldsymbol{\theta}$. Gray points show a random sample from the prior distribution and the blue points show a random sample from the target posterior distribution. For standard elliptical slice sampling (a), the ellipses used to update parameters are always centered at the prior mean. This algorithm can be modified to allow the ellipse centers to also vary (b).

large number of species (Ovaskainen et al., 2016; Ovaskainen and Abrego, 2020). However, there are limitations to this approach and room for improving the statistical models used to analyze community data in spatial ecology.

One limitation of latent factor models for modeling ecological communities is that they can be unable to account for many negative correlations among species. For instance, consider using two latent factors to model correlations among four species. One way to interpret the factor loadings is to plot them as vectors in the two-dimensional space corresponding to the latent factors. Species that have factor loadings pointing in opposite directions are negatively correlated, while species with factor loadings in similar directions are positively correlated (Figure 5.2). Specifically, the covariance for a pair of species is the dot product of the corresponding factor loadings. While every pair of these four species could be strongly negatively correlated (corresponding to a valid covariance matrix), this is not possible with only two latent factors (e.g., Figure 5.2). Consequently, this approach may not be suitable for modeling communities where many of the species are

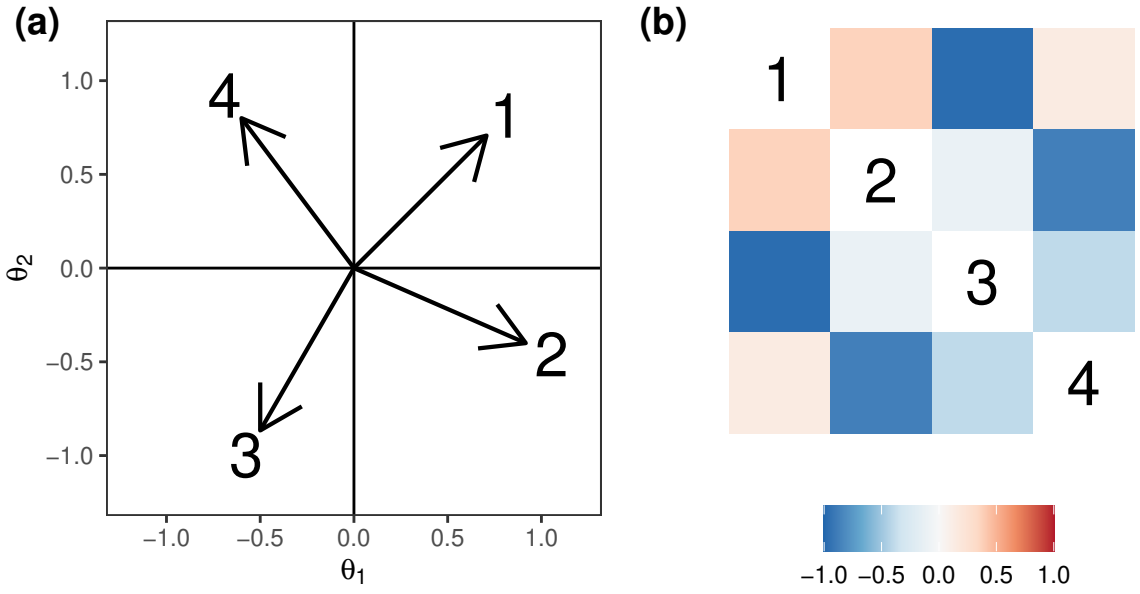

Figure 5.2: Example of latent factor model for four species and two latent factors. The factor loadings (a) have a geometric interpretation that can help explain the corresponding covariance among species (b). Species that have vectors pointing in similar directions are positively correlated while species with vectors pointing in opposite directions are negatively correlated. In lower dimensions, it can be difficult to include many negative correlations among species.

competitors of one another. There may be empirical examples of this limitation as well — Tobler et al. (2019) analyzed bird communities using a latent factor model and noted that some groups of species were estimated to be positively correlated even though they were competitors and should be excluding each other. Additionally, in Chapter 3 I found evidence of strong negative correlations among species and my model required a large number of latent factors to adequately account for the community structure. These examples suggest that other statistical methods may be needed for modeling some ecological communities.

Mechanistically motivated models may provide an alternative approach for modeling the dependence among species. For instance, Tang et al. (2023) developed a mechanistic model for a fish community that included dependence among species based on knowledge of the food web in that study system. Other mechanistic models could be developed for describing ecological communities more generally. Competitive exclusion, that should result in many negative correlations among species, can be explained by limited resources that prevent multiple species from occurring at the same location. Including this idea in a statistical model would provide a new way to mechanistically model the dependence among species. Space itself may be a shared resource and limit the number of different species that can co-occur. Accounting for this may be useful for modeling the dependence among plant species (e.g., Chapter 3) at small spatial scales if many negative correlations among species are expected.

Another consideration for modeling ecological communities is that dependence among species may occur at multiple spatial scales. The model in Chapter 3 assumed that the latent factors were correlated at a single spatial scale. Consequently, the dependence among species was only modeled at this scale. This may be unrealistic for some applications. For instance, the occurrence of two species may be positively correlated at large spatial scales (e.g., species co-occur regionally) but negatively correlated at small spatial scales (e.g., they compete locally). Additionally, the model I developed for plant cover allows the correlation among species to vary with predictor variables but there could be additional variability across space. New statistical methods that can account for

and model these complex spatial structures in species-to-species associations are needed for better understanding ecological communities.

Bibliography

Agresti, A. (2012). *Categorical Data Analysis*. John Wiley & Sons.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.

Bailey, L. L., MacKenzie, D. I., and Nichols, J. D. (2014). Advances and applications of occupancy models. *Methods in Ecology and Evolution*, 5(12):1269–1279.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(4):825–848.

Berg, T. and Steinnes, E. (1997). Use of mosses (*Hylocomium splendens* and *Pleurozium schreberi*) as biomonitors of heavy metal deposition: from relative to absolute deposition values. *Environmental Pollution*, 98:61–71.

Berger, J. O. (2013). *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media.

Berrett, C. and Calder, C. A. (2016). Bayesian spatial binary classification. *Spatial Statistics*, 16:72–102.

Bertness, M. D. and Callaway, R. (1994). Positive interactions in communities. *Trends in Ecology & Evolution*, 9(5):191–193.

Bradley, J. R., Wikle, C. K., and Holan, S. H. (2016). Bayesian spatial change of support for count-valued survey data with application to the American Community Survey. *Journal of the American Statistical Association*, 111(514):472–487.

Briffa, J., Sinagra, E., and Blundell, R. (2020). Heavy metal pollution in the environment and their toxicological effects on humans. *Heliyon*, 6(9):e04691.

Broms, K. M., Johnson, D. S., Altwegg, R., and Conquest, L. L. (2014). Spatial occupancy models applied to atlas data show Southern Ground Hornbills strongly depend on protected areas. *Ecological Applications*, 24(2):363–374.

Brumbaugh, W. G., Mora, M. A., May, T. W., and Phalen, D. N. (2010). Metal exposure and effects in voles and small birds near a mining haul road in Cape Krusenstern National Monument, Alaska. *Environmental Monitoring and Assessment*, 170:73 – 86.

Brumbaugh, W. G., Morman, S. A., and May, T. W. (2011). Concentrations and bioaccessibility of metals in vegetation and dust near a mining road, Cape Krusenstern National Monument, Alaska. *Environmental Monitoring and Assessment*, 182:325–340.

Buckland, S. T., Oedekoven, C. S., and Borchers, D. L. (2016). Model-based distance sampling. *Journal of Agricultural, Biological, and Environmental Statistics*, 21:58–75.

Burnham, K. P., Anderson, D. R., and Laake, J. L. (1980). Estimation of density from line transect sampling of biological populations. *Wildlife monographs*, 72:3–202.

Carlin, B. P. and Louis, T. A. (2008). *Bayesian Methods for Data Analysis*. CRC press.

Chambert, T., Miller, D. A., and Nichols, J. D. (2015). Modeling false positive detections in species occurrence data under different study designs. *Ecology*, 96(2):332–339.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, 85(2):347–361.

Cressie, N. (2015). *Statistics for Spatial Data*. John Wiley & Sons.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70:209–226.

Cressie, N. and Wikle, C. K. (2015). *Statistics for Spatio-Temporal Data*. John Wiley & Sons.

Cressie, N. A. (1996). Change of support and the modifiable areal unit problem. *Geographical Systems*, 3:159–180.

Csavina, J., Field, J., Taylor, M. P., Gao, S., Landázuri, A., Betterton, E. A., and Sáez, A. E. (2012). A review on the importance of metals and metalloids in atmospheric dust and aerosol from mining operations. *Science of the Total Environment*, 433:58–73.

Damgaard, C., Strandberg, B., Ehlers, B., Hansen, R. R., and Strandberg, M. T. (2022). Effect of nitrogen and glyphosate on the plant community composition in a simulated field margin ecosystem: Model-based ordination of pin-point cover data. *Environmental Pollution*, 315:120377.

Damgaard, C. F. and Irvine, K. M. (2019). Using the beta distribution to analyse plant cover data. *Journal of Ecology*, 107(6):2747–2759.

Datta, A. (2022). Nearest-neighbor sparse Cholesky matrices in spatial statistics. *Wiley Interdisciplinary Reviews: Computational Statistics*, 14(5):e1574.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812.

De Oliveira, V. (2000). Bayesian prediction of clipped Gaussian random fields. *Computational Statistics and Data Analysis*, 34:299–314.

De Oliveira, V. (2020). Models for geostatistical binary data: Properties and connections. *The American Statistician*, 74(1):72–79.

Diggle, P. J. and Ribeiro Jr, P. J. (2002). Bayesian inference in Gaussian model-based geostatistics. *Geographical and Environmental Modelling*, 6(2):129–146.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 47:299–350.

Donovan, G. H., Jovan, S. E., Gatziolis, D., Burstyn, I., Michael, Y. L., Amacher, M. C., and Monleon, V. J. (2016). Using an epiphytic moss to identify previously unknown sources of atmospheric cadmium pollution. *Science of The Total Environment*, 559:84–93.

Dorazio, R. M. and Royle, J. A. (2005). Estimating size and composition of biological communities by modeling the occurrence of species. *Journal of the American Statistical Association*, 100(470):389–398.

Doser, J. W., Finley, A. O., and Banerjee, S. (2023). Joint species distribution models with imperfect detection for high-dimensional spatial data. *Ecology*, 104(9):e4137.

Doser, J. W., Finley, A. O., Kéry, M., and Zipkin, E. F. (2022). spOccupancy: An R package for single-species, multi-species, and integrated spatial occupancy models. *Methods in Ecology and Evolution*, 13:1670–1678.

Drezner, T. D. and Drezner, Z. (2021). Informed cover measurement: Guidelines and error for point-intercept approaches. *Applications in Plant Sciences*, 9(9-10):e11446.

Eddelbuettel, D. and Balamuta, J. J. (2018). Extending R with C++: A Brief Introduction to Rcpp. *The American Statistician*, 72(1):28–36.

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.

Eddelbuettel, D., François, R., Bates, D., Ni, B., and Sanderson, C. (2023). *RcppArmadillo: 'Rcpp' Integration for the 'Armadillo' Templated Linear Algebra Library*. R package version 0.12.4.0.0.

Eddelbuettel, D. and Sanderson, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis*, 71:1054–1063.

Efford, M. G. and Dawson, D. K. (2012). Occupancy in continuous habitat. *Ecosphere*, 3(4):1–15.

Elith, J. and Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, 40:677–697.

Ersoy, A., Yunsel, T., and Cetin, M. (2004). Characterization of land contaminated by past heavy metal mining using geostatistical methods. *Archives of Environmental Contamination and Toxicology*, 46(2):162–175.

Fagan, F., Bhandari, J., and Cunningham, J. P. (2016). Elliptical slice sampling with expectation propagation. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, page 172–181.

Ferreira, M. A. R. and Lee, H. K. (2007). *Multiscale Modeling: A Bayesian Perspective*, volume 2. Springer.

Finley, A. O., Datta, A., Cook, B. D., Morton, D. C., Andersen, H. E., and Banerjee, S. (2019). Efficient algorithms for Bayesian nearest neighbor Gaussian processes. *Journal of Computational*

and Graphical Statistics, 28(2):401–414.

Fonseca, T. C. and Ferreira, M. A. (2017). Dynamic multiscale spatiotemporal models for Poisson data. *Journal of the American Statistical Association*, 112(517):215–234.

Gelfand, A. E. (2022). Spatial modeling for the distribution of species in plant communities. *Spatial Statistics*, 50:100582.

Gelfand, A. E. and Schliep, E. M. (2016). Spatial statistics and Gaussian processes: A beautiful marriage. *Spatial Statistics*, 18:86–104.

Gelfand, A. E. and Shirota, S. (2019). Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. *Ecological Monographs*, 83(3):e01372.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Gelfand, A. E., Zhu, L., and Carlin, B. P. (2001). On the change of support problem for spatio-temporal data. *Biostatistics*, 2(1):31–45.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*. Chapman & Hall/CRC.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, PAMI-6(6):721–741.

Ghosh, J. and Dunson, D. B. (2009). Default prior distributions and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics*, 18(2):306–320.

Godínez-Alvarez, H., Herrick, J., Mattocks, M., Toledo, D., and Van Zee, J. (2009). Comparison of three vegetation monitoring methods: Their relative utility for ecological assessment and monitoring. *Ecological Indicators*, 9(5):1001–1008.

Gotway, C. A. and Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, 97:632–648.

Guinness, J. (2018). Permutation and grouping methods for sharpening Gaussian process approximations. *Technometrics*, 60(4):415–429.

Guisan, A. and Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, 8(9):993–1009.

Hanks, E. M. (2017). Modeling spatial covariance using the limiting distribution of spatio-temporal random walks. *Journal of the American Statistical Association*, 112:497–507.

Hanks, E. M., Schliep, E. M., Hooten, M. B., and Hoeting, J. A. (2015). Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification. *Environmetrics*, 26(4):243–254.

Hasselbach, L., Ver Hoef, J., Ford, J., Neitlich, P., Crecelius, E., Berryman, S., Wolk, B., and Bohle, T. (2005). Spatial patterns of cadmium and lead deposition on and adjacent to National Park Service lands in the vicinity of Red Dog Mine, Alaska. *Science of The Total Environment*, 348:211–230.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., et al. (2019). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 24:398–425.

Hefley, T. J. and Hooten, M. B. (2016). Hierarchical species distribution models. *Current Landscape Ecology Reports*, 1:87–97.

Hefley, T. J., Hooten, M. B., Hanks, E. M., Russell, R. E., and Walsh, D. P. (2017a). The Bayesian group lasso for confounded spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 22:42–59.

Hefley, T. J., Hooten, M. B., Hanks, E. M., Russell, R. E., and Walsh, D. P. (2017b). Dynamic spatio-temporal models for spatial data. *Spatial Statistics*, 20:206–220.

Hefley, T. J., Hooten, M. B., Russell, R. E., Walsh, D. P., and Powell, J. A. (2017c). When mechanism matters: Bayesian forecasting using models of ecological diffusion. *Ecology Letters*, 20:640–650.

Higgs, M. D. and Hoeting, J. A. (2010). A clipped latent variable model for spatially correlated ordered categorical data. *Computational Statistics & Data Analysis*, 54(8):1999–2011.

Hodges, J. S. and Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, 64(4):325–334.

Hoegh, A., Ferreira, M. A., and Leman, S. (2016). Spatiotemporal model fusion: multiscale modelling of civil unrest. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 65(4):529–545.

Hoeting, J. A., Leecaster, M., and Bowden, D. (2000). An improved model for spatially correlated binary responses. *Journal of Agricultural, Biological, and Environmental Statistics*, 5:102–114.

Hollister, J., Shah, T., Nowosad, J., Robitaille, A. L., Beck, M. W., and Johnson, M. (2023). *elevatr: Access Elevation Data from Various APIs*. R package version 0.99.0.

Hooten, M. B. and Hefley, T. (2019). *Bringing Bayesian Models to Life*. CRC Press.

Hooten, M. B., Larsen, D. R., and Wikle, C. K. (2003). Predicting the spatial distribution of ground flora on large domains using a hierarchical Bayesian model. *Landscape Ecology*, 18:487–502.

Hooten, M. B. and Wikle, C. K. (2008). A hierarchical Bayesian non-linear spatio-temporal model for the spread of invasive species with application to the Eurasian Collared-Dove. *Environmental and Ecological Statistics*, 15:59–70.

Hooten, M. B., Wikle, C. K., Sheriff, S. L., and Rushin, J. W. (2009). Optimal spatio-temporal hybrid sampling designs for ecological monitoring. *Journal of Vegetation Science*, 20:639–649.

Huber, P., Ronchetti, E., and Victoria-Feser, M.-P. (2004). Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(4):893–908.

Irvine, K. M., Wright, W. J., Shanahan, E. K., and Rodhouse, T. J. (2019). Cohesive framework for modelling plant cover class data. *Methods in Ecology and Evolution*, 10(10):1749–1760.

Johnson, D. S., Conn, P. B., Hooten, M. B., Ray, J. C., and Pond, B. A. (2013). Spatial occupancy models for large data sets. *Ecology*, 94(4):801–808.

Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S.-K., Hnilo, J., Fiorino, M., and Potter, G. (2002). NCEP–DOE AMPI-II Reanalysis (R-2). *Bulletin of the American Meteorological Society*, 83(11):1631–1644.

Katzfuss, M. (2017). A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, 112(517):201–214.

Katzfuss, M. and Gong, W. (2020). A class of multi-resolution approximations for large spatial datasets. *Statistica Sinica*, 30(4):2203–2226.

Kéry, M. and Royle, J. A. (2016). *Applied Hierarchical Modeling in Ecology: Analysis of Distribution, Abundance, and Species Richness in R and BUGS*. Academic Press.

Kéry, M., Royle, J. A., Plattner, M., and Dorazio, R. M. (2009). Species richness and occupancy estimation in communities subject to temporary emigration. *Ecology*, 90(5):1279–1290.

Kleiven, E. F., Barraquand, F., Gimenez, O., Henden, J.-A., Ims, R. A., Soininen, E. M., and Yoccoz, N. G. (2023). A dynamic occupancy model for interacting species with two spatial scales. *Journal of Agricultural, Biological and Environmental Statistics*, 28(3):466–482.

Knijnenburg, T. A., Ramsey, S. A., Berman, B. P., Kennedy, K. A., Smit, A. F., Wessels, L. F., Laird, P. W., Aderem, A., and Shmulevich, I. (2014). Multiscale representation of genomic signals. *Nature Methods*, 11(6):689–694.

Latimer, A. M., Wu, S., Gelfand, A. E., and Silander Jr, J. A. (2006). Building statistical models to analyze species distributions. *Ecological Applications*, 16:33–50.

Leach, C. B., Williams, P. J., Eisaguirre, J. M., Womble, J. N., Bower, M. R., and Hooten, M. B. (2022). Recursive bayesian computation facilitates adaptive optimal design in ecological studies. *Ecology*, 103:e03573.

Legendre, P. and Fortin, M. J. (1989). Spatial pattern and ecological analysis. *Vegetatio*, 80:107–138.

Levin, S. A. (1992). The problem of pattern and scale in ecology: The Robert H. MacArthur award lecture. *Ecology*, 73(6):1943–1967.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of*

the Royal Statistical Society: Series B (Statistical Methodology), 73:423–498.

Liu, F. and West, M. (2009). A dynamic modelling strategy for Bayesian computer model emulation. *Bayesian Analysis*, 4:393–411.

Liu, X., Yeo, K., Hwang, Y., Singh, J., and Kalagnanam, J. (2016). A statistical modeling approach for air quality data based on physical dispersion processes and its application to ozone modeling. *The Annals of Applied Statistics*, 10:756 – 785.

Lu, X., Hooten, M. B., Raiho, A. M., Swanson, D. K., Roland, C. A., and Stehn, S. E. (2023). Latent trajectory models for spatio-temporal dynamics in Alaskan ecosystems. *Biometrics*, 79:3664–3675.

Lu, X., Williams, P. J., Hooten, M. B., Powell, J. A., Womble, J. N., and Bower, M. R. (2020). Nonlinear reaction–diffusion process models improve inference for population dynamics. *Environmetrics*, 31:e2604.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Royle, J. A., and Langtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8):2248–2255.

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L., and Hines, J. E. (2018). *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence*. Elsevier.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21:1087–1092.

Mishra, S., Bharagava, R. N., More, N., Yadav, A., Zainith, S., Mani, S., and Chowdhary, P. (2019). Heavy metal contamination: an alarming threat to environment and human health. In *Environmental Biotechnology: For Sustainable Future*, pages 103–125. Springer.

Murray, I., Adams, R., and MacKay, D. (2010). Elliptical slice sampling. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, pages 541–548. PMLR.

Murray, I. and Adams, R. P. (2010). Slice sampling covariance hyperparameters of latent Gaussian models. In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.

Neal, R. M. (1999). Regression and classification using Gaussian process priors (with discussion). *Bayesian Statistics*, 6:475–501.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, 31:705–767.

Neitlich, P. N., Berryman, S., Geiser, L. H., Mines, A., and Shiel, A. E. (2022). Impacts on tundra vegetation from heavy metal-enriched fugitive dust on National Park Service lands along the Red Dog Mine haul road, Alaska. *PLoS ONE*, 17(6):e0269801.

Neitlich, P. N., Ver Hoef, J. M., Berryman, S. D., Mines, A., Geiser, L. H., Hasselbach, L. M., and Shiel, A. E. (2017). Trends in spatial patterns of heavy metal deposition on national park service lands along the Red Dog Mine haul road, Alaska, 2001–2006. *PLoS ONE*, 12:1–35.

Neitlich, P. N., Wright, W., Di Meglio, E., Shiel, A. E., Hampton-Miller, C. J., and Hooten, M. B. (2024). Mixed trends in heavy metal-enriched fugitive dust on National Park Service lands along the Red Dog Mine haul road, Alaska, 2006–2017. *PLoS ONE*, 19(2):e0297777.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135:370–384.

Nichols, J. D., Bailey, L. L., O'Connell Jr, A. F., Talancy, N. W., Campbell Grant, E. H., Gilbert, A. T., Annand, E. M., Husband, T. P., and Hines, J. E. (2008). Multi-scale occupancy estimation and modelling using multiple detection methods. *Journal of Applied Ecology*, 45:1321–1329.

Nishihara, R., Murray, I., and Adams, R. P. (2014). Parallel MCMC with generalized elliptical slice sampling. *The Journal of Machine Learning Research*, 15(1):2087–2112.

Noon, B. R., Bailey, L. L., Sisk, T. D., and McKelvey, K. S. (2012). Efficient species-level monitoring at the landscape scale. *Conservation Biology*, 26(3):432–441.

Ovaskainen, O. and Abrego, N. (2020). *Joint Species Distribution Modelling: with Applications in R*. Cambridge University Press.

Ovaskainen, O., Roy, D. B., Fox, R., and Anderson, B. J. (2016). Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. *Methods in Ecology and Evolution*, 7(4):428–436.

Papastamoulis, P. and Ntzoufras, I. (2022). On the identifiability of Bayesian factor analytic models. *Statistics and Computing*, 32:23.

Phillips, S. J., Anderson, R. P., and Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3-4):231–259.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian Processes for Machine Learning*. MIT press Cambridge, MA.

Ren, Q. and Banerjee, S. (2013). Hierarchical factor models for large spatially misaligned data: A low-rank predictive process approach. *Biometrics*, 69(1):19–30.

Reza, S., Baruah, U., Singh, S., and Das, T. (2015). Geostatistical and multivariate analysis of soil heavy metal contamination near coal mining area, Northeastern India. *Environmental earth sciences*, 73(9):5425–5433.

Ribeiro Jr, P. J. and Diggle, P. J. (2007). The geoR package. *R News*, 1(2):15–18.

Roberts, D. W. (2020). Comparison of distance-based and model-based ordinations. *Ecology*, 101(1):e02908.

Rodenhouse, N. L. and Sillet, S. (2019). Valleywide bird survey, Hubbard Brook Experimental Forest, 1999-2016 (ongoing) ver 3. Environmental Data Initiative. https://doi.org/10.6073/pasta/faca2b2cf2db9d415c39b695cc7fc217.

Royle, J. A. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, 60(1):108–115.

Royle, J. A. (2006). Site occupancy models with heterogeneous detection probabilities. *Biometrics*, 62:97–102.

Royle, J. A. and Nichols, J. D. (2003). Estimating abundance from repeated presence–absence data or point counts. *Ecology*, 84:777–790.

Ruiz-Gutierrez, V., Hooten, M. B., and Campbell Grant, E. H. (2016). Uncertainty in biological monitoring: a framework for data collection and analysis to account for multiple sources of sampling bias. *Methods in Ecology and Evolution*, 7:900–909.

Saha, A. and Datta, A. (2018). BRISC: Bootstrap for rapid inference on spatial covariances. *Stat*, 7(1):e184.

Sang, H., Jun, M., and Huang, J. Z. (2011). Covariance approximation for large multivariate spatial data sets with an application to multiple climate model errors. *The Annals of Applied Statistics*, pages 2519–2548.

Schabenberger, O. and Gotway, C. A. (2017). *Statistical Methods for Spatial Data Analysis*. CRC press.

Schliep, E. M. and Hoeting, J. A. (2013). Multilevel latent Gaussian process model for mixed discrete and continuous multivariate response data. *Journal of Agricultural, Biological, and Environmental Statistics*, 18(4):492–513.

Shi, H., Kang, E. L., Konomi, B. A., Vemaganti, K., and Madireddy, S. (2017). Uncertainty quantification using the nearest neighbor Gaussian process. In Chen, D.-G., Jin, Z., Li, G., Li, Y., Liu, A., and Zhao, Y., editors, *New Advances in Statistics and Data Science*, pages 89–107. Springer, Cham.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Science & Business Media.

Stein, M. L., Chi, Z., and Welty, L. J. (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(2):275–296.

Stockie, J. M. (2011). The mathematics of atmospheric dispersion modeling. *Society for Industrial and Applied Mathematics*, 53:349–372.

Tang, B., Roberts, S. M., Clark, J. S., and Gelfand, A. E. (2023). Mechanistic modeling of climate effects on redistribution and population growth in a community of fish species. *Global Change Biology*, 29(22):6399–6414.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, pages 1701–1728.

Tikhonov, G., Abrego, N., Dunson, D., and Ovaskainen, O. (2017). Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. *Methods in Ecology and Evolution*, 8(4):443–452.

Tobler, M. W., Kéry, M., Hui, F. K., Guillera-Arroita, G., Knaus, P., and Sattler, T. (2019). Joint species distribution models with species correlations and imperfect detection. *Ecology*, 100(8):e02754.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(sup1):234–240.

Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. (2021). Modelling species presence-only data with random forests. *Ecography*, 44(12):1731–1742.

Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 50(2):297–312.

Walker, N. B., Hefley, T. J., Ballmann, A. E., Russell, R. E., and Walsh, D. P. (2021). Recovering individual-level spatial inference from aggregated binary data. *Spatial Statistics*, 44:100514.

Walker, N. B., Hefley, T. J., and Walsh, D. P. (2020). Bias correction of bounded location error in binary data. *Biometrics*, 76(2):530–539.

Wang, W. and Yan, J. (2021). Shape-restricted regression splines with R package splines2. *Journal of Data Science*, 19(3).

Wikle, C. K. (2003). Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology*, 84:1382–1394.

Wikle, C. K. and Hooten, M. B. (2010). A general science-based framework for dynamical spatio-temporal models. *Test*, 19:417–451.

Wikle, C. K., Milliff, R. F., Nychka, D., and Berliner, L. M. (2001). Spatiotemporal hierarchical Bayesian modeling: tropical ocean surface winds. *Journal of the American Statistical Association*, 96:382–397.

Wikle, C. K. and Royle, J. A. (2005). Dynamic design of ecological monitoring networks for non-Gaussian spatio-temporal data. *Environmetrics*, 16(5):507–522.

Wikle, N. B., Hanks, E. M., Henneman, L. R. F., and Zigler, C. M. (2022). A mechanistic model of annual sulfate concentrations in the United States. *Journal of the American Statistical Association*, 117:1082–1093.

Williams, P. J. and Hooten, M. B. (2016). Combining statistical inference and decisions in ecology. *Ecological Applications*, 26(6):1930–1942.

Williams, P. J., Hooten, M. B., Womble, J. N., Esslinger, G. G., and Bower, M. R. (2018). Monitoring dynamic spatio-temporal ecological processes optimally. *Ecology*, 99:524–535.

Wright, W. J., Irvine, K. M., Rodhouse, T. J., and Litt, A. R. (2021). Spatial Gaussian processes improve multi-species occupancy models when range boundaries are uncertain and nonoverlapping. *Ecology and Evolution*, 11(13):8516–8527.

Wright, W. J., Irvine, K. M., Warren, J. M., and Barnett, J. K. (2017). Statistical design and analysis for plant cover studies with multiple sources of observation errors. *Methods in Ecology and Evolution*, 8(12):1832–1841.

Wright, W. J., Neitlich, P. N., Shiel, A. E., and Hooten, M. B. (2022). Mechanistic spatial models for heavy metal pollution. *Environmetrics*, 33(8):e2760.

Yoo, M. and Wikle, C. K. (2024). A Bayesian spatio-temporal level set dynamic model and application to fire front propagation. *The Annals of Applied Statistics*, 18(1):404–423.

Zhang, L. and Banerjee, S. (2022). Spatial factor modeling: A Bayesian matrix-normal approach for misaligned data. *Biometrics*, 78(2):560–573.

Zhu, Z. and Stein, M. L. (2006). Spatial sampling design for prediction with estimated parameters. *Journal of Agricultural, Biological, and Environmental Statistics*, 11:24–44.

Zimmerman, D. L. (2006). Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics*, 17(6):635–652.

Appendix A

Supplemental Information for Chapter 2

A.1 Prior and posterior distributions

We assumed the following prior distributions:

$$\begin{aligned}
\mu_{\beta_0} &\sim \text{Normal}(5, 5^2), \\
\mu_{\beta_1} &\sim \text{Normal}(0, 3^2), \\
\mu_{\gamma} &\sim \text{Normal}(0, 5^2), \\
\mu_{\theta_1} &\sim \text{Normal}(0, 3^2), \\
\mu_{\theta_2} &\sim \text{Normal}(0, 5^2), \\
\sigma^2_{\beta_0} &\sim \text{Inverse Gamma}(2, 1), \\
\sigma^2_{\beta_1} &\sim \text{Inverse Gamma}(2, 1), \\
\sigma^2_{\gamma} &\sim \text{Inverse Gamma}(2, 1), \\
\sigma^2_{\theta_1} &\sim \text{Inverse Gamma}(2, 1), \\
\sigma^2_{\theta_2} &\sim \text{Inverse Gamma}(2, 1), \\
\theta_{0j} &\sim \text{Normal}(5, 5^2), \\
\log(\alpha_j) &\sim \text{Normal}(4, 3^2), \\
\sigma^2_j &\sim \text{Inverse Gamma}(2, 1).
\end{aligned}$$

The resulting posterior distribution is

$$
\begin{aligned}
&[\boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\gamma}, \boldsymbol{\sigma^2}, \mu_{\beta_0}, \mu_{\beta_1}, \mu_\gamma, \mu_{\theta_1}, \mu_{\theta_2}, \sigma_{\beta_0}, \sigma_{\beta_1}, \sigma_\gamma, \sigma_{\theta_1}, \sigma_{\theta_2} \mid \mathbf{y}] \propto \\
&\prod_{j=1}^{3} \prod_{i=1}^{n} [y_{ij} \mid \alpha_j, \beta_{0j}, \beta_{1j}, \gamma_j, \theta_{0j}, \theta_{1j}, \theta_{2j}, \sigma_j^2] \times \\
&\prod_{j=1}^{3} [\beta_{0j} \mid \mu_{\beta_0}, \sigma_{\beta_0}^2][\beta_{1j} \mid \mu_{\beta_1}, \sigma_{\beta_1}^2] \times \\
&\prod_{j=1}^{3} [\theta_{1j} \mid \mu_{\theta_1}, \sigma_{\theta_1}^2][\theta_{2j} \mid \mu_{\theta_2}, \sigma_{\theta_2}^2] \times \\
&\prod_{j=1}^{3} [\log(\gamma_j) \mid \mu_\gamma, \sigma_\gamma^2][\log(\alpha_j)][\theta_{0j}][\sigma_j^2] \times \\
&[\mu_{\beta_0}][\sigma_{\beta_0}][\mu_{\beta_1}][\sigma_{\beta_1}][\mu_{\theta_1}][\sigma_{\theta_1}][\mu_{\theta_2}][\sigma_{\theta_2}][\mu_\gamma][\sigma_\gamma].
\end{aligned}
$$

A.2 Computational details

First, consider the partial differential equation with only Fickian diffusion:

$$
\frac{\partial \lambda(\mathbf{s}, t)}{\partial t} = \nabla \cdot (\delta(\mathbf{s}) \nabla \lambda(\mathbf{s}, t)) \,,
$$

where $\delta(\mathbf{s})$ is the spatially varying diffusion coefficent, $\nabla \cdot$ denotes the divergence operator, and ∇ denotes the gradient operator. In this case we are considering two spatial dimensions so that $\mathbf{s} \equiv (s_1, s_2)'$.

Observe that this PDE can be rewritten as

$$\begin{aligned}
\frac{\partial \lambda(\mathbf{s},t)}{\partial t} &= \nabla \cdot (\delta(\mathbf{s}) \nabla \lambda(\mathbf{s},t)) \\
&= \nabla \cdot \left(\delta(\mathbf{s}) \begin{pmatrix} \frac{\partial \lambda(\mathbf{s},t)}{\partial s_1} \\ \frac{\partial \lambda(\mathbf{s},t)}{\partial s_2} \end{pmatrix} \right) \\
&= \frac{\partial}{\partial s_1} \left(\delta(\mathbf{s}) \frac{\partial \lambda(\mathbf{s},t)}{\partial s_1} \right) + \frac{\partial}{\partial s_2} \left(\delta(\mathbf{s}) \frac{\partial \lambda(\mathbf{s},t)}{\partial s_2} \right) \\
&= \delta(\mathbf{s}) \left(\left(\frac{\partial^2}{\partial s_1^2} + \frac{\partial^2}{\partial s_2^2} \right) \lambda(\mathbf{s},t) \right) + \left(\frac{\partial \delta(\mathbf{s})}{\partial s_1} \right) \left(\frac{\partial \lambda(\mathbf{s},t)}{\partial s_1} \right) + \left(\frac{\partial \delta(\mathbf{s})}{\partial s_2} \right) \left(\frac{\partial \lambda(\mathbf{s},t)}{\partial s_2} \right).
\end{aligned}$$

Then use the forward difference operator to approximate all derivatives with respect to time (left-hand side) and the centered difference operators to approximate the derivatives with respect to space (right-hand side). This results in the following approximations:

$$\begin{aligned}
\frac{\partial \lambda(\mathbf{s},t)}{\partial t} &\approx \frac{\lambda(\mathbf{s},t+\Delta t) - \lambda(\mathbf{s},t)}{\Delta t}, \\
\frac{\partial^2 \lambda(\mathbf{s},t)}{\partial s_1^2} &\approx \frac{\lambda(s_1+\Delta s_1, s_2, t) - 2\lambda(\mathbf{s},t) + \lambda(s_1-\Delta s_1, s_2, t)}{(\Delta s_1)^2}, \\
\frac{\partial^2 \lambda(\mathbf{s},t)}{\partial s_2^2} &\approx \frac{\lambda(s_1, s_2+\Delta s_2, t) - 2\lambda(\mathbf{s},t) + \lambda(s_1, s_2-\Delta s_2, t)}{(\Delta s_2)^2}, \\
\frac{\partial \lambda(\mathbf{s},t)}{\partial s_1} &\approx \frac{\lambda(s_1+\Delta s_1, s_2, t) - \lambda(s_1-\Delta s_1, s_2, t)}{2\Delta s_1}, \\
\frac{\partial \lambda(\mathbf{s},t)}{\partial s_2} &\approx \frac{\lambda(s_1, s_2+\Delta s_2, t) - \lambda(s_1, s_2-\Delta s_2, t)}{2\Delta s_2}, \\
\frac{\partial \delta(\mathbf{s})}{\partial s_1} &\approx \frac{\delta(s_1+\Delta s_1, s_2) - \delta(s_1-\Delta s_1, s_2)}{2\Delta s_1}, \\
\frac{\partial \delta(\mathbf{s})}{\partial s_2} &\approx \frac{\delta(s_1, s_2+\Delta s_2) - \delta(s_1, s_2-\Delta s_2)}{2\Delta s_2}.
\end{aligned}$$

It then follows that we can approximate the PDE of interest as

$$\frac{\lambda(\mathbf{s}, t+\Delta t) - \lambda(\mathbf{s}, t)}{\Delta t} = \delta(\mathbf{s}) \left(\frac{\lambda(s_1 + \Delta s_1, s_2, t) - 2\lambda(\mathbf{s}, t) + \lambda(s_1 - \Delta s_1, s_2, t)}{(\Delta s_1)^2} \right) + \delta(\mathbf{s}) \left(\frac{\lambda(s_1, s_2 + \Delta s_2, t) - 2\lambda(\mathbf{s}, t) + \lambda(s_1, s_2 - \Delta s_2, t)}{(\Delta s_2)^2} \right) + \left(\frac{\delta(s_1 + \Delta s_1, s_2) - \delta(s_1 - \Delta s_1, s_2)}{2\Delta s_1} \right) \left(\frac{\lambda(s_1 + \Delta s_1, s_2, t) - \lambda(s_1 - \Delta s_1, s_2, t)}{2\Delta s_1} \right) + \left(\frac{\delta(s_1, s_2 + \Delta s_2) - \delta(s_1, s_2 - \Delta s_2)}{2\Delta s_2} \right) \left(\frac{\lambda(s_1, s_2 + \Delta s_2, t) - \lambda(s_1, s_2 - \Delta s_2, t)}{2\Delta s_2} \right).$$

Finally, solving for $\lambda(\mathbf{s}, t+\Delta t)$ and grouping the remaining λ terms results in using

$$\begin{aligned}
\lambda(\mathbf{s}, t+\Delta t) = & \left(1 - 2\delta(\mathbf{s})\Delta t \left(\frac{1}{(\Delta s_1)^2} + \frac{1}{(\Delta s_2)^2} \right) \right) \lambda(\mathbf{s}, t) + \\
& \left(\frac{\Delta t}{(\Delta s_1)^2} \left(\delta(\mathbf{s}) + \frac{\delta(s_1 + \Delta s_1, s_2) - \delta(s_1 - \Delta s_1, s_2)}{4} \right) \right) \lambda(s_1 + \Delta s_1, s_2, t) + \\
& \left(\frac{\Delta t}{(\Delta s_1)^2} \left(\delta(\mathbf{s}) - \frac{\delta(s_1 + \Delta s_1, s_2) - \delta(s_1 - \Delta s_1, s_2)}{4} \right) \right) \lambda(s_1 - \Delta s_1, s_2, t) + \\
& \left(\frac{\Delta t}{(\Delta s_2)^2} \left(\delta(\mathbf{s}) + \frac{\delta(s_1, s_2 + \Delta s_2) - \delta(s_1, s_2 - \Delta s_2)}{4} \right) \right) \lambda(s_1, s_2 + \Delta s_2, t) + \\
& \left(\frac{\Delta t}{(\Delta s_2)^2} \left(\delta(\mathbf{s}) - \frac{\delta(s_1, s_2 + \Delta s_2) - \delta(s_1, s_2 - \Delta s_2)}{4} \right) \right) \lambda(s_1, s_2 - \Delta s_2, t)
\end{aligned}$$

to solve the Fickian diffusion PDE. As described in the main text, by letting $\boldsymbol{\lambda}(t)$ denote a vector of dimension N that contains $\lambda(\mathbf{s}, t)$ values at time t for the spatial locations used to approximate the PDE. A solution to the PDE for times $t = 1, \ldots, T$ can then be found as $\boldsymbol{\lambda}(t+1) = \mathbf{H}\boldsymbol{\lambda}(t)$ where $\mathbf{H}$ is a "propagator" matrix which is defined using finite difference approximation described above. Note that $\mathbf{H}$ is sparse and utilizing sparse matrix operations make solving this PDE more efficient.

To incorporate advection into this PDE, we used "upwind" difference operators to approximate the spatial derivatives because they are more numerically stable than the centered difference operator in this case. For advection component

$$\gamma u(t) \frac{\partial \lambda(\mathbf{s}, t)}{\partial s_1} + \gamma v(t) \frac{\partial \lambda(\mathbf{s}, t)}{\partial s_2},$$

we approximated $\partial\lambda(\mathbf{s},t)/\partial s_1$ using the backward difference operator if $u(t) > 0$ and the forward difference operator if $u(t) < 0$. Similarly, the backwards or forwards difference operator was used to approximate $\partial\lambda(\mathbf{s},t)/\partial s_2$ depending on whether $v(t)$ was greater than zero or not. For instance, if $u(t), v(t) > 0$ and we only consider the advection component, the PDE is approximated by

$$\begin{aligned}\lambda(\mathbf{s}, t+\Delta t) = &\left(1 - \gamma u(t)\frac{\Delta t}{\Delta s_1} - \gamma v(t)\frac{\Delta t}{\Delta s_2}\right)\lambda(\mathbf{s},t) + \\ &\left(\gamma u(t)\frac{\Delta t}{\Delta s_1}\right)\lambda(s_1 - \Delta s_1, s_2, t) + \\ &\left(\gamma v(t)\frac{\Delta t}{\Delta s_2}\right)\lambda(s_1, s_2 - \Delta s_2, t).\end{aligned}$$

This approximation can be used to modify the propagator matrix, which now varies with time because it includes $u(t)$ and $v(t)$. The specification of propagator matrix $\mathbf{H}(t)$ that was used to solve the PDE when fitting our model is completed after including the deposition component as well.

Appendix B

Supplemental Information for Chapter 3

B.1 MCMC details

We derive the full-conditional distributions needed in the MCMC algorithm. These results are for the hierarchical formulation of the multi-species model.

B.1.1 Updating the latent process

For plot i and species j, the vector of binary observations $\mathbf{y}_{ij}$ is related to the latent process $\tilde{\mathbf{y}}_{ij}$ which needs to be updated at each iteration of the MCMC algorithm. Our model assumes

$$\tilde{\mathbf{y}}_{ij} \sim \mathrm{N}(\mu_{ij}\mathbf{1}_D + \mathbf{\Lambda}_i\boldsymbol{\theta}_j, \sigma_{j2}^2\tilde{\mathbf{R}}_{j2}),$$

and that

$$y_{ijd} = \mathbb{1}(\tilde{y}_{ijd} \geq 0),$$

for grid points $d = 1, \ldots, D$. It follows that the full-conditional distribution of $\tilde{\mathbf{y}}_{ij}$ is a truncated multivariate normal distribution where the truncation bounds are defined by y_{ijd} so that $\tilde{y}_{ijd} \in (-\infty, 0]$ if $y_{ijd} = 0$ and $\tilde{y}_{ijd} \in (0, \infty)$ if $y_{ijd} = 1$. We used a Gibbs sampler to update each component of $\tilde{\mathbf{y}}_{ij}$ as follows. Let $\mathbf{H} \equiv \sigma_{j2}^{-2}\tilde{\mathbf{R}}_{j2}^{-1}$, where we have suppressed the species subscript j and denote partitions of this matrix using subscripts (e.g., $\mathbf{H}_{dd}$ or $\mathbf{H}_{d,-d}$). Conditional on all other values of the latent process, denoted $\tilde{\mathbf{y}}_{ij,-d}$, the mean of $\tilde{y}_{ijd}$ is

$$\mu_{ijd|-d} = \mu_{ij} + \mathbf{\Lambda}_{id}\boldsymbol{\theta}_j - (\mathbf{H}_{dd})^{-1}\mathbf{H}_{d,-d}(\tilde{\mathbf{y}}_{ij,-d} - \mu_{ij}\mathbf{1}_{D-1} - \mathbf{\Lambda}_{i,-d}\boldsymbol{\theta}_j)$$

and the variance is $(\mathbf{H}_{dd})^{-1}$. These results follow from basic properties of the multivariate normal distribution. Then the full-conditional distribution for each $\tilde{y}_{ijd}$ is

$$\tilde{y}_{ijd} \mid \cdot \sim \text{TN}\left(\mu_{ijd|-d}, (\mathbf{H}_{dd})^{-1}, a_{ijd}, b_{ijd}\right),$$

where $TN(\mu, \sigma^2, a, b)$ denotes a normal distribution with mean μ and variance σ^2 that has been truncated to the interval (a, b). The truncation boundaries (a_{ijd}, b_{ijd}) are defined by y_{ijd} so that $\tilde{y}_{ijd} \in (-\infty, 0]$ if $y_{ijd} = 0$ and $\tilde{y}_{ijd} \in (0, \infty)$ if $y_{ijd} = 1$.

B.1.2 Updating the plot means

We consider the full-conditional distribution of the vector of plot means $\boldsymbol{\mu}$ for species j where we have suppressed the species subscript in this section for clarity. Note that this update is performed for each species. Let $\tilde{\mathbf{y}}^*$ denote a vector of the latent process $\tilde{\mathbf{y}}_i$ across all plots $i = 1, \ldots, n$ after subtracting the effects of the latent factors. That is,

$$\tilde{\mathbf{y}}^* \equiv \begin{pmatrix} \tilde{\mathbf{y}}_1 - \boldsymbol{\Lambda}_1\boldsymbol{\theta} \\ \vdots \\ \tilde{\mathbf{y}}_n - \boldsymbol{\Lambda}_n\boldsymbol{\theta} \end{pmatrix}.$$

Then our model results in

$$\tilde{\mathbf{y}}^* \sim \text{N}(\mathbf{Z}\boldsymbol{\mu}, \mathbf{I}_n \otimes \sigma_2^2\tilde{\mathbf{R}}_2),$$
$$\boldsymbol{\mu} \sim \text{N}(\tilde{\mathbf{X}}\boldsymbol{\beta}, \sigma_1^2\tilde{\mathbf{R}}_1),$$

where $\mathbf{Z}$ is a $nD \times n$ matrix of zeroes and ones that selects the appropriate element of $\boldsymbol{\mu}$ for each element of $\tilde{\mathbf{y}}^*$. Specifically, we can define $\mathbf{Z}$ to be

$$\mathbf{Z} = \begin{pmatrix} \mathbf{1}_D & \mathbf{0}_D & \mathbf{0}_D & \cdots & \mathbf{0}_D \\ \mathbf{0}_D & \mathbf{1}_D & \mathbf{0}_D & \cdots & \mathbf{0}_D \\ \vdots & & \ddots & & \vdots \\ \mathbf{0}_D & \cdots & \mathbf{0}_D & \mathbf{1}_D & \mathbf{0}_D \\ \mathbf{0}_D & \cdots & \mathbf{0}_D & \mathbf{0}_D & \mathbf{1}_D \end{pmatrix},$$

where $\mathbf{1}_D$ and $\mathbf{0}_D$ denote D-dimensional vectors of ones and zeroes, respectively. Then the full-conditional distribution for $\boldsymbol{\mu}$ is

$$\begin{aligned} p(\boldsymbol{\mu} \mid \cdot) &\propto p(\tilde{\mathbf{y}}^* \mid \boldsymbol{\mu}, \sigma_2^2 \tilde{\mathbf{R}}_2) p(\boldsymbol{\mu}) \\ &\propto \exp\left\{ -\frac{1}{2} \left(\tilde{\mathbf{y}}^* - \mathbf{Z}\boldsymbol{\mu}\right)' \left(\mathbf{I}_n \otimes \sigma_2^2 \tilde{\mathbf{R}}_2\right)^{-1} \left(\tilde{\mathbf{y}}^* - \mathbf{Z}\boldsymbol{\mu}\right) \right\} \times \\ &\quad \exp\left\{ -\frac{1}{2} \left(\boldsymbol{\mu} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right)' (\sigma_1^2 \tilde{\mathbf{R}}_1)^{-1} \left(\boldsymbol{\mu} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right) \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \left[\boldsymbol{\mu}' \mathbf{Z}' \left(\mathbf{I}_n \otimes \sigma_2^2 \tilde{\mathbf{R}}_2\right)^{-1} \mathbf{Z}\boldsymbol{\mu} - 2\boldsymbol{\mu}' \mathbf{Z}' \left(\mathbf{I}_n \otimes \sigma_2^2 \tilde{\mathbf{R}}_2\right)^{-1} \tilde{\mathbf{y}}^* \right] \right\} \times \\ &\quad \exp\left\{ -\frac{1}{2} \left[\boldsymbol{\mu}' (\sigma_1^2 \tilde{\mathbf{R}}_1)^{-1} \boldsymbol{\mu} - 2\boldsymbol{\mu}' (\sigma_1^2 \tilde{\mathbf{R}}_1)^{-1} \tilde{\mathbf{X}}\boldsymbol{\beta} \right] \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \left[\boldsymbol{\mu}' \left((\sigma_1^2 \tilde{\mathbf{R}}_1)^{-1} + \mathbf{Z}' \left(\mathbf{I}_n \otimes \sigma_2^2 \tilde{\mathbf{R}}_2\right)^{-1} \mathbf{Z} \right) \boldsymbol{\mu} \right] \right\} \times \\ &\quad \exp\left\{ \boldsymbol{\mu}' \left((\sigma_1^2 \tilde{\mathbf{R}}_1)^{-1} \tilde{\mathbf{X}}\boldsymbol{\beta} + \mathbf{Z}' \left(\mathbf{I}_n \otimes \sigma_2^2 \tilde{\mathbf{R}}_2\right)^{-1} \tilde{\mathbf{y}}^* \right) \right\}, \end{aligned}$$

which is the kernel of a multivariate normal distribution with mean $\mathbf{A}^{-1}\mathbf{b}$ and covariance $\mathbf{A}^{-1}$ where

$$\begin{aligned} \mathbf{A} &= (\sigma_1^2 \tilde{\mathbf{R}}_1)^{-1} + \mathbf{Z}' \left(\mathbf{I}_n \otimes \sigma_2^2 \tilde{\mathbf{R}}_2\right)^{-1} \mathbf{Z}, \\ \mathbf{b} &= (\sigma_1^2 \tilde{\mathbf{R}}_1)^{-1} \tilde{\mathbf{X}}\boldsymbol{\beta} + \mathbf{Z}' \left(\mathbf{I}_n \otimes \tilde{\mathbf{R}}_2\right)^{-1} \tilde{\mathbf{y}}^*. \end{aligned}$$

Calculating the required mean and variance to update $\boldsymbol{\mu}$ can be simplified. First, note that $\mathbf{Z} = \mathbf{I}_n \otimes \mathbf{1}_D$. Then we can see that

$$\begin{aligned}
\mathbf{Z}'(\mathbf{I}_n \otimes \sigma_2^2 \tilde{\mathbf{R}}_2)^{-1}\mathbf{Z} &= \mathbf{Z}' \left(\mathbf{I}_n \otimes \sigma_2^{-2}\tilde{\mathbf{R}}_2^{-1}\right)\mathbf{Z} \\
&= (\mathbf{I}_n \otimes \mathbf{1}_D') \left(\mathbf{I}_n \otimes \sigma_2^{-2}\tilde{\mathbf{R}}_2^{-1}\right)(\mathbf{I}_n \otimes \mathbf{1}_D) \\
&= \mathbf{I}_n \otimes \sigma_2^{-2}\mathbf{1}_D'\tilde{\mathbf{R}}_2^{-1}\mathbf{1}_D,
\end{aligned}$$

which is a diagonal matrix with each element equal to the (scalar) $\sigma_2^{-2}\mathbf{1}_D'\tilde{\mathbf{R}}_2^{-1}\mathbf{1}_D$. This result simplifies the calculation of $\mathbf{A}$ above. The calculation of $\mathbf{b}$ can also be simplified as

$$\begin{aligned}
\mathbf{Z}' \left(\mathbf{I}_n \otimes \tilde{\mathbf{R}}_2\right)^{-1} \tilde{\mathbf{y}}^* &= \left(\mathbf{I}_n \otimes \sigma_2^{-2}\mathbf{1}_D'\tilde{\mathbf{R}}_2^{-1}\right)\tilde{\mathbf{y}}^* \\
&= \begin{pmatrix} \sigma_2^{-2}\mathbf{1}_D'\tilde{\mathbf{R}}_2^{-1}(\tilde{\mathbf{y}}_1 - \boldsymbol{\Lambda}_1\boldsymbol{\theta}) \\ \vdots \\ \sigma_2^{-2}\mathbf{1}_D'\tilde{\mathbf{R}}_2^{-1}(\tilde{\mathbf{y}}_n - \boldsymbol{\Lambda}_n\boldsymbol{\theta}) \end{pmatrix}.
\end{aligned}$$

We can further simplify $\mathbf{b}$ if we rearrange the terms of $\tilde{\mathbf{y}}^*$ into a $n \times D$ matrix $\tilde{\mathbf{Y}}^*$ such that $\tilde{\mathbf{y}}^* = \text{vec}\left(\tilde{\mathbf{Y}}^{*\prime}\right)$ (so that row i of $\tilde{\mathbf{Y}}^*$ is equal to the row-vector $\tilde{\mathbf{y}}_i' - \boldsymbol{\theta}'\boldsymbol{\Lambda}_i'$). Then it follows from above that

$$\begin{aligned}
\left(\mathbf{I}_n \otimes \sigma_2^{-2}\mathbf{1}_D'\tilde{\mathbf{R}}_2^{-1}\right)\tilde{\mathbf{y}}^* &= \left(\mathbf{I}_n \otimes \sigma_2^{-2}\mathbf{1}_D'\tilde{\mathbf{R}}_2^{-1}\right)\text{vec}\left(\tilde{\mathbf{Y}}^{*\prime}\right) \\
&= \text{vec}\left(\sigma_2^{-2}\mathbf{1}_D'\tilde{\mathbf{R}}_2^{-1}\tilde{\mathbf{Y}}^{*\prime}\right) \\
&= \tilde{\mathbf{Y}}^*\sigma_2^{-2}\tilde{\mathbf{R}}_2^{-1}\mathbf{1}_D,
\end{aligned}$$

using the properties of the vectorization operator and Kronecker product. Rewriting the mean and variance in this way shows that we only need to invert the $D \times D$ matrix $\tilde{\mathbf{R}}_2$ once each iteration.

B.1.3 Updating the plot means and factor loadings

The plot means $\boldsymbol{\mu}$ and factor loadings $\boldsymbol{\theta}_p$ for species j can be updated simultaneously. Note that we have suppressed the species subscript in this section for clarity. Let $\tilde{\mathbf{y}}^*$ denote the vector of the latent process $\tilde{\mathbf{y}}_i$ across all plots $i = 1, \ldots, n$. That is, $\tilde{\mathbf{y}}^* = (\tilde{\mathbf{y}}_1' \ \tilde{\mathbf{y}}_2' \ \ldots \ \tilde{\mathbf{y}}_n')'$. Additionally, let $\boldsymbol{\theta}^* \equiv (\boldsymbol{\theta}_1' \ \ldots \boldsymbol{\theta}_{P_2}')'$ and

$$\boldsymbol{\Lambda}^* \equiv \begin{pmatrix} z_{1,1}\boldsymbol{\Lambda}_1 & \cdots & z_{1,P_2}\boldsymbol{\Lambda}_1 \\ \vdots & & \vdots \\ z_{n,1}\boldsymbol{\Lambda}_n & \cdots & z_{n,P_2}\boldsymbol{\Lambda}_n \end{pmatrix},$$

where $z_{i,p}$ is predictor variable p at plot i. Then our model assumes that

$$\begin{aligned} \tilde{\mathbf{y}}^* &\sim \mathrm{N}(\mathbf{H}\boldsymbol{\mu} + \boldsymbol{\Lambda}^*\boldsymbol{\theta}^*, \mathbf{I}_n \otimes \sigma_2^2\tilde{\mathbf{R}}_2), \\ \boldsymbol{\mu} &\sim \mathrm{N}(\tilde{\mathbf{X}}\boldsymbol{\beta}, \sigma_1^2\tilde{\mathbf{R}}_1), \\ \boldsymbol{\theta}^* &\sim \mathrm{N}(\mathbf{0}, \mathbf{I}_{MP_2}), \end{aligned}$$

where $\mathbf{H}$ is a $nD \times n$ matrix of zeroes and ones that selects the appropriate element of $\boldsymbol{\mu}$ for each element of $\tilde{\mathbf{y}}^*$. Specifically, we can define $\mathbf{H}$ to be

$$\mathbf{H} = \mathbf{I}_n \otimes \mathbf{1}_D = \begin{pmatrix} \mathbf{1}_D & \mathbf{0}_D & \mathbf{0}_D & \cdots & \mathbf{0}_D \\ \mathbf{0}_D & \mathbf{1}_D & \mathbf{0}_D & \cdots & \mathbf{0}_D \\ \vdots & & \ddots & & \vdots \\ \mathbf{0}_D & \cdots & \mathbf{0}_D & \mathbf{1}_D & \mathbf{0}_D \\ \mathbf{0}_D & \cdots & \mathbf{0}_D & \mathbf{0}_D & \mathbf{1}_D \end{pmatrix},$$

where $\mathbf{1}_D$ and $\mathbf{0}_D$ denote D-dimensional vectors of ones and zeroes, respectively. The likelihood for $\tilde{\mathbf{y}}^*$ can be rewritten as

$$\tilde{\mathbf{y}}^* \sim \mathrm{N}\left(\mathbf{H}^* \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\theta}^* \end{pmatrix}, \mathbf{I}_n \otimes \sigma_2^2 \tilde{\mathbf{R}}_2\right),$$

where $\mathbf{H}^* = [\mathbf{H}\ \boldsymbol{\Lambda}^*]$. Similarly, the joint prior distribution for $\boldsymbol{\mu}$ and $\boldsymbol{\theta}^*$ is

$$\begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\theta}^* \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \tilde{\mathbf{X}}\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma_1^2 \tilde{\mathbf{R}}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{MP_2} \end{pmatrix}\right) \stackrel{d}{=} \mathrm{N}(\boldsymbol{\nu}_{\mu,\theta}, \boldsymbol{\Sigma}_{\mu,\theta}).$$

Standard calculations show that the full-conditional distribution is

$$\begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\theta}^* \end{pmatrix} \mid \cdot \sim \mathrm{N}(\mathbf{A}^{-1}\mathbf{b}, \mathbf{A}^{-1}),$$

where $\mathbf{A} = \mathbf{H}^{*-1}\left(\mathbf{I}_n \otimes \sigma_2^{-2}\tilde{\mathbf{R}}_2^{-1}\right)\mathbf{H}^* + \boldsymbol{\Sigma}_{\mu,\theta}^{-1}$ and $\mathbf{b} = \mathbf{H}^{*-1}\left(\mathbf{I}_n \otimes \sigma_2^{-2}\tilde{\mathbf{R}}_2^{-1}\right)\tilde{\mathbf{y}}^* + \boldsymbol{\Sigma}_{\mu,\theta}^{-1}\boldsymbol{\nu}_{\mu,\theta}$.

B.1.4 Updating the regression coefficients and hyperparameters

For each species j, we assumed that

$$\begin{aligned} \boldsymbol{\beta}_j &\sim \mathrm{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta), \\ \boldsymbol{\mu}_j &\sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}_j, \sigma_{j1}^2 \tilde{\mathbf{R}}_{j1}). \end{aligned}$$

Standard posterior calculations show that we can update the regression coefficients for each species using the conditionally conjugate update

$$\boldsymbol{\beta}_j \sim \mathrm{N}(\mathbf{A}^{-1}\mathbf{b}, \mathbf{A}^{-1}),$$

where

$$\mathbf{A} = \mathbf{X}'\sigma_{j1}^{-2}\tilde{\mathbf{R}}_{j1}^{-1}\mathbf{X} + \boldsymbol{\Sigma}_{\beta}^{-1},$$
$$\mathbf{b} = \mathbf{X}'\sigma_{j1}^{-2}\tilde{\mathbf{R}}_{j1}^{-1}\boldsymbol{\mu}_j + \boldsymbol{\Sigma}_{\beta}^{-1}\boldsymbol{\mu}_{\beta}.$$

Next, we consider updating the hyperparameters associated with these regression coefficients, $\boldsymbol{\mu}_{\beta}$ and $\boldsymbol{\Sigma}_{\beta}$. For $p = 1, \ldots, P$, we assigned exchangeable prior distributions as

$$\mu_{\beta,p} \overset{iid}{\sim} \text{N}(0, \tau_{\beta}^2),$$
$$\sigma_{\beta,p}^2 \overset{iid}{\sim} \text{IG}(a, b),$$

where $\boldsymbol{\Sigma}_{\beta} \equiv \text{Diag}(\sigma_{\beta,1}^2, \ldots, \sigma_{\beta,P}^2)$. These hyperparameters can then be updated using standard conditionally conjugate updates.

B.1.5 Updating the latent factors

The matrix of latent factors for plot i, denoted $\boldsymbol{\Lambda}_i$, has a conditionally conjugate update within our MCMC algorithm. Note that $\boldsymbol{\Lambda}_i$ is a $D \times M$ matrix and that

$$\text{vec}(\boldsymbol{\Lambda}_i) \sim \text{N}(\mathbf{0}, \mathbf{I}_m \otimes \tilde{\boldsymbol{\Sigma}}_{\lambda}),$$

where $\tilde{\boldsymbol{\Sigma}}_{\lambda}$ is a block diagonal matrix with blocks $\psi_m \boldsymbol{\Sigma}_{\lambda_m}$ for $m = 1, \ldots, M$. Each $\boldsymbol{\Sigma}_m$ is defined by the spatial covariance function $K_{\rho_{\lambda_m}}$ using the grid point distances $\mathbf{v}_1, \ldots, \mathbf{v}_D$. The ψ_m are working parameters that are assigned prior distributions $\psi_m \sim \text{InverseGamma}(a_{\psi}, b_{\psi})$. This parameter expansion induces the t prior distributions for each θ_{jpm}. Let $\tilde{\mathbf{y}}_i^*$ denote the vector of latent variables

at plot i for all species after they have been centered by their corresponding means. That is,

$$\tilde{\mathbf{y}}_i^* = \begin{pmatrix} \tilde{\mathbf{y}}_{i1} - \mu_{i1}\mathbf{1}_D \\ \tilde{\mathbf{y}}_{i2} - \mu_{i2}\mathbf{1}_D \\ \vdots \\ \tilde{\mathbf{y}}_{iJ} - \mu_{iJ}\mathbf{1}_D \end{pmatrix},$$

which is distributed

$$\tilde{\mathbf{y}}_i^* \sim \mathrm{N}(\mathrm{vec}(\boldsymbol{\Lambda}_i\boldsymbol{\Theta}), \mathbf{R}^*),$$

where $\mathbf{R}^*$ is a block diagonal matrix created using the elements $\tilde{\mathbf{R}}_{j2}$ for $j = 1, \ldots, J$. Note that we can rewrite the mean of $\tilde{\mathbf{y}}_i^*$ as

$$\mathrm{vec}(\boldsymbol{\Lambda}_i\boldsymbol{\Theta}) = (\boldsymbol{\Theta}' \otimes \mathbf{I}_D)\mathrm{vec}(\boldsymbol{\Lambda}_i)$$

using known properties of the Kronecker product. Then $\mathrm{vec}(\boldsymbol{\Lambda}_i)$ can be updated using the full-conditional distribution

$$\begin{aligned} p(\boldsymbol{\Lambda}_i \mid \cdot) &\propto \exp\left\{-\frac{1}{2}\left(\tilde{\mathbf{y}}_i^* - (\boldsymbol{\Theta}' \otimes \mathbf{I}_D)\mathrm{vec}(\boldsymbol{\Lambda}_i)\right)' \mathbf{R}^{*-1}\left(\tilde{\mathbf{y}}_i^* - (\boldsymbol{\Theta}' \otimes \mathbf{I}_D)\mathrm{vec}(\boldsymbol{\Lambda}_i)\right)\right\} \times \\ &\qquad \exp\left\{-\frac{1}{2}\mathrm{vec}(\boldsymbol{\Lambda}_i)'\tilde{\boldsymbol{\Sigma}}_\lambda^{-1}\mathrm{vec}(\boldsymbol{\Lambda}_i)\right\} \\ &\propto \exp\left\{-\frac{1}{2}\mathrm{vec}(\boldsymbol{\Lambda}_i)'\left((\boldsymbol{\Theta}' \otimes \mathbf{I}_D)'\mathbf{R}^{*-1}(\boldsymbol{\Theta}' \otimes \mathbf{I}_D) + \tilde{\boldsymbol{\Sigma}}_\lambda^{-1}\right)\mathrm{vec}(\boldsymbol{\Lambda}_i)\right\} \times \\ &\qquad \exp\left\{\mathrm{vec}(\boldsymbol{\Lambda}_i)'(\boldsymbol{\Theta}' \otimes \mathbf{I}_D)'\mathbf{R}^{*-1}\tilde{\mathbf{y}}_i^*\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(\mathrm{vec}(\boldsymbol{\Lambda}_i) - \mathbf{A}^{-1}\mathbf{b}\right)'\mathbf{A}\left(\mathrm{vec}(\boldsymbol{\Lambda}_i) - \mathbf{A}^{-1}\mathbf{b}\right)\right\}, \end{aligned}$$

where

$$\begin{aligned} \mathbf{A} &= (\boldsymbol{\Theta}' \otimes \mathbf{I}_D)'\mathbf{R}^{*-1}(\boldsymbol{\Theta}' \otimes \mathbf{I}_D) + \tilde{\boldsymbol{\Sigma}}_\lambda^{-1}, \\ \mathbf{b} &= (\boldsymbol{\Theta}' \otimes \mathbf{I}_D)'\mathbf{R}^{*-1}\tilde{\mathbf{y}}_i^*. \end{aligned}$$

This results in

$$\boldsymbol{\Lambda}_i \mid \cdot \sim \mathrm{N}(\mathbf{A}^{-1}\mathbf{b}, \mathbf{A}^{-1}),$$

for $\mathbf{A}$ and $\mathbf{b}$ defined above.

The matrix calculations required for this update can be simplified using the following results. First, note that

$$\boldsymbol{\Theta}' \otimes \mathbf{I}_D = \begin{pmatrix} \boldsymbol{\theta}'_1 \\ \boldsymbol{\theta}'_2 \\ \vdots \\ \boldsymbol{\theta}'_J \end{pmatrix} \otimes \mathbf{I}_D = \begin{pmatrix} \boldsymbol{\theta}'_1 \otimes \mathbf{I}_D \\ \boldsymbol{\theta}'_2 \otimes \mathbf{I}_D \\ \vdots \\ \boldsymbol{\theta}'_J \otimes \mathbf{I}_D \end{pmatrix},$$

and that

$$(\boldsymbol{\Theta}' \otimes \mathbf{I}_D)' = \begin{pmatrix} \boldsymbol{\theta}_1 \otimes \mathbf{I}_D & \boldsymbol{\theta}_2 \otimes \mathbf{I}_D & \cdots & \boldsymbol{\theta}_J \otimes \mathbf{I}_D \end{pmatrix}.$$

Then we can rewrite the first part of $\mathbf{A}$ as

$$\begin{aligned}
(\boldsymbol{\Theta}' \otimes \mathbf{I}_D)'\mathbf{R}^{*-1}(\boldsymbol{\Theta}' \otimes \mathbf{I}_D) &= \begin{pmatrix} (\boldsymbol{\theta}_1 \otimes \mathbf{I}_D)\sigma_{12}^{-2}\tilde{\mathbf{R}}_{12}^{-1} & \cdots & (\boldsymbol{\theta}_J \otimes \mathbf{I}_D)\sigma_{J2}^{-2}\tilde{\mathbf{R}}_{J2}^{-1} \end{pmatrix} (\boldsymbol{\Theta}' \otimes \mathbf{I}_D) \\
&= (\boldsymbol{\theta}_1 \otimes \mathbf{I}_D)\sigma_{12}^{-2}\tilde{\mathbf{R}}_{12}^{-1}(\boldsymbol{\theta}'_1 \otimes \mathbf{I}_D) + \ldots + \\
&\quad (\boldsymbol{\theta}_J \otimes \mathbf{I}_D)\sigma_{J2}^{-2}\tilde{\mathbf{R}}_{J2}^{-1}(\boldsymbol{\theta}'_J \otimes \mathbf{I}_D) \\
&= \boldsymbol{\theta}_1\boldsymbol{\theta}'_1 \otimes \sigma_{12}^{-2}\tilde{\mathbf{R}}_{12}^{-1} + \ldots + \boldsymbol{\theta}_J\boldsymbol{\theta}'_J \otimes \sigma_{J2}^{-2}\tilde{\mathbf{R}}_{J2}^{-1},
\end{aligned}$$

where the last line follows from

$$\begin{aligned}
(\boldsymbol{\theta}_j \otimes \mathbf{I}_D)\sigma_{j2}^{-2}\tilde{\mathbf{R}}_{j2}^{-1}(\boldsymbol{\theta}'_j \otimes \mathbf{I}_D) &= (\boldsymbol{\theta}_j \otimes \mathbf{I}_D)(1 \otimes \sigma_{j2}^{-2}\tilde{\mathbf{R}}_{j2}^{-1})(\boldsymbol{\theta}'_j \otimes \mathbf{I}_D) \\
&= (\boldsymbol{\theta}_j \otimes \sigma_{j2}^{-2}\tilde{\mathbf{R}}_{j2}^{-1})(\boldsymbol{\theta}'_j \otimes \mathbf{I}_D) \\
&= \boldsymbol{\theta}_j\boldsymbol{\theta}'_j \otimes \sigma_{j2}^{-2}\tilde{\mathbf{R}}_{j2}^{-1}
\end{aligned}$$

for all j. Using this result, $\mathbf{A}$ is equivalent to

$$\mathbf{A} = \boldsymbol{\theta}_1\boldsymbol{\theta}_1' \otimes \sigma_{12}^{-2}\tilde{\mathbf{R}}_{12}^{-1} + \ldots + \boldsymbol{\theta}_J\boldsymbol{\theta}_J' \otimes \sigma_{J2}^{-2}\tilde{\mathbf{R}}_{J2}^{-1} + \tilde{\boldsymbol{\Sigma}}_{\lambda}^{-1}.$$

Similarly, we can rewrite $\mathbf{b}$ as

$$\begin{aligned}
\mathbf{b} = (\boldsymbol{\Theta}' \otimes \mathbf{I}_D)'\mathbf{R}^{*-1}\tilde{\mathbf{y}}_i^* &= \left((\boldsymbol{\theta}_1 \otimes \mathbf{I}_D)\sigma_{12}^{-2}\tilde{\mathbf{R}}_{12}^{-1} \quad \cdots \quad (\boldsymbol{\theta}_J \otimes \mathbf{I}_D)\sigma_{J2}^{-2}\tilde{\mathbf{R}}_{J2}^{-1}\right)\tilde{\mathbf{y}}_i^* \\
&= (\boldsymbol{\theta}_1 \otimes \mathbf{I}_D)\sigma_{12}^{-2}\tilde{\mathbf{R}}_{12}^{-1}(\tilde{\mathbf{y}}_{i1} - \mu_{i1}\mathbf{1}_D) + \ldots + \\
&\quad (\boldsymbol{\theta}_J \otimes \mathbf{I}_D)\sigma_{J2}^{-2}\tilde{\mathbf{R}}_{J2}^{-1}(\tilde{\mathbf{y}}_{iJ} - \mu_{iJ}\mathbf{1}_D) \\
&= \boldsymbol{\theta}_1 \otimes \sigma_{12}^{-2}\tilde{\mathbf{R}}_{12}^{-1}(\tilde{\mathbf{y}}_{i1} - \mu_{i1}\mathbf{1}_D) + \ldots + \boldsymbol{\theta}_J \otimes \sigma_{J2}^{-2}\tilde{\mathbf{R}}_{J2}^{-1}(\tilde{\mathbf{y}}_{iJ} - \mu_{iJ}\mathbf{1}_D).
\end{aligned}$$

This shows how both $\mathbf{A}$ and $\mathbf{b}$ can be computed from the sum of J smaller matrices. This is computationally more efficient than performing the larger matrix operations in the original formulations above.

The working parameters ψ_m are each updated using conditionally conjugate steps. Standard calculations show that the full-conditional distribution is

$$\psi_m \mid \cdot \sim \text{InverseGamma}\left(a_\psi + nD/2, b_\psi + 0.5\sum_{i=1}^{n}\boldsymbol{\lambda}_{im}'\boldsymbol{\Sigma}_{\lambda_m}^{-1}\boldsymbol{\lambda}_{im}\right).$$

Note that in this parameter expanded version of the model, the species coefficients $\boldsymbol{\theta}_j$ and factor loadings $\boldsymbol{\lambda}_{im}$ are not identifiable. Inferences for the interspecies correlation structure are of interest and are defined by the species coefficients. Thus, during post processing, the posterior draws of the factor loadings are transformed so that $\tilde{\theta}_{jpm} = \psi_m^{\frac{1}{2}}\theta_{jpm}$ and then the additional orthogonality constraints are imposed. The transformed loadings $\tilde{\theta}_{jpm}$ are identifiable and can be used for inference and to assess MCMC convergence.

B.2 Discrete support for spatial range parameters

We used a simulated data example to demonstrate how discretizing the support for the spatial range parameters can improve the computational efficiency of our model. Data were generated from our multispecies model. To reduce the computation time, we did not include the latent factor structure. That is, we assumed $\boldsymbol{\theta}_j = \mathbf{0}$ for all j when generating data and fitting models. We modeled the regression coefficients, $\boldsymbol{\beta}_j$ and the spatial scale parameters, ρ_{1j} and ρ_{2j}, hierarchically across species. For the regression coefficients, we assumed

$$\beta_{pj} \overset{iid}{\sim} \text{Normal}(\mu_{\beta_p}, \sigma^2_{\beta_p})$$

for coefficients $p = 1, \ldots, P$. For each spatial range parameter, we created a grid of evenly spaced values defined by $\rho_{\text{grid}}[k]$ for index $k = 0, \ldots, K$. We have excluded the subscript for the different range parameters for clarity. As described in the main text, we let

$$\Pr(\rho_j = \rho_{\text{grid},k}) = \begin{cases} \binom{K}{k} \frac{B(\alpha_1+k, \alpha_2+K-k)}{B(\alpha_1,\alpha_2)}, & k = 0, \ldots, K, \\ 0, & \text{otherwise}, \end{cases}$$

and

$$\nu_\rho \sim \text{Uniform}(0, 1),$$
$$\phi_\rho \sim \text{Gamma}(\gamma_1, \gamma_2),$$

where $\nu_\rho = \alpha_1/(\alpha_1 + \alpha_2)$ and $\phi_\rho = \alpha_1 + \alpha_2$. Different grids and hyperparameters (γ_1, γ_2) can be chosen for both spatial scales as needed for the specific study design.

We simulated data for $J = 20$ species, $N = 100$ plots, and $D = 100$ grid points per plot. We assumed a single predictor variable x_i was available for each plot and generated this predictor variable from a Uniform(-1, 1) distribution. For each species, values for the regression coefficients were independently generated from Normal distributions (as described above) with means 0 and -1,

variances 0.5^2 and 0.25^2, for the intercept and slope parameters, respectively. The data-generating values for the spatial range parameters were random draws from log-Normal distributions. We fit one model which assumed a discrete support for the range parameters (as described above) and we fit a second model which assumed a continuous support for the range parameters (data-generating model).

We first assessed the inferences from the model which assumed a discrete support. Even though the population-level model was different from that used to generate the data, this approach recovered the population-level distribution for each range parameter (Figure B.1). Similarly, the discrete-support model was able to recover the data-generating values for both range parameters (Figure B.2). While the population-level models differ, inferences from the continuous-support and discrete-support models are also similar to one another (Figure B.2). This example shows how the beta-binomial distribution is flexible enough to allow the discrete-support model to approximate other hierarchical models for continuous parameters.

We also compared the computational efficiency of the two models by examining the effective sample sizes per computation time. This quantity is related to the autocorrelation of the MCMC

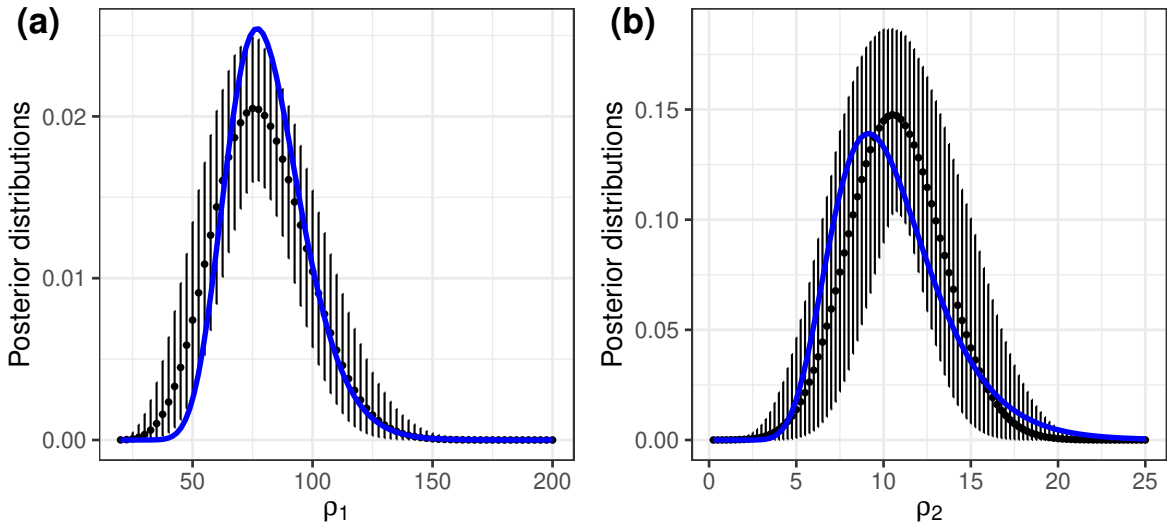

Figure B.1: Posterior distributions for the population-level probability mass functions of ρ_1 (a) and ρ_2 (b). Black points show the posterior means and lines show the 95% CIs for probabilities at each value in the assumed discrete support. The blue lines indicate the data-generating densities for each parameter.

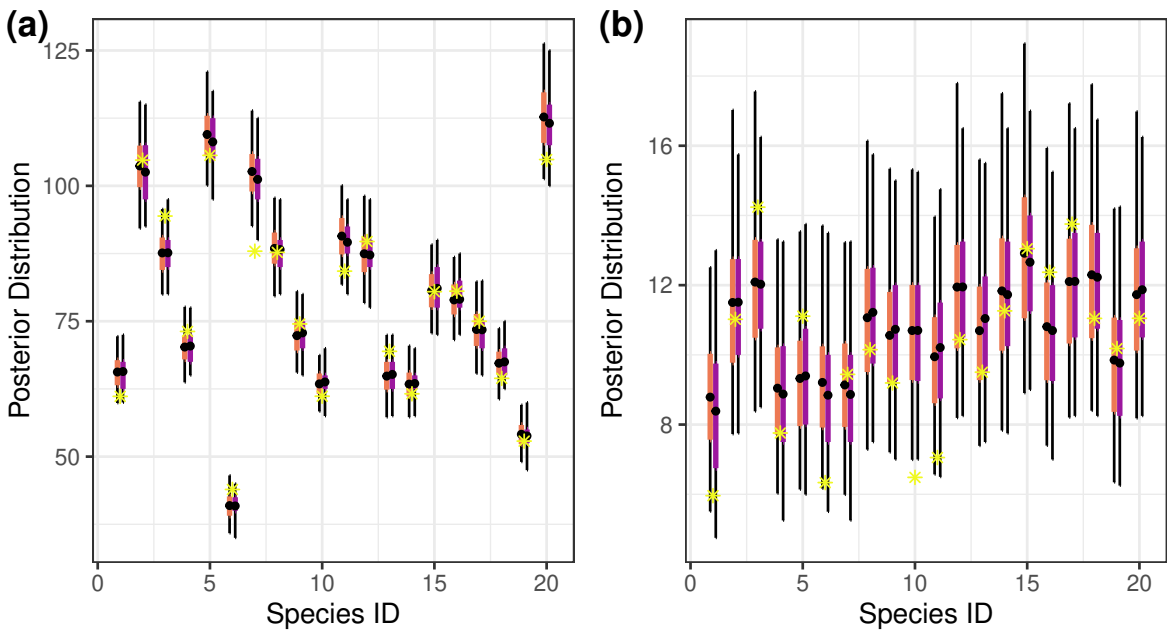

Figure B.2: Posterior distributions for the range parameters ρ_1 (a) and ρ_2 (b) for the continuous-support model (orange) and discrete-support model (purple). Black points show the posterior means, wide lines show the 50% CIs, and thin lines show the 95% CIs. Data-generating values are shown by the yellow stars.

chain and quantifies the efficiency of our algorithm. In general, the effective sample sizes were greater for the small-scale range parameter. Across both range parameters and all species, the discrete-support model had larger effective sample sizes than those of the continuous-support model (Figure B.3). Specifically, the effective sample size per computation time was over five times greater for some individual parameters. Because inferences are comparable and effective sample sizes are greater, we used the discrete-support model for the spatial range parameters to implement our model.

B.3 Model comparisons

B.3.1 Simulations

We used posterior predictive checks to assess the spatial latent factor component of our model. These posterior predictive checks are based on the cross-covariance of model residuals. We consider residuals conditional on the large-scale spatial variation to assess the model structure describing interspecies dependence within a plot. First consider the single-species model defined by main text

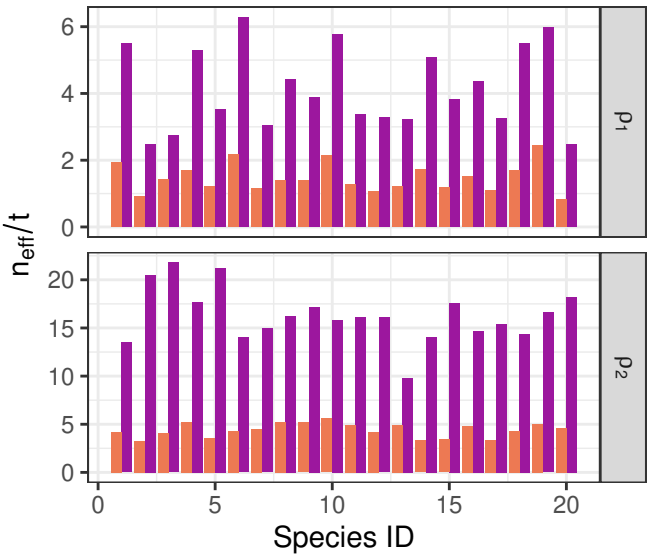

Figure B.3: From a simulated example, effective sample sizes per computation time for range parameters based on fitting the continuous-support model (orange) and discrete-support model (purple).

equations (2, 3, 4). For plot i and species j, we defined the residual for grid point d as

$$r_{ijd} = y_{ijd} - \mathrm{E}(y_{ijd} \mid \mu_{ij}), \tag{B.1}$$

where $\mathrm{E}(y_{ijd} \mid \mu_{ij})$ denotes the expected value of y_{ijd} conditional on its plot-level mean μ_{ij}. For our model, this conditional expectation is

$$\mathrm{E}(y_{ijd} \mid \mu_{ij}) = \Pr(y_{ijd} = 1 \mid \mu_{ij}) = \Phi(-\mu_{ij}), \tag{B.2}$$

where $\Phi(\cdot)$ denotes the CDF of a standard normal random variable. In Bayesian analyses, the residual in (B.1) has a posterior distribution and thus we calculate the residual for every posterior draw obtained from our MCMC algorithm.

We use these residuals to construct a posterior predictive check based on the cross-covariance for a pair of species. Binary residuals are typically binned when used in model assessments and here we bin by distance between points to construct cross-covariograms. Each distance bin $t = 1, \ldots, T$ is defined by lower bounds l_t and upper bounds u_t. The cross-covariance test quantity for two

species j and j' in bin t is

$$Q_{tjj'} = n^{-1} n_t^{-1} \sum_{i=1}^{n} \sum_{d=1}^{D} \sum_{d'=1}^{D} \mathbb{1}(\|\mathbf{v}_{id} - \mathbf{v}_{id'}\| \in [l_t, u_t)) r_{ijd} r_{ij'd'}, \tag{B.3}$$

where n_t is the number of pairs of grid points within distance bin t and $\|\mathbf{v}_{id} - \mathbf{v}_{id'}\|$ denotes the Euclidean distance between grid points d and d'. Again the test quantity $Q_{tjj'}$ is calculated using the residuals from every posterior draw to approximate its posterior distribution. Additionally, we calculated the same test quantity for replicated datasets, denoted $Q^{\text{rep}}_{tjj'}$, where the residuals are calculated based on data generated under the model based on a particular draw from the posterior distribution. Comparing $Q_{tjj'}$ and $Q^{\text{rep}}_{tjj'}$ allows us to assess whether the cross-covariances in the residuals are reasonable under the posterior predictive distribution.

We used simulated datasets to demonstrate and evaluate this posterior predictive check. We used the same predictor variables and sample sizes as in the real data example, but created data for $J = 5$ species and $M = 2$ latent factors to reduce computation times for fitting models. Data generating values for the model parameters were chosen to be similar to those in the real data. In the first scenario, we fixed the covariances among species in the simulated datasets by assuming

$$\boldsymbol{\Theta} = \begin{pmatrix} 1 & 0.75 & -1 & 0 & 0 \\ 0 & 0.75 & -0.25 & 0 & 0 \end{pmatrix}, \tag{B.4}$$

which implies species 1 and 2 are positively correlated and both of these species are negatively correlated with species 3. Data were generated such that species 4 and 5 were both uncorrelated with all the other species. We fit a model that assumed independence among species (i.e., $\boldsymbol{\theta}_j = \mathbf{0}$ for all j) and assessed model fit using the posterior predictive check defined in (B.3). The graphical assessment based on the cross-covariogram shows the direction of dependence and for which distance bins the model may be inadequate (Figure B.4). We consider the test quantity at the smallest distance bin ($\|\mathbf{v}_{id} - \mathbf{v}_{id'}\| = 0$) to summarize simulation results more broadly. Across 100 simulated datasets, the posterior predictive probabilities that $Q_{tjj'} > Q^{\text{rep}}_{tjj'}$ for distance bin at zero

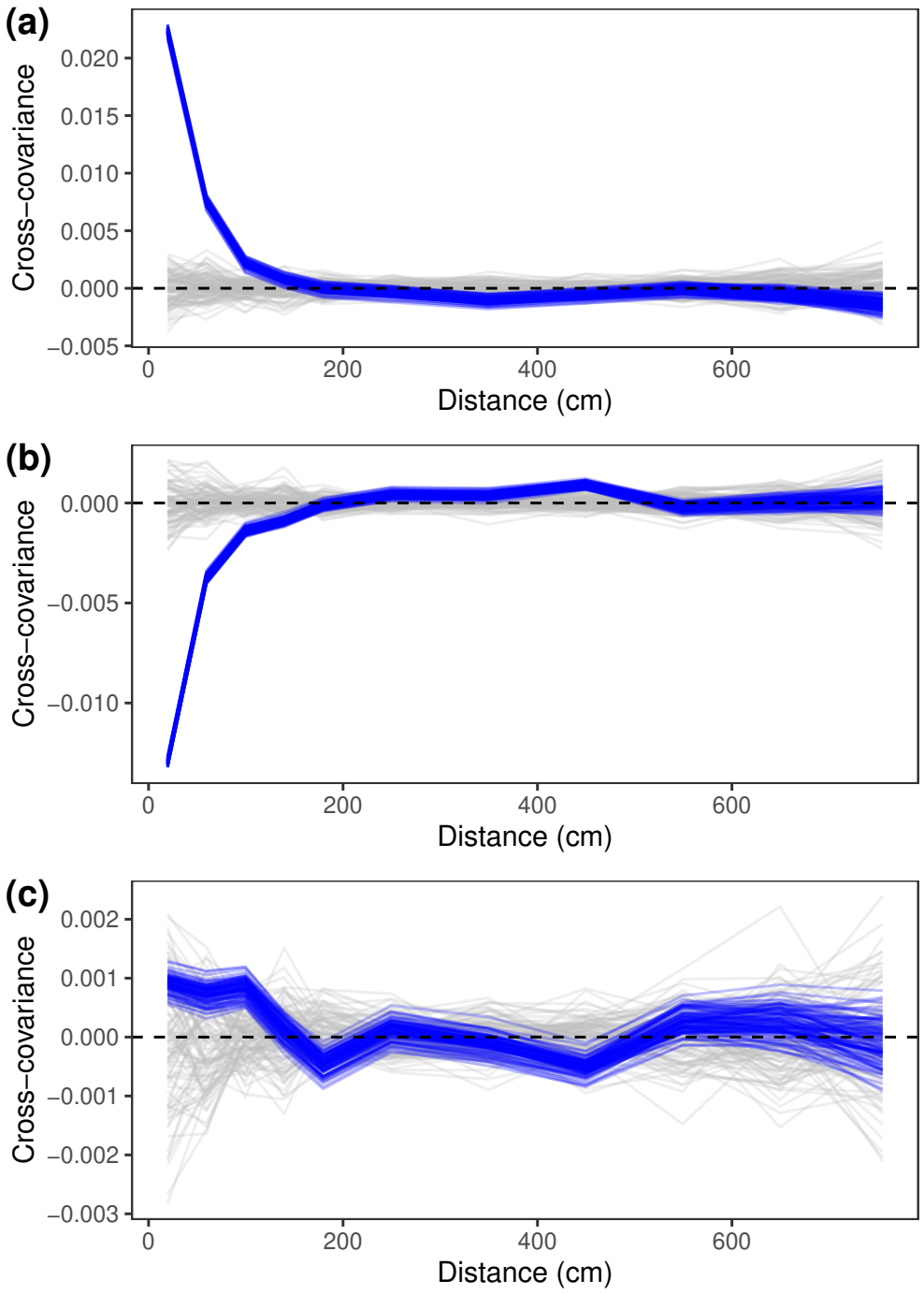

Figure B.4: For a single simulated dataset, graphical posterior predictive checks based on the cross-covariance at different distance bins. Pairs of species were generated with positive covariance (a), negative covariance (b), and no covariance (c). In each plot, the gray lines denote the test quantity from replicated datasets and the blue lines indicated test quantities from the observed data. Different lines correspond to different draws (100 random realizations) from the posterior distribution of the fitted model.

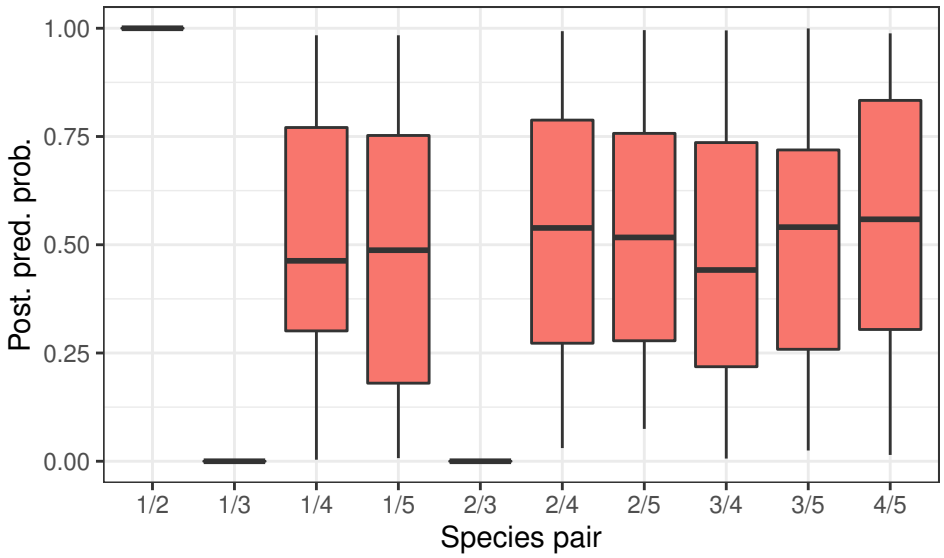

Figure B.5: Across 100 simulated datasets, the posterior predictive probabilities that $Q_{tjj'} > Q^{\text{rep}}_{tjj'}$ at distance bin zero for every species pair. Data were generated such that species 1 and 2 were positively correlated and both of these species were negatively correlated with species 3. Species 4 and 5 were uncorrelated with all other species.

was able to identify which pairs of species were inadequately described by the fitted model at a high frequency (Figure B.5). Note that posterior predictive probabilities close to zero or one suggest the observed data are inconsistent with the replicated datasets.

In the second scenario, we allowed the covariance among species to vary with pollution concentration. Data were generated the same as in scenario 1, but now with $\boldsymbol{\theta}_{12} = \boldsymbol{\theta}_{13} = [-2\ 0]'$ where these coefficients correspond to the B-spline basis matrix of the zinc concentrations, as described in main text Section 5. For all other species, the additional coefficients were all zero. Under this scenario, the covariance among species pairs 1/2 and 1/3 will vary with zinc concentration.

To modify the posterior predictive check to assess for changes in interspecies dependence, we also bin residuals by zinc concentration. Using the quantiles of the observed zinc concentrations, we considered plots with low, medium, or high zinc. Then the cross-covariance test quantity in (B.3) can be calculated separately for plots within each zinc level. We again fit a model assuming independence across species and used this posterior predictive check to assess model fit. Variability in the cross-covariograms across the different zinc levels indicates that the covariance for that pair

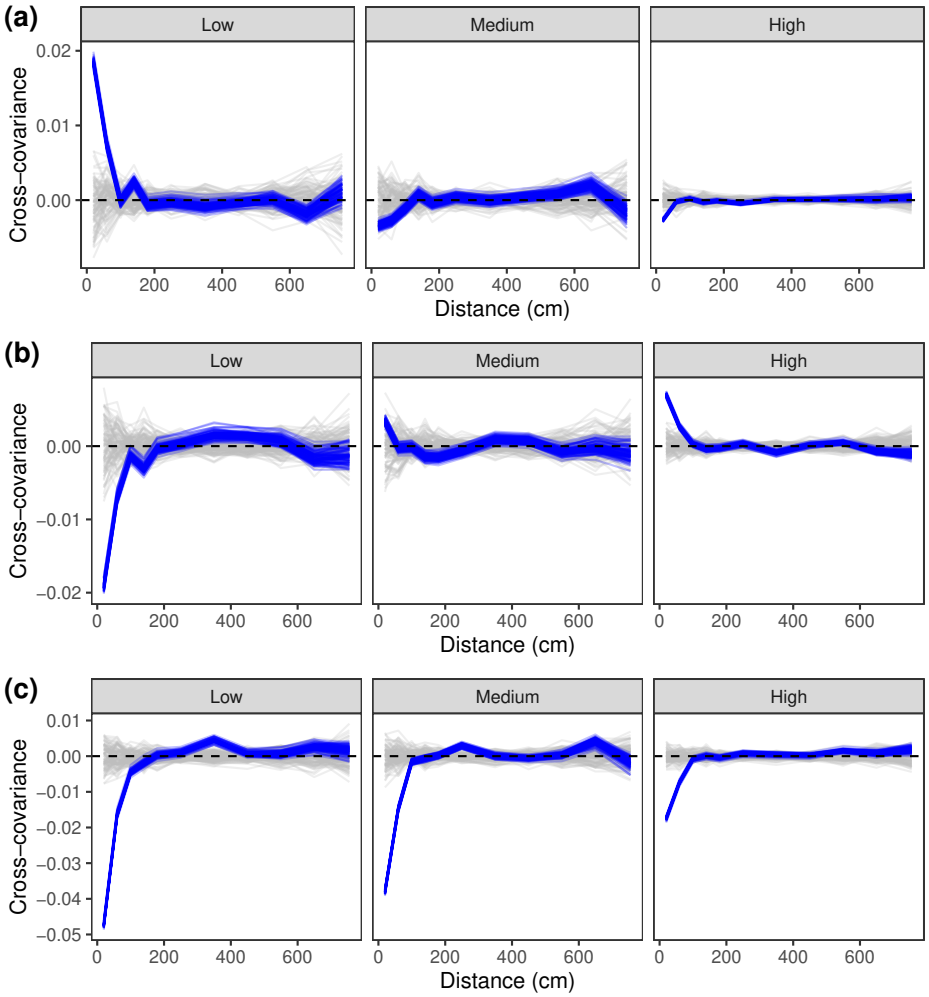

Figure B.6: For a single simulated dataset, graphical posterior predictive checks based on the cross-covariance at different distance bins and zinc concentrations. Pairs of species were generated with changing covariance (a, b) and constant covariance (c) as a function of zinc. In each plot, the gray lines denote the test quantity from replicated datasets and the blue lines indicated test quantities from the observed data. Different lines correspond to different draws (100 random realizations) from the posterior distribution of the fitted model and facets show the different zinc levels.

of species changes (Figure B.6). To summarize the performance of this assessment across multiple simulated datasets, we considered the posterior predictive probability that

$$(Q_{tjj'}^{\text{low}} - Q_{tjj'}^{\text{high}}) > (Q_{tjj'}^{\text{rep, low}} - Q_{tjj'}^{\text{rep, high}}), \tag{B.5}$$

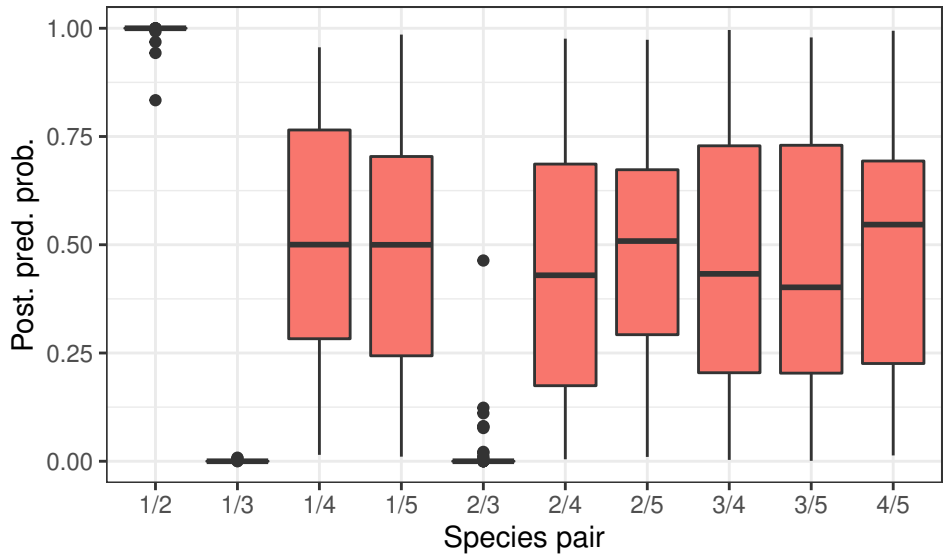

Figure B.7: Across 100 simulated datasets, the posterior predictive probabilities for whether the cross-covariance varied with zinc at distance bin zero for every species pair. Data were generated such that the covariance between species 1/2 and 1/3 changed by zinc level.

where the low and high superscripts on Q now denote plots with different zinc levels. Again this assessment appears very sensitive to model misspecification (Figure B.7). In this case, however, despite the covariance between species 2 and 3 being constant as a function of zinc, the assessment still consistently identifies that pair of species as being inconsistent with the fitted model. Data were generated such that the covariance between these species and species 1 changes. The changing covariances with species 1 appears to affect the covariance among species 2 and 3 indirectly (see also Figure B.6c).

B.3.2 CAKR data

We applied these posterior predictive assessments to models fit to the CAKR data. We first considered a simple model that assumed independence among species (i.e., $\boldsymbol{\theta}_{jp} = \mathbf{0}$ for all j and p). Then we included the latent factor structure described in the main text. We performed the posterior predictive checks for both fitted models. For the model assuming independence, we found evidence of residual correlation for many pairs of species (e.g., left column of Figure B.8). The assumed latent factor structure was able to account for this dependence (Figure B.8). For some species pairs,

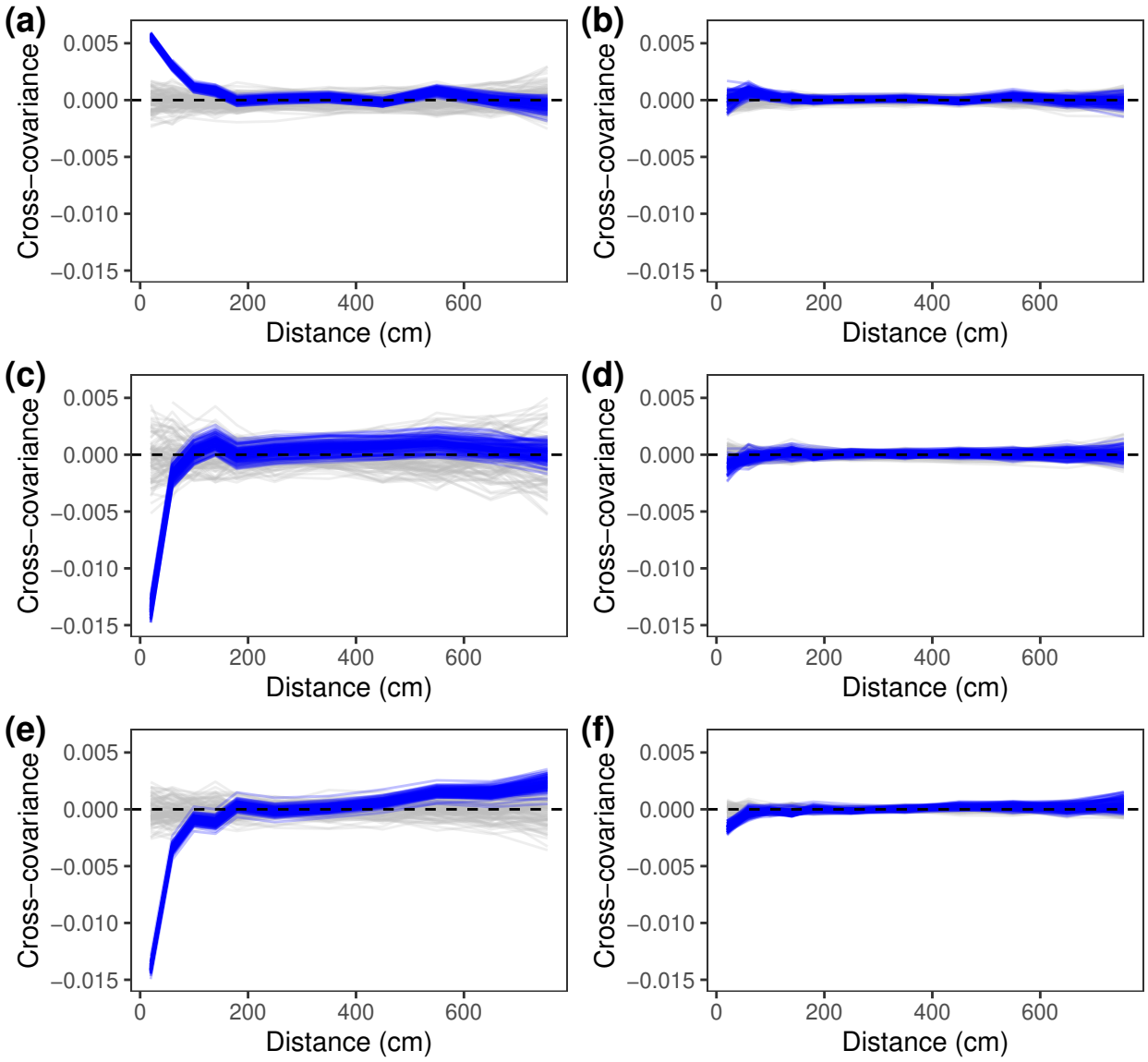

Figure B.8: For the CAKR analysis, posterior predictive checks for three pairs of species (a different pair for each row). Diagnostics for the model assuming independence among species (a, c, e) and for the model including latent factors (b, d, f).

there is still evidence of a small amount of negative residual cross-covariance at very small distances (e.g., see Figure B.8f).

For the model assuming independence, the posterior predictive checks also provided evidence that the correlation among species could vary with pollution level (Figure B.9a). The latent factor structure accounted for the varying cross-covariance among species (Figure B.9b).

B.4 Additional CAKR results

There is also interest in inferences for community-level measures and how these are related to pollutant concentrations. While our approach models cover for individual species, we can still obtain community-level inferences from our model. For instance, inferences for overall lichen cover can be obtained by aggregating across the individual lichen species. We used this strategy to obtain posterior inferences on how expected lichen species richness, expected lichen cover, and expected *Sphagnum* cover is related to zinc concentrations. These quantities are obtained numerically using a posterior predictive approach. Specifically, we generate grid point data at plots with different zinc concentrations and aggregate the predictions as necessary for the community-level inferences.

At the species-level, cover of the lichen and moss species decreased as zinc concentration increased. This leads to declines in the community-level quantities as zinc increases as well (Figure

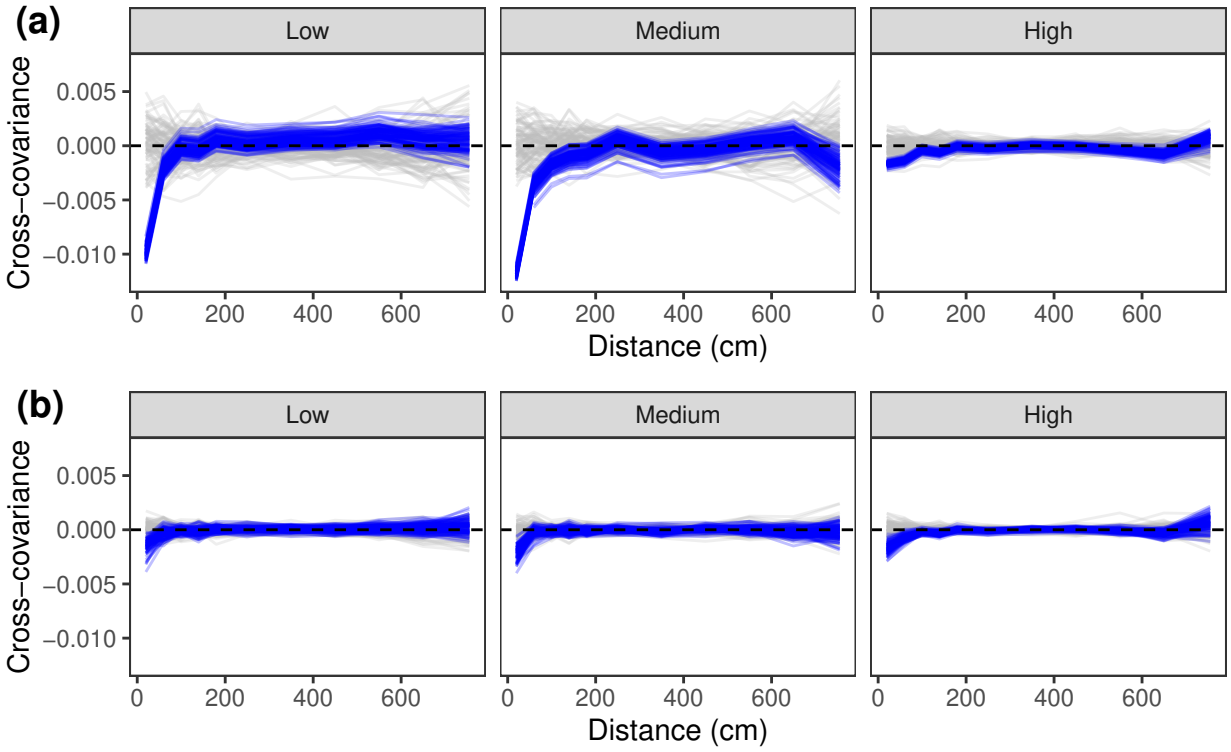

Figure B.9: Posterior predictive checks for a model assuming independence among species (a) and for a model including the latent factor structure (b). This shows an example for one pair of species and examines the residual diagnostic across the different levels of zinc concentration.

B.10). We also considered the posterior distributions for various thresholds that define what zinc concentration results in the quantity of interest dropping to 95%, 75%, 50%, 25%, and 5% of its maximum (Table B.1). These thresholds can be used to help identify pollution levels that are needed to maintain healthy vegetation communities within CAKR.

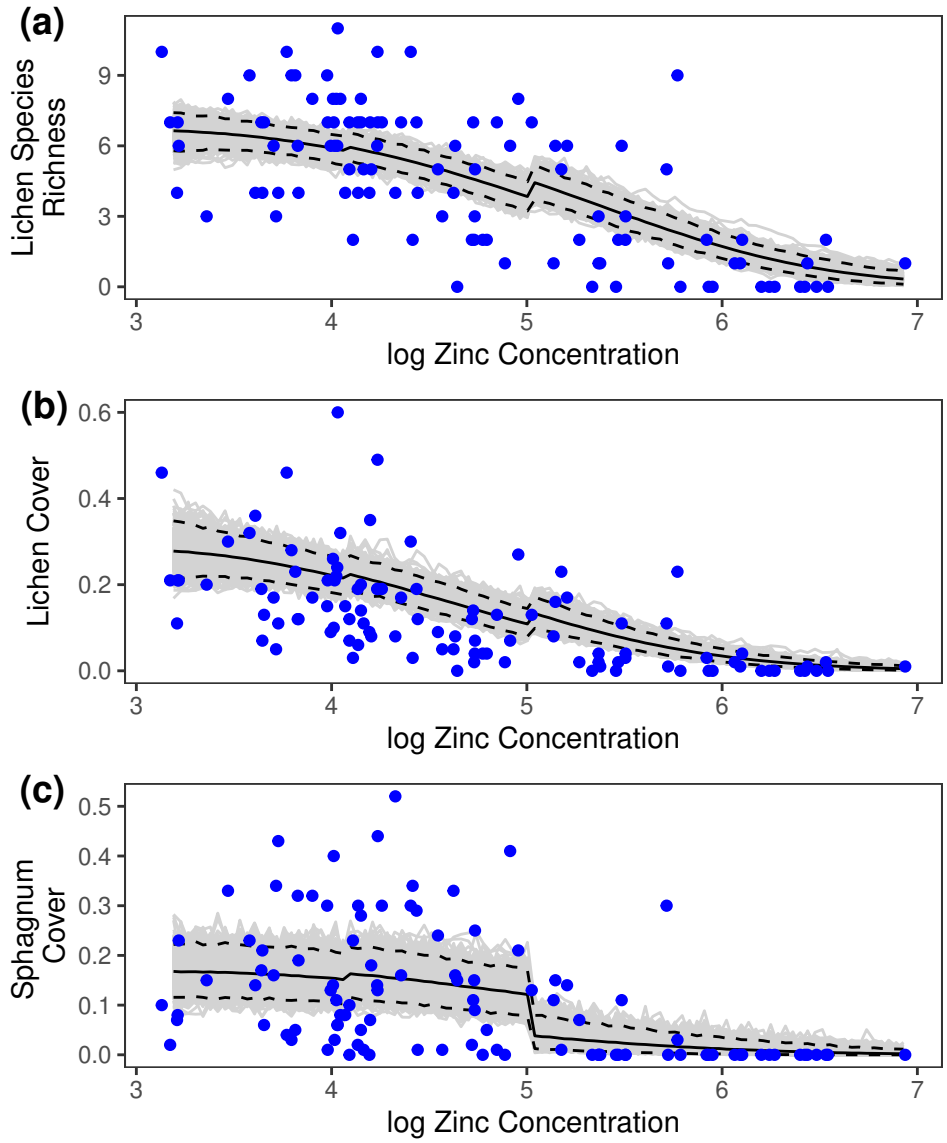

Figure B.10: Posterior distributions of expected lichen species richness (a), expected lichen cover (b), and expected Sphagnum cover (c) versus log zinc concentration. Gray lines show random samples from the posterior distribution, solid black lines show the posterior mean, and the dashed black lines show the pointwise 95% CIs. The observed data are indicated by the blue points.

Table B.1: Posterior distributions of zinc thresholds for expected lichen species richness, expected lichen cover, and expected *Sphagnum* cover. Each threshold is calculated as the zinc concentration that results in the quantity of interest dropping to that percent of its maximum.

Lichen species richness		
Threshold	Post. mean	95% CI
95%	40.9	(27.2, 64.9)
75%	96.9	(67.3, 160.7)
50%	218.2	(166.7, 272.4)
25%	404.9	(328.7, 498.4)
5%	942.9	(727.0, 1021.3)

Lichen cover		
Threshold	Post. mean	95% CI
95%	33.8	(24.3, 55.8)
75%	64.7	(41.2, 94.6)
50%	126.0	(81.3, 186.7)
25%	251.6	(193.9, 329.0)
5%	633.1	(479.9, 911.9)

***Sphagnum* cover**		
Threshold	Post. mean	95% CI
95%	53.2	(25.2, 98.2)
75%	111.2	(57.9, 148.8)
50%	147.6	(132.9, 148.8)
25%	174.1	(148.8, 316.8)
5%	559.5	(272.4, 1021.3)

Appendix C

Supplemental Information for Chapter 4

C.1 Additional MCMC details

Our MCMC algorithm uses a Gibbs sampling approach and we detail the necessary full-conditional distributions here. To update the detection-level parameters, we introduce latent variables $\tilde{y}_{ij}$ such that $y_{ij} = \mathbb{1}(\tilde{y}_{ij} \geq 0)$ (Albert and Chib, 1993). The full-conditional distributions for these latent variables are

$$(\tilde{y}_{ij} \mid y_{ij}, \boldsymbol{\alpha}, \gamma, \tilde{\mathbf{z}}) \sim \begin{cases} \text{TN}\left(\mu_{\tilde{y}_{ij}}, 1, 0, \infty\right), & \text{if } y_{ij} = 1, \\ \text{TN}\left(\mu_{\tilde{y}_{ij}}, 1, -\infty, 0\right), & \text{if } y_{ij} = 0, \end{cases} \tag{C.1}$$

where $\mu_{\tilde{y}_{ij}} = \mathbf{w}'_{ij}\boldsymbol{\alpha} + \gamma D_i^{-1} \sum_{\mathbf{s}_d \in \mathcal{A}_i} \mathbb{1}(\tilde{z}(\mathbf{s}_d) \geq 0)$ and $\text{TN}(\mu, \sigma^2, a, b)$ denotes a normal distribution with mean μ and variance σ^2 that has been truncated to the interval (a, b).

The detection-level parameters $(\boldsymbol{\alpha}, \gamma)$ can be updated jointly and their full-conditional distribution is conjugate given the latent variables $\tilde{y}_{ij}$ and spatial effects $\tilde{\mathbf{z}}$. We let $\mathbf{y}$ and $\tilde{\mathbf{y}}$ denote vectors of the observed detection data and corresponding latent variables, respectively. These vectors must be restricted to only include observations from sites that are occupied (i.e., (i, j) such that $\max_{\mathbf{s}_d \in \mathcal{A}_i} \mathbb{1}(\tilde{z}(\mathbf{s}_d) \geq 0)$). This restriction is necessary because unoccupied sites provide no information about the detection parameters (species cannot be detected there) and using these nondetections would bias inferences for the detection coefficients. Then, conditional on $\tilde{\mathbf{z}}$, we define the design matrix $\mathbf{W}$ to have rows equal to $(\mathbf{w}'_{ij}, D_i^{-1} \sum_{\mathbf{s}_d \in \mathcal{A}_i} \mathbb{1}(\tilde{z}(\mathbf{s}_d) \geq 0))$. The matrix $\mathbf{W}$ is also restricted to only include rows corresponding to occupied sites. We assume that the joint prior distribution for $(\boldsymbol{\alpha}, \gamma)$ is $\text{N}(\boldsymbol{\mu}_{\alpha,\gamma}, \boldsymbol{\Sigma}_{\alpha,\gamma})$. Standard calculations show that the full-conditional distribution for $(\boldsymbol{\alpha}, \gamma)$ is

$$(\boldsymbol{\alpha}, \gamma \mid \tilde{\mathbf{y}}, \tilde{\mathbf{z}}) \sim \text{N}(\mathbf{A}^{-1}\mathbf{b}, \mathbf{A}^{-1}), \tag{C.2}$$

where $\mathbf{A} = \mathbf{W}'\mathbf{W} + \boldsymbol{\Sigma}_{\alpha,\gamma}^{-1}$ and $\mathbf{b} = \mathbf{W}'\tilde{\mathbf{y}} + \boldsymbol{\Sigma}_{\alpha,\gamma}^{-1}\boldsymbol{\mu}_{\alpha,\gamma}$.

We implement our MCMC algorithm by integrating $\boldsymbol{\beta}$ out of the model. Conditional on the draws of $\tilde{\mathbf{z}}$, we can sample $\boldsymbol{\beta}$ from its full-conditional distribution after the MCMC algorithm has completed. We assume a normal prior distribution for $\boldsymbol{\beta}$ so that $\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$. The resulting full-conditional distribution is also conjugate and it can be shown to be

$$(\boldsymbol{\beta} \mid \tilde{\mathbf{z}}, \boldsymbol{\theta}) \sim \mathrm{N}(\mathbf{A}^{-1}\mathbf{b}, \mathbf{A}^{-1}), \tag{C.3}$$

where $\mathbf{A} = \mathbf{X}'\boldsymbol{\Sigma}_\eta^{-1}\mathbf{X} + \boldsymbol{\Sigma}_\beta^{-1}$ and $\mathbf{b} = \mathbf{X}'\boldsymbol{\Sigma}_\eta^{-1}\tilde{\mathbf{z}} + \boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta$.

The auxiliary variables introduced for our surrogate data slice sampler preserve the joint distribution of the partitioned spatial terms $(\tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2)$. Additionally, this construction defines the joint distribution of $(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \mathbf{g}, \tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2)$ to be multivariate normal. The first step of our surrogate data slice sampler updates the surrogate data $\mathbf{g}$ from $[\mathbf{g} \mid \tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2, \boldsymbol{\theta}]$. To determine this distribution, first note that

$$\mathrm{Cov}(\mathbf{g}, \tilde{\mathbf{z}}_1) = \mathrm{Cov}(\mathbf{g}, \boldsymbol{\Sigma}_{11}(\boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_g)^{-1}\mathbf{g}) = \boldsymbol{\Sigma}_{11}, \tag{C.4}$$

and

$$\mathrm{Cov}(\mathbf{g}, \tilde{\mathbf{z}}_2) = \mathrm{Cov}(\mathbf{g}, \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{11}(\boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_g)^{-1}\mathbf{g}) = \boldsymbol{\Sigma}_{12}. \tag{C.5}$$

Using these covariances, the conditional distribution $[\mathbf{g} \mid \tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2, \boldsymbol{\theta}]$ is normal with mean

$$\mathrm{E}(\mathbf{g} \mid \tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2, \boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}^{-1} \begin{pmatrix} \tilde{\mathbf{z}}_1 \\ \tilde{\mathbf{z}}_2 \end{pmatrix} \tag{C.6}$$

$$= \begin{pmatrix} \mathbf{I} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}^{-1} \begin{pmatrix} \tilde{\mathbf{z}}_1 \\ \tilde{\mathbf{z}}_2 \end{pmatrix} \tag{C.7}$$

$$= \tilde{\mathbf{z}}_1, \tag{C.8}$$

and variance

$$\text{Var}(\mathbf{g} \mid \tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2, \boldsymbol{\theta}) = \boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_g - \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\Sigma}_{11} \\ \boldsymbol{\Sigma}_{21} \end{pmatrix} \tag{C.9}$$

$$= \boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_g - \begin{pmatrix} \mathbf{I} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I} \\ \mathbf{0} \end{pmatrix} \tag{C.10}$$

$$= \boldsymbol{\Sigma}_g, \tag{C.11}$$

using standard properties of the multivariate normal distribution. Thus we can update the surrogate data directly using this conditional distribution.

C.2 Details of NNGP

We use a NNGP to model the spatial terms $\tilde{\mathbf{z}}$ at the point locations after integrating $\boldsymbol{\beta}$ out of the model. This results in the spatial terms being modeled as

$$\tilde{\mathbf{z}} \sim \text{N}(\boldsymbol{\mu}_{\tilde{z}}, \tilde{\boldsymbol{\Sigma}}_{\tilde{z}}), \tag{C.12}$$

where $\tilde{\boldsymbol{\Sigma}}_{\tilde{z}}$ is the NNGP approximation for the full covariance matrix $\boldsymbol{\Sigma}_{\tilde{z}}$. Datta et al. (2016) show that the corresponding precision matrix $\tilde{\boldsymbol{\Sigma}}_{\tilde{z}}^{-1}$ is sparse by construction. Specifically,

$$\tilde{\boldsymbol{\Sigma}}_{\tilde{z}}^{-1} = (\mathbf{I} - \mathbf{B})' \mathbf{F}^{-1} (\mathbf{I} - \mathbf{B}), \tag{C.13}$$

where the rows of $\mathbf{B}$ are defined as

$$\mathbf{b}_d' = K(\mathbf{s}_{c(d)}, \mathbf{s}_{c(d)})^{-1} K(\mathbf{s}_{c(d)}, \mathbf{s}_d), \tag{C.14}$$

and $\mathbf{F}$ is a diagonal matrix with elements

$$f_d = K(\mathbf{s}_d, \mathbf{s}_d) - K(\mathbf{s}_d, \mathbf{s}_{c(d)})\mathbf{b}_d'. \tag{C.15}$$

In (C.14) and (C.15) the function $K(\cdot, \cdot)$ corresponds to the spatial covariance function that generates $\boldsymbol{\Sigma}_{\tilde{z}}$. This function depends on parameters $\boldsymbol{\theta}$ as well as spatial predictor variables $\mathbf{X}$ and the prior distribution for $\boldsymbol{\beta}$. Additionally, note that in (C.14), the elements of $\mathbf{b}_d$ corresponding to observations that are not neighbors of location d are equal to zero (Datta et al., 2016; Datta, 2022).

Our MCMC algorithm requires finding $\tilde{\mathbf{z}}$ by solving the system $\mathbf{L}^{-1}\tilde{\mathbf{z}} = \boldsymbol{\nu}$ where $\mathbf{L}$ is defined to be the square root matrix of $\tilde{\boldsymbol{\Sigma}}_{\tilde{z}}$ such that $\tilde{\boldsymbol{\Sigma}}_{\tilde{z}} = \mathbf{L}\mathbf{L}'$. This step is conditional on values of the covariance parameters $\boldsymbol{\theta}$ and decorrelated spatial terms $\boldsymbol{\nu}$. Saha and Datta (2018) and Datta (2022) developed an efficient algorithm for finding this solution and we describe that approach here. First, we let $l_{dd'}$ denote the elements of $\mathbf{L}^{-1}$. Then the correlated spatial terms $\tilde{\mathbf{z}}$ can be found as

$$\begin{aligned}
\tilde{z}_1 &= \nu_1 / l_{11}, \\
\tilde{z}_2 &= (\nu_2 - l_{21}\tilde{z}_1)\, l_{22}, \\
&\vdots \\
\tilde{z}_d &= \left(\nu_d - \sum_{d' < d;\, l_{dd'} \neq 0} l_{dd'} \tilde{z}_{d'} \right) / l_{dd}.
\end{aligned} \tag{C.16}$$

These calculations can be simplified by observing that

$$\tilde{\boldsymbol{\Sigma}}_{\tilde{z}}^{-1} = (\mathbf{I} - \mathbf{B})'\mathbf{F}^{-1}(\mathbf{I} - \mathbf{B}) = (\mathbf{L}')^{-1}\mathbf{L}^{-1}, \tag{C.17}$$

implies that $\mathbf{L}^{-1} = \mathbf{F}^{-1/2}(\mathbf{I} - \mathbf{B}) = \mathbf{F}^{-1/2} - \mathbf{F}^{-1/2}\mathbf{B}$. Because $\mathbf{F}$ is diagonal and $\mathbf{B}$ is sparse, the elements of $\mathbf{L}^{-1}$ are equal to

$$l_{dd'} = \begin{cases} f_d^{-1/2}, & d = d', \\ -f_d^{-1/2} b_{dd'}, & d \neq d', \end{cases} \tag{C.18}$$

where $b_{dd'}$ denotes the elements of $\mathbf{B}$. Then the last line of (C.16) can be rewritten as

$$\tilde{z}_d = \nu_d f_d^{1/2} + \sum_{d'<d;\, b_{dd'}\neq 0} b_{dd'}\tilde{z}_{d'}, \tag{C.19}$$

and there is no need to explicitly construct $\mathbf{L}^{-1}$. We used this approach to efficiently solve for $\tilde{\mathbf{z}}$ in our MCMC algorithm. These same equations can also be used to find the decorrelated spatial terms $\boldsymbol{\nu}$ conditional on the spatial terms $\tilde{\mathbf{z}}$.

C.3 Alternative occupancy models

We used a simulation study to compare our continuous-space occupancy model to other spatial occupancy models. We provide additional details about these alternative approaches and their implementation in what follows.

C.3.1 Areal occupancy model

The first alternative approach assumes that the species occupancy process is defined for the same areal survey units where the detection/nondetection data are collected. This is a standard occupancy model that includes spatial dependence among the areal survey units. We assume an approximate ICAR structure for the spatial terms as described by Johnson et al. (2013) and Hooten and Hefley (2019).

We model the indicator for species occurrence at site i as $z_i = \mathbb{1}(\tilde{z}_i \geq 0)$ where

$$\tilde{\mathbf{z}} \sim \text{N}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta}, \mathbf{I}), \tag{C.20}$$

$$\boldsymbol{\eta} \sim \text{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{C.21}$$

for $i = 1, \ldots, n$. The spatial terms $\boldsymbol{\eta}$ have covariance $\boldsymbol{\Sigma}$ that is assumed to have the form

$$\boldsymbol{\Sigma} = \sigma^2(\text{diag}(\mathbf{H1}) - \rho\mathbf{H})^{-1}, \tag{C.22}$$

where $\mathbf{H}$ is an adjacency matrix, σ^2 is a variance parameter, and $\rho = 0.99$ to approximate an ICAR covariance. For our simulations on a regular grid, we used the queen definition to define neighbors in the adjacency matrix. We model the detection/nondetection data for visits $j = 1, \ldots, J_i$ as

$$y_{ij} \sim \begin{cases} \mathbb{1}(y_{ij} = 0), & z_i = 0, \\ \text{Bernoulli}(p_{ij}), & z_i = 1, \end{cases} \tag{C.23}$$

where $\Phi^{-1}(p_{ij}) = \mathbf{w}'_{ij}\boldsymbol{\alpha}$. As described in the main text and Appendix C.1, we can introduce latent variables $\tilde{y}_{ij}$ to implement the detection component of this model.

We assume prior distributions

$$\boldsymbol{\beta} \sim \text{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta), \tag{C.24}$$

$$\boldsymbol{\alpha} \sim \text{N}(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha), \tag{C.25}$$

$$\sigma^2 \sim \text{IG}(a, b). \tag{C.26}$$

The following Gibbs updates can then be used to implement this model (Johnson et al., 2013; Hooten and Hefley, 2019):

1. Update $\boldsymbol{\beta}$ from

$$\boldsymbol{\beta} \mid \cdot \sim \text{N}(\mathbf{A}^{-1}\mathbf{b}, \mathbf{A}^{-1}), \tag{C.27}$$

   where $\mathbf{A} = \mathbf{X}'\mathbf{X} + \boldsymbol{\Sigma}_\beta^{-1}$ and $\mathbf{b} = \mathbf{X}'(\tilde{\mathbf{z}} - \boldsymbol{\eta}) + \boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta$.

2. Update each z_i from

$$z_i \mid \cdot \sim \text{Bernoulli}(\tilde{\psi}_i), \tag{C.28}$$

   where

$$\tilde{\psi}_i = \frac{\psi_i \prod_{j=1}^{J_i}(1 - p_{ij})}{\psi_i \prod_{j=1}^{J_i}(1 - p_{ij}) + (1 - \psi_i)}, \tag{C.29}$$

   and $\psi_i = \Phi^{-1}(\mathbf{x}'_i\boldsymbol{\beta} + \eta_i)$.

3. Update each $\tilde{z}_i$ from

$$\tilde{z}_i \mid \cdot \sim \begin{cases} \text{TN}(\mathbf{x}_i'\boldsymbol{\beta} + \eta_i, 1, -\infty, 0), & z_i = 0, \\ \text{TN}(\mathbf{x}_i'\boldsymbol{\beta} + \eta_i, 1, 0, \infty), & z_i = 1, \end{cases} \tag{C.30}$$

where TN denotes a truncated normal distribution.

4. Update $\boldsymbol{\eta}$ from

$$\boldsymbol{\eta} \mid \cdot \sim \text{N}(\mathbf{A}^{-1}\mathbf{b}, \mathbf{A}^{-1}), \tag{C.31}$$

where $\mathbf{A} = \mathbf{I} + \boldsymbol{\Sigma}^{-1}$ and $\mathbf{b} = \tilde{\mathbf{z}} - \mathbf{X}\boldsymbol{\beta}$.

5. Update σ^2 from

$$\sigma^2 \mid \cdot \sim \text{IG}\left(a + n/2, b + \boldsymbol{\eta}'(\text{diag}(\mathbf{H1}) - \rho\mathbf{H})\boldsymbol{\eta}/2\right), \tag{C.32}$$

where $\text{IG}(a, b)$ denotes an inverse-gamma distribution with shape a and scale b.

6. Update $\boldsymbol{\alpha}$ from

$$\boldsymbol{\alpha} \mid \cdot \sim \text{N}(\mathbf{A}^{-1}\mathbf{b}, \mathbf{A}^{-1}), \tag{C.33}$$

where $\mathbf{A} = \mathbf{W}_2'\mathbf{W}_2 + \boldsymbol{\Sigma}_\alpha^{-1}$ and $\mathbf{b} = \mathbf{W}_2'\tilde{\mathbf{y}}_2 + \boldsymbol{\Sigma}_\alpha^{-1}\boldsymbol{\mu}_\alpha$. Here $\tilde{\mathbf{y}}_2$ and $\mathbf{W}_2$ are restricted to only include indices i where $z_i = 1$ for a given update.

7. Update each $\tilde{y}_{ij}$ from

$$\tilde{y}_{ij} \mid \cdot \sim \begin{cases} \text{TN}\left(\mathbf{w}_{ij}'\boldsymbol{\alpha}, 1, 0, \infty\right), & \text{if } y_{ij} = 1, \\ \text{TN}\left(\mathbf{w}_{ij}'\boldsymbol{\alpha}, 1, -\infty, 0\right), & \text{if } y_{ij} = 0. \end{cases} \tag{C.34}$$

C.3.2 Centroid occupancy model

The second alternative model we considered assumes that occupancy is modeled in continuous space, but ignores the areal nature of the detection/nondetection data. That is, it treats the observed

data as corresponding to the centroid of the areal survey unit to avoid accounting for the change of spatial support. We also define this approach using a clipped Gaussian process to make it more comparable to our continuous-space approach that includes a change of support.

For this approach, we consider species occurrence at spatial points $\mathbf{s}_i$ corresponding to the centroids of sites $i = 1, \ldots, n$. We model the indicator for species occurrence at these locations as $z_i = \mathbb{1}(\tilde{z}_i \geq 0)$ where

$$\tilde{\mathbf{z}} \sim \text{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}). \tag{C.35}$$

Here $z_i \equiv z(\mathbf{s}_i)$ and $\tilde{\mathbf{z}} \equiv (\tilde{z}(\mathbf{s}_1), \ldots, \tilde{z}(\mathbf{s}_n))'$. As described in the main text, we assume a Gaussian covariance function to define $\boldsymbol{\Sigma}$ so this covariance depends on a single range parameter ρ. The detection model is then defined as that for standard occupancy models where

$$y_{ij} \sim \begin{cases} \mathbb{1}(y_{ij} = 0), & z_i = 0, \\ \text{Bernoulli}(p_{ij}), & z_i = 1, \end{cases} \tag{C.36}$$

and $\Phi^{-1}(p_{ij}) = \mathbf{w}'_{ij}\boldsymbol{\alpha}$. Note that this approach assumes that whenever the species is detected within a site, it occurs exactly at point location $\mathbf{s}_i$ (site centroid). Consequently, even though the occupancy process is modeled in continuous space, this model does not account for the change of spatial support that is needed to use areal survey data.

For this model we assume prior distributions

$$\boldsymbol{\beta} \sim \text{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta), \tag{C.37}$$

$$\boldsymbol{\alpha} \sim \text{N}(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha), \tag{C.38}$$

$$\rho \sim \text{Lognormal}(\mu_\rho, \sigma^2_\rho). \tag{C.39}$$

The point occupancy model can be implemented using the following Gibbs sampling steps:

1. Update each $\tilde{z}_i$ from

$$\tilde{z}_i \mid \cdot \sim \begin{cases} \text{TN}(\tilde{\mu}_i, \tilde{\sigma}_i^2, 0, \infty), & \text{if } \sum_{j=1}^{J_i} y_{ij} > 0, \\ \tilde{\psi}_i \text{TN}(\tilde{\mu}_i, \tilde{\sigma}_i^2, 0, \infty) + (1 - \tilde{\psi}_i)\text{TN}(\tilde{\mu}_i, \tilde{\sigma}_i^2, -\infty, 0), & \text{if } \sum_{j=1}^{J_i} y_{ij} = 0, \end{cases} \tag{C.40}$$

   where $\tilde{\mu}_i$ and $\tilde{\sigma}_i^2$ are the conditional mean $\text{E}(\tilde{z}_i \mid \tilde{\mathbf{z}}_{-i})$ and variance $\text{Var}(\tilde{z}_i \mid \tilde{\mathbf{z}}_{-i})$, respectively. These can be found using standard properties of the multivariate normal distribution. Using Bayes' Theorem, the conditional probability of occurrence $\tilde{\psi}_i$ is

$$\tilde{\psi}_i = \frac{\psi_i \prod_{j=1}^{J_i}(1 - p_{ij})}{(1 - \psi_i) + \psi_i \prod_{j=1}^{J_i}(1 - p_{ij})}, \tag{C.41}$$

   where $\psi_i = \int_0^\infty \text{N}(x \mid \tilde{\mu}_i, \tilde{\sigma}_i^2)dx$. Note that when the species is undetected, we update $\tilde{z}_i$ with a random draw from a mixture of truncated normal distributions with mixture probability $\tilde{\psi}_i$. Also, this step effectively updates $\tilde{z}_i$ and z_i simultaneously.

2. Update $\boldsymbol{\beta}$ from

$$\boldsymbol{\beta} \sim \text{N}(\mathbf{A}^{-1}\mathbf{b}, \mathbf{A}^{-1}), \tag{C.42}$$

   where $\mathbf{A} = \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X} + \boldsymbol{\Sigma}_\beta$ and $\mathbf{b} = \mathbf{X}'\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{z}} + \boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta$.

3. Update $\boldsymbol{\alpha}$ using step 6 of the MCMC algorithm for the areal occupancy model.

4. Update each $\tilde{y}_{ij}$ using step 7 of the MCMC algorithm for the areal occupancy model.

5. Update ρ using a Metropolis-Hastings step.

To approximate the proportion of the study region occupied we considered posterior predictions of $\tilde{z}$ over a grid of additional points. This grid consisted of the same points as that used to implement our continuous-space occupancy model. We fit this model without approximating the spatial covariance structure among the site centroids. However, we again used a NNGP to more efficiently implement the posterior predictions for this model.