Received: 6 April 2022 | Revised: 26 May 2023 | Accepted: 13 June 2023

DOI: 10.1002/ecy.4137

STATISTICAL REPORT

ECOLOGY
ECOLOGICAL SOCIETY OF AMERICA

Joint species distribution models with imperfect detection for high-dimensional spatial data

Jeffrey W. Doser[1,2] | **Andrew O. Finley**[2,3] | **Sudipto Banerjee**[4]

[1]Department of Integrative Biology, Michigan State University, East Lansing, Michigan, USA

[2]Ecology, Evolution, and Behavior Program, Michigan State University, East Lansing, Michigan, USA

[3]Department of Forestry, Michigan State University, East Lansing, Michigan, USA

[4]Department of Biostatistics, University of California, Los Angeles, California, USA

Correspondence
Jeffrey W. Doser
Email: doserjef@msu.edu

Funding information
National Science Foundation, Grant/Award Numbers: DEB-2213565, DMS-1916395, EF-1253225

Handling Editor: Viviana Ruiz-Gutierrez

Abstract

Determining the spatial distributions of species and communities is a key task in ecology and conservation efforts. Joint species distribution models are a fundamental tool in community ecology that use multi-species detection–nondetection data to estimate species distributions and biodiversity metrics. The analysis of such data is complicated by residual correlations between species, imperfect detection, and spatial autocorrelation. While many methods exist to accommodate each of these complexities, there are few examples in the literature that address and explore all three complexities simultaneously. Here we developed a spatial factor multi-species occupancy model to explicitly account for species correlations, imperfect detection, and spatial autocorrelation. The proposed model uses a spatial factor dimension reduction approach and Nearest Neighbor Gaussian Processes to ensure computational efficiency for data sets with both a large number of species (e.g., >100) and spatial locations (e.g., 100,000). We compared the proposed model performance to five alternative models, each addressing a subset of the three complexities. We implemented the proposed and alternative models in the *spOccupancy* software, designed to facilitate application via an accessible, well documented, and open-source R package. Using simulations, we found that ignoring the three complexities when present leads to inferior model predictive performance, and the impacts of failing to account for one or more complexities will depend on the objectives of a given study. Using a case study on 98 bird species across the continental US, the spatial factor multi-species occupancy model had the highest predictive performance among the alternative models. Our proposed framework, together with its implementation in *spOccupancy*, serves as a user-friendly tool to understand spatial variation in species distributions and biodiversity while addressing common complexities in multi-species detection–nondetection data.

KEYWORDS
Bayesian, latent factor, Nearest Neighbor Gaussian Process, occupancy model

INTRODUCTION

Understanding the spatial distributions of species and communities is a fundamental task in ecology research and conservation efforts. Species distribution models (SDMs) are popular for predicting species distributions and understanding species–habitat relationships across space and time (Guisan & Zimmermann, 2000), which have informed key developments in ecological theory as well as conservation and management decisions (Bateman et al., 2020). While SDMs can use different data types, they most commonly use binary detection–nondetection data. Advances in hierarchical modeling have addressed many issues encountered when modeling multi-species detection–nondetection data. In particular, the three major complexities are (1) residual species correlations (Ovaskainen et al., 2010), (2) imperfect detection (MacKenzie et al., 2002), and (3) spatial autocorrelation (Banerjee et al., 2014; Latimer et al., 2009).

Joint species distribution models (JSDMs) are regression-based approaches that extend SDMs to jointly model multiple species simultaneously (Latimer et al., 2009; Ovaskainen et al., 2010). Many JSDMs jointly model species within a single model by explicitly accommodating residual species correlations, which facilitates co-occurrence hypothesis testing (Ovaskainen et al., 2010) and increases the precision of both individual species distributions and community metrics. However, most JSDMs typically do not accommodate imperfect detection (but see Hogg et al., 2021; Tobler et al., 2019). Failure to account for imperfect detection in detection–nondetection data can lead to biases in estimates of both species distributions and the effects of environmental drivers on species occurrence (MacKenzie et al., 2002). Occupancy models, a specific type of SDM, explicitly account for imperfect detection separately from the true species occurrence process using replicated detection–nondetection data. Multi-species occupancy models are an extension to single-species occupancy models that use detection–nondetection data from multiple species by treating species as random effects arising from a community-level distribution (Dorazio & Royle, 2005; Gelfand et al., 2005). Multi-species occupancy models can be viewed as a specific type of JSDM that accommodate imperfect detection, but they traditionally do not include residual co-occurrence associations between species as in other JSDMs that lack imperfect detection (but see Tobler et al., 2019).

Accounting for spatial autocorrelation in SDMs is often necessary when modeling species distributions across large spatial extents or a large number of observed locations (Latimer et al., 2009). Spatially explicit SDMs account for spatial autocorrelation by including spatially structured random effects (Banerjee et al., 2014; Shirota et al., 2019). Such spatially explicit approaches have been used in JSDMs to simultaneously account for residual species correlations and spatial autocorrelation (Thorson et al., 2015), and in multi-species occupancy models that model imperfect detection (Doser et al., 2022).

Despite separate development of JSDMs that account for residual correlations and imperfect detection, only recently have approaches emerged that incorporate both of these complexities in JSDMs for large communities (Hogg et al., 2021; Tobler et al., 2019). Further, these approaches can become computationally intensive as both the number of spatial locations and species in the community increase, and no approaches exist that simultaneously incorporate species correlations, imperfect detection, and spatial autocorrelation, despite the well recognized impacts of ignoring these complexities. Here we develop a JSDM that explicitly accounts for species correlations, imperfect detection, and spatial autocorrelation. Analogous to Tikhonov et al. (2020), we build an ecological process model that uses a spatial factor model together with Nearest Neighbor Gaussian Processes (NNGPs; Datta et al., 2016) to ensure computational efficiency for large species assemblages (e.g., >100 species) across a large number of spatial locations (e.g., $\sim 10^5$). We extend the model of Tikhonov et al. (2020) by incorporating an observation submodel that separately models imperfect detection from the latent ecological process. We use simulations and a case study on 98 bird species across the continental US to compare the performance of our proposed model with five alternative models that fail to address all three complexities. Our proposed modeling framework, and its user-friendly implementation in the *spOccupancy* R package (Doser et al., 2022), provides a computationally efficient approach that explicitly accounts for imperfect detection, residual correlations between species, and spatial autocorrelation to deliver inference on individual species distributions, species co-occurrence patterns, and overall biodiversity metrics.

MODELING FRAMEWORK

Process model

Let $\boldsymbol{s}_j$ denote the spatial coordinates of site j, for all $j = 1,\ldots,J$ sites. Define $z_i(\boldsymbol{s}_j)$ as the true latent presence (1) or absence (0) of species i at site j for $i = 1,\ldots,N$ species. We assume $z_i(\boldsymbol{s}_j)$ arises from a Bernoulli distribution following

$$z_i(\boldsymbol{s}_j) \sim \text{Bernoulli}(\psi_i(\boldsymbol{s}_j)), \quad (1)$$

where $\psi_i(\boldsymbol{s}_j)$ is the probability of occurrence for species i at site j. We model $\psi_i(\boldsymbol{s}_j)$ as

$$\text{logit}(\psi_i(\boldsymbol{s}_j)) = (\beta_{i,1} + w_i^*(\boldsymbol{s}_j)) + \sum_{t=2}^{p_\psi} x_t(\boldsymbol{s}_j)\beta_{i,t}, \quad (2)$$

where $x_t(\boldsymbol{s}_j)$, for each $t = 2,...,p_\psi$, is an environmental covariate at site j, $\beta_{i,t}$ is a regression coefficient corresponding to $x_t(\boldsymbol{s}_j)$ for species i, $\beta_{i,1}$ is the species-specific intercept, and $w_i^*(\boldsymbol{s}_j)$ is a species-specific latent spatial process. We seek to jointly model the species-specific spatial processes to account for residual correlations between species. For a small number of species (e.g., <10), such a process can be estimated via a linear model of coregionalization framework (Gelfand et al., 2004; Latimer et al., 2009). However, when the number of species is even moderately large (e.g., >10), estimating such a joint process becomes computationally intractable. A viable solution to this problem is to use a spatial factor model (Hogan & Tchernis, 2004; Ren & Banerjee, 2013; Zhang & Banerjee, 2021), a dimension reduction approach that can account for correlations among a large number of species. We decompose $w_i^*(\boldsymbol{s}_j)$ into a linear combination of q latent variables (i.e., factors) and their associated species-specific coefficients (i.e., factor loadings). In particular, we have

$$w_i^*(\boldsymbol{s}_j) = \lambda_i^T \mathbf{w}(\boldsymbol{s}_j), \quad (3)$$

where λ_i^T is the ith row of factor loadings from an $N \times q$ loading matrix Λ, and $\mathbf{w}(\boldsymbol{s}_j)$ is a $q \times 1$ vector of independent spatial factors at site j. We achieve computational improvements and dimension reduction by setting $q << N$, where often a small number of factors (e.g., $q = 5$) is sufficient (Taylor-Rodriguez et al., 2019; Zhang & Banerjee, 2021). We account for residual species correlations using individual responses (i.e., loadings) to the q latent spatial factors. Factor loadings explain the occurrence of multiple species at the same location beyond what is explained by the covariates included in the model; co-occurring species will have similar species-specific factor loadings (i.e., they will have the same sign). The residual interspecies covariance matrix $\Sigma = \Lambda\Lambda^T$ has rank $q << N$ and, hence, is singular. Shirota et al. (2019) discuss its use and interpretation in detecting species clustering.

Let $w_r(\boldsymbol{s}_j)$ denote the value of the rth spatial factor at site j, where $r = 1,...,q$. Following Taylor-Rodriguez et al. (2019) and Tikhonov et al. (2020), we model $w_r(\boldsymbol{s}_j)$ using an NNGP (Datta et al., 2016) to achieve computational efficiency when modeling a large number of spatial locations. Thus,

$$w_r(\boldsymbol{s}_j) \sim N\left(\mathbf{0}, \widetilde{\mathbf{C}}_r(\boldsymbol{\theta}_r)\right), \quad (4)$$

where $\widetilde{\mathbf{C}}_r(\boldsymbol{\theta}_r)$ is the NNGP-derived covariance matrix for the rth spatial factor. The vector θ_r consists of parameters governing the spatial process according to a spatial correlation function (Banerjee et al., 2014). For many correlation functions (e.g., exponential, spherical, Gaussian), $\boldsymbol{\theta}_r$ includes a spatial variance parameter, σ_r^2, and a spatial decay parameter, ϕ_r, while the Matérn correlation function includes an additional spatial smoothness parameter, ν_r.

We assume that all species-specific parameters ($\beta_{i,t}$ for all $t = 1,...,p_\psi$) arise from community-level distributions to enable information sharing across species (Dorazio & Royle, 2005; Gelfand et al., 2005). Specifically, we assign a normal prior with mean and variance hyperparameters that represent the community-level average and variance among species-specific effects across the community, respectively. For example, we model the species-specific occurrence intercept, $\beta_{i,1}$, following:

$$\beta_{i,1} \sim N\left(\mu_{\beta_1}, \tau_{\beta_1}^2\right), \quad (5)$$

where μ_{β_1} and $\tau_{\beta_1}^2$ are the community-level average and variance, respectively.

Observation model

To estimate $\psi_i(\boldsymbol{s}_j)$ and $z_i(\boldsymbol{s}_j)$ while explicitly accounting for imperfect detection, we obtain $k = 1,...,K_j$ sampling replicates at each site j. Let $y_{i,k}(\boldsymbol{s}_j)$ denote the detection (1) or nondetection (0) of species i during replicate k at site j. We model the observed data $y_{i,k}(\boldsymbol{s}_j)$ conditional on the true species-specific occurrence $z_i(\boldsymbol{s}_j)$ at site j following

$$y_{i,k}(\boldsymbol{s}_j) \mid z_i(\boldsymbol{s}_j) \sim \text{Bernoulli}(\pi_{i,k}(\boldsymbol{s}_j) z_i(\boldsymbol{s}_j)), \quad (6)$$

where $\pi_{i,k}(\boldsymbol{s}_j)$ is the probability of detecting species i at site j during replicate k given the species is present at the site (i.e., $z_i(\boldsymbol{s}_j) = 1$). Note that when the species is not present at site j (i.e., $z_i(\boldsymbol{s}_j) = 0$), (6) implies $y_{i,k}(\boldsymbol{s}_j) = 0$ (i.e., we assume no false-positive detections). We model $\pi_{i,k}(\boldsymbol{s}_j)$ as a function of site and/or replicate-level covariates that may influence species-specific detection probability. Specifically,

$$\text{logit}(\pi_{i,k}(\boldsymbol{s}_j)) = \alpha_{i,1} + \sum_{t=2}^{p_\pi} v_{t,k}(\boldsymbol{s}_j)\alpha_{i,t}, \quad (7)$$

where $v_{t,k}(\boldsymbol{s}_j)$ is the value of covariate t at site j during replicate k, $\alpha_{i,t}$ is a regression coefficient corresponding to $v_{t,k}(\boldsymbol{s}_j)$, and $\alpha_{i,1}$ is a species-specific intercept.

Analogous to the species-specific occurrence effects (5), we assume all species-specific detection parameters (i.e., $\alpha_{i,t}$ for all $t = 1, \ldots, p_\pi$) arise from community-level normal distributions.

Prior specification and identifiability considerations

We assume normal priors for mean parameters and inverse-Gamma priors for variance parameters. Following Taylor-Rodriguez et al. (2019), we set all elements in the upper triangle of the factor loadings matrix Λ equal to 0 and its diagonal elements equal to 1 to ensure identifiability of the spatial factors. Additionally, we fix the spatial variance parameters σ_r^2 to 1. We assign standard normal priors for elements in Λ below the upper diagonal and assign each spatial decay parameter ϕ_r an independent uniform prior.

Model implementation and alternative models

We implement the spatial factor multi-species occupancy model in a Bayesian framework in the function sfMsPGOcc within our open-source *spOccupancy* R package (Doser et al., 2022). We employ the computational algorithms discussed in Finley et al. (2022) to ensure that spatially explicit models are computationally feasible for large data sets. The Bayesian framework allows us to easily calculate biodiversity metrics, with fully propagated uncertainty, as derived quantities. For example, we can estimate species richness of the entire community (or a subset of species in the community) by summing up the latent occurrence state $z_i(\boldsymbol{s}_j)$ at each site j for all species of interest at each iteration to yield a full posterior distribution for species richness. We use a Pólya-Gamma data augmentation scheme (Polson et al., 2013) to yield an efficient Gibbs sampler (see Appendix S1 for full details).

We compare the spatial factor multi-species occupancy model to five alternative models, each of which addresses a subset of the three complexities (Table 1). We provide functionality for all five alternative models in the *spOccupancy* R package, and subsequently refer to all models by their *spOccupancy* function name (Table 1). Our first alternative model is a nonspatial latent factor JSDM (lfJSDM) that does not account for imperfect detection, analogous to many standard JSDM approaches (Wilkinson et al., 2019). Our second alternative model is a spatial factor JSDM (sfJSDM) that does not account for imperfect detection, similar to the NNGP model of Tikhonov et al. (2020). Our third alternative model is the basic nonspatial multi-species occupancy model (msPGOcc) that does not incorporate residual species correlations (Dorazio & Royle, 2005). Our fourth alternative model is a spatial multi-species occupancy model (spMsPGOcc) that does not incorporate residual species correlations and estimates a separate NNGP spatial process for each species (Doser et al., 2022). Finally, our fifth alternative model is a nonspatial latent factor multi-species occupancy model (lfMsPGOcc) that accounts for residual species correlations and imperfect detection, analogous to the model of Tobler et al. (2019), except we use a logit formulation of the model. See Appendices S1 and S2 for full model details.

TABLE 1 Characteristics of the six models used in the simulation study and case study, as well as the function name for model implementation in the *spOccupancy* R package (Doser et al., 2022).

spOccupancy function	Species correlations	Spatial autocorrelation	Imperfect detection
lfJSDM	✓		
sfJSDM	✓	✓	
msPGOcc			✓
spMsPGOcc		✓	✓
lfMsPGOcc	✓		✓
sfMsPGOcc	✓	✓	✓

Abbreviations: lfJSDM, latent factor joint species distribution model; lfMsPGOcc, latent factor multi-species occupancy model; msPGOcc, multi-species occupancy model; sfJSDM, spatial factor joint species distribution model; sfMsPGOcc, spatial factor multi-species occupancy model; spMsPGOcc, spatial multi-species occupancy model.

SIMULATION STUDY

We used simulations to compare estimates from the spatial factor multi-species occupancy model to estimates from the five alternative models (Table 1). We generated 100 detection–nondetection data sets for each of six simulation scenarios, where the data were simulated with different combinations of the three complexities. We simulated data under situations that roughly corresponded to the six alternative models to assess how each model performed under "ideal" data conditions for that model, as well as when the data did not meet all the assumptions of the modeling framework. More specifically, we generated data with (1) residual species correlations and constant imperfect detection, (2) residual species correlations, constant imperfect detection, and spatial autocorrelation, (3) imperfect detection only, (4) imperfect detection and spatial autocorrelation, (5) residual species correlations and imperfect detection,

and (6) residual species correlations, imperfect detection, and spatial autocorrelation.

We simulated detection–nondetection data from $N = 10$ species at $J = 225$ sites with $K = 3$ replicates at each site for each of the 100 data sets for the six simulation scenarios. We used an exponential correlation function for spatially explicit data generation scenarios (Scenarios 2, 4, 6). For scenarios leveraging a factor model (Scenarios 1, 2, 5, 6), we generated the data using $q = 3$ latent factors. As there are often many potential covariates that explain multi-species occurrence patterns in empirical data sets, we simulated data with 15 spatially varying occurrence covariates for all scenarios and five observational-level detection covariates for scenarios where detection probability was not constant (Scenarios 3–6). We specified reasonable values for all parameters in the model (see Appendix S2 for full details). For each data set in each scenario, we ran three chains each of 15,000 samples, with a burn-in of 10,000 samples and a thinning rate of 5, resulting in a total of 3000 Markov chain Monte Carlo (MCMC) samples for each of the six alternative models. We fit all models using the *spOccupancy* R package (Doser et al., 2022). We assessed the performance of the models by comparing the root mean squared error and 95% coverage rates for the species-specific occurrence probabilities and the occurrence covariate effect.

CASE STUDY

We applied the spatial factor multi-species occupancy model to detection–nondetection data from the North American Breeding Bird Survey (Pardieck et al., 2020) in 2018 on $N = 98$ bird species at $J = 2619$ routes (i.e., sites) across the continental US. The 98 species belong to two distinct biogeographical communities following the definitions in Bateman et al. (2020), with 66 species in the eastern forest bird community and 32 species in the grassland bird community. Our objectives for this case study were to (1) develop spatially explicit maps of species richness for the two communities across the continental US, (2) determine if the latent spatial factors ($\mathbf{w}$) and the species-specific factor loadings (Λ) distinguish the two communities of birds, and (3) assess the benefits of accounting for species correlations, imperfect detection, and spatial autocorrelation. At 50 points along each route (called "stops"), observers performed a 3-min point count survey of all birds seen or heard within a 0.4 km radius. We summarized the data for each species at each site into $K = 5$ spatial replicates (each comprising data from 10 of the 50 stops), where each spatial replicate took value 1 if the species was detected at any of the 10 stops in that replicate, and value 0 if the species was not detected. Using five replicates was more computationally efficient than treating each of the 50 stops as spatial replicates, and exploratory analyses revealed minimal differences between models using the full 50 stop data (Appendix S2).

Using the spatial factor multi-species occupancy model, we modeled the route-level occurrence of the 98 species as a function of five bioclimatic variables and eight land cover variables (Appendix S2). We modeled detection as a function of the day of the survey (linear and quadratic), the start time of the first survey (linear), and a random observer effect. Note that all detection covariates only varied across Breeding Bird Survey (BBS) routes, not across spatial replicates within a route. We standardized all variables to have a mean of 0 and a standard deviation of 1. We fit the model using 15 nearest neighbors, an exponential correlation function, and $q = 5$ latent spatial factors. We subsequently predicted occurrence for the 98 species across the continental US to generate spatially explicit maps of species richness, with associated uncertainty, for the two bird communities.

To determine whether the spatial factor multi-species occupancy model provided benefits for predicting species distributions and biodiversity metrics, we fit four additional alternative models (msPGOcc, lfMsPGOcc, lfJSDM, sfJSDM). For the models that did not explicitly model imperfect detection (lfJSDM and sfJSDM), we collapsed the data with five replicates at each site into a single binary value, which takes value 1 if the species was detected in any of the five replicates and 0 if not. Additionally, because the detection covariates we included in the model only varied by site and not by replicate, we included the detection covariates together with the occurrence covariates in the two JSDMs without a distinct submodel, which is a common approach used to account for sampling variability in models that do not explicitly account for imperfect detection (Ovaskainen et al., 2017). We used the Widely Applicable Information Criterion (WAIC; Watanabe, 2010) to compare the performance of the three occupancy models (msPGOcc, lfMsPGOcc, and sfMsPGOcc) and the two JSDMs without imperfect detection (lfJSDM and sfJSDM). However, as the two JSDMs without imperfect detection used a collapsed form of the data used in the occupancy models, we could not directly compare all five models using WAIC. Thus, we additionally fit all models using 75% of the data points and kept the remaining 25% of the data points for evaluation of model predictive performance. We assessed out-of-sample predictive performance using the observed data at the hold-out locations as well as latent occupancy predictions at the hold-out locations generated from models fit with only the hold-out

locations. See Appendix S2: Section S3 for details. All models were fit using functions in *spOccupancy*. In all cases, model parameter estimates were based on three chains, each with 150,000 iterations, a burn-in period of 100,000 iterations, and a thinning rate of 50. We assessed convergence using visual assessment of trace plots and the Gelman–Rubin (Brooks & Gelman, 1998) diagnostic using the *coda* package (Plummer et al., 2006). See Appendix S3 for a detailed overview and recommendations for convergence assessment using our proposed modeling approach.

RESULTS

Simulation study

Failing to account for residual species correlations had negative impacts on both the accuracy and the precision of model estimates (Table 2; Appendix S2: Tables S1 and S2). Estimates from msPGOcc, which did not account for residual species correlations, had larger bias (Appendix S2: Tables S1 and S2), and low coverage rates (Table 2) for both latent occurrence and covariate effects when data were simulated with residual correlations between species. spMsPGOcc, which accounts for spatial autocorrelation but ignores species correlations, had less bias and better coverage rates than msPGOcc in these scenarios, but still had higher bias in occurrence probabilities and lower coverage rates than models that did account for species correlations and imperfect detection. Therefore, accounting for spatial autocorrelation mitigates some, but not all, of the negative impacts of incorrectly assuming independence between species.

When data were simulated with imperfect detection that varied across sites and replicates, ignoring imperfect detection resulted in higher bias and low coverage rates for both occurrence probability and covariate effects (Table 2; Appendix S2: Tables S1 and S2). However, when detection was high and constant over sites and replicates (Scenarios 1 and 2), bias in lfJSDM and sfJSDM was comparable with models that address imperfect detection and coverage rates were closer to the expected 95%, in particular for the latent occurrence probability (Appendix S2: Tables S1 and S2). Notably, the decreased coverage rates were less drastic for estimating occurrence probability when failing to account for imperfect detection compared with estimates from a standard multi-species occupancy

TABLE 2 Estimated coverage rates of simulated species-specific occurrence probabilities and covariate effects for six different simulation scenarios and six models of varying complexity, as well as average run time.

		Model					
Parameter	**Scenario**	**lfJSDM**	**sfJSDM**	**msPGOcc**	**spMsPGOcc**	**lfMsPGOcc**	**sfMsPGOcc**
$\psi_i(\boldsymbol{s}_j)$	1	91.5	90.8	68.9	88.1	95.6	95.3
	2	91.6	91.0	69.1	89.1	95.5	95.4
	3	85.6	84.8	95.0	96.4	95.5	95.5
	4	77.5	76.4	80.2	93.1	95.7	95.5
	5	75.3	74.2	71.3	88.5	95.5	95.3
	6	76.0	75.0	72.2	89.6	95.3	95.2
β_i	1	88.7	88.2	82.0	91.1	95.2	95.1
	2	88.8	88.2	82.2	91.7	94.9	94.9
	3	73.8	73.1	95.1	94.4	90.4	90.8
	4	65.9	65.0	89.1	94.0	94.7	94.7
	5	64.2	63.6	83.6	91.7	95.2	95.0
	6	65.7	64.6	85.1	92.7	94.9	94.9
Run time		1.55	3.17	3.00	6.17	3.31	5.24

Note: Coverage rates are defined as the percentage of species-specific occurrence probabilities ($\psi_i(\boldsymbol{s}_j)$) or covariate effects contained within the 95% credible interval, averaged across the 10 species and 100 simulated data sets. Run time is the number of minutes for the model to complete 15,000 Markov chain Monte Carlo (MCMC) iterations, averaged across all six simulation scenarios and 100 simulated data sets. Data were generated with the following characteristics for the six simulation scenarios: (1) residual species correlations and constant, high detection; (2) residual species correlations, constant and high detection, spatial autocorrelation; (3) imperfect detection; (4) imperfect detection and spatial autocorrelation; (5) residual species correlations and imperfect detection; and (6) residual species correlations, imperfect detection, and spatial autocorrelation.

Abbreviations: lfJSDM, latent factor joint species distribution model; lfMsPGOcc, latent factor multi-species occupancy model; msPGOcc, multi-species occupancy model; sfJSDM, spatial factor joint species distribution model; sfMsPGOcc, spatial factor multi-species occupancy model; spMsPGOcc, spatial multi-species occupancy model.

model (msPGOcc) when ignoring residual correlations when present. Alternatively, failing to account for imperfect detection when present resulted in larger bias and lower coverage rates in occurrence covariate effect estimates compared with a model that ignored residual correlations and/or spatial autocorrelation when present. Ignoring spatial autocorrelation had minimal impacts on the average bias, particularly when accounting for residual correlations, but coverage rates were substantially low for both latent occurrence and the covariate effect for msPGOcc (Table 2).

Case study

The spatial factor multi-species occupancy model predicted high species richness for the eastern forest bird community across the eastern US and high species richness for the grassland bird community in the Northern Great Plains region (Figure 1). Further, the model distinguished between the two bird communities via the species-specific factor loadings and the spatial factors (Appendix S2: Figures S1–S5). Compared with the standard multi-species occupancy model (msPGOcc), incorporating residual species correlations (lfMsPGOcc) yielded a lower WAIC (417,954 vs. 395,094), while additionally accounting for spatial autocorrelation (sfMsPGOcc) further reduced the WAIC (390,607; Appendix S2: Table S3). Failing to account for spatial autocorrelation led to unreasonable species richness estimates for the two communities across large portions of the US (Figure 2A,B). Additionally, the spatially explicit JSDM without imperfect detection (sfJSDM) outperformed the nonspatial JSDM without imperfect detection (lfJSDM) according to WAIC (84,192 vs. 87,615).

Analogous to model comparison using WAIC, the two models that accounted for spatial autocorrelation (sfJSDM and sfMsPGOcc) had the smallest out-of-sample model deviance, with sfJSDM outperforming sfMsPGOcc when assessing performance based on the raw detection–nondetection data. However, when estimating predictive performance using estimates of species occurrence generated from three occupancy model fits, sfMsPGOcc

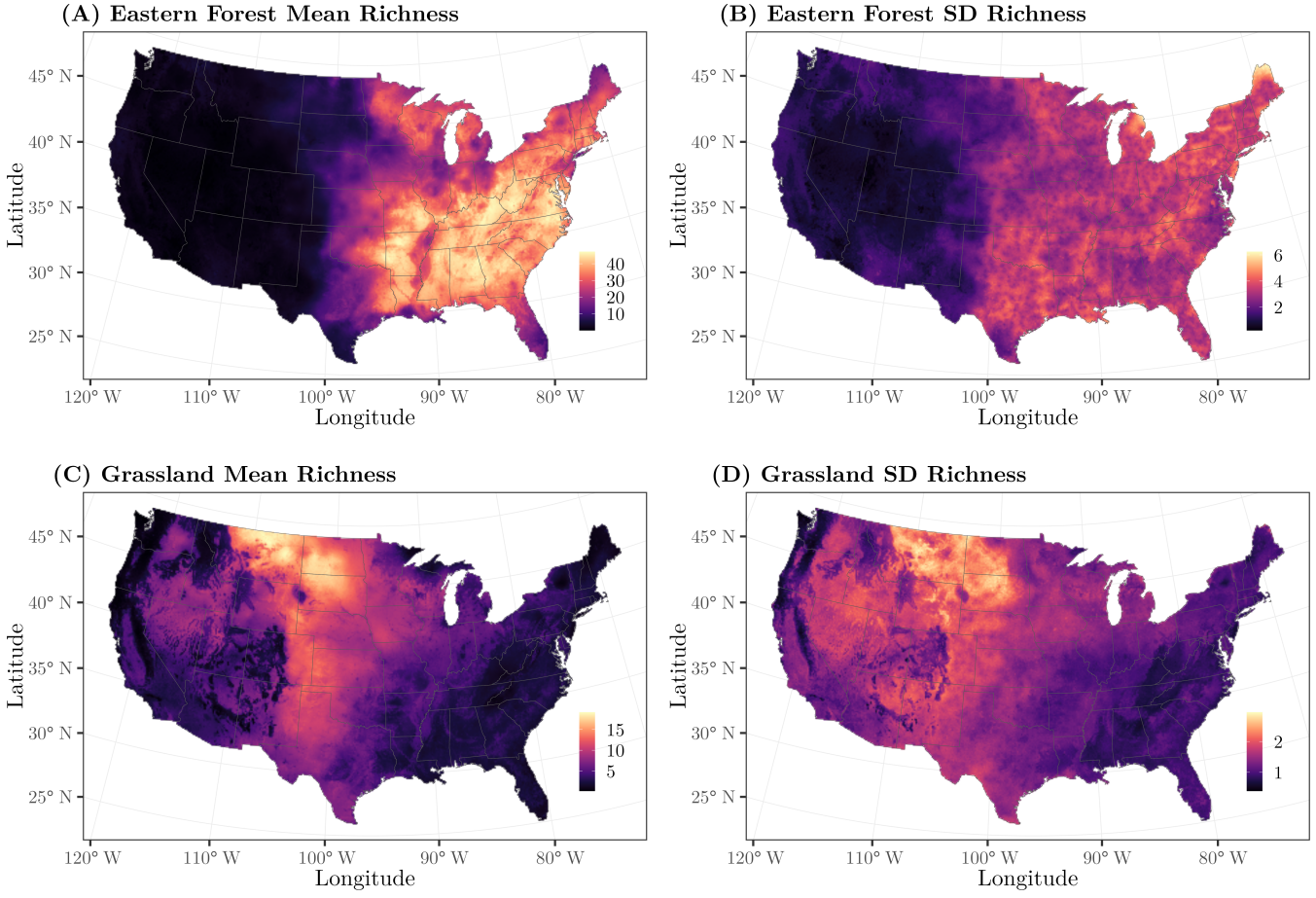

FIGURE 1 Predicted mean species richness for the eastern forest bird community (A) and the grassland bird community (C), as well as their associated standard deviations (B, D) using a spatial latent factor multi-species occupancy model (sfMsPGOcc).

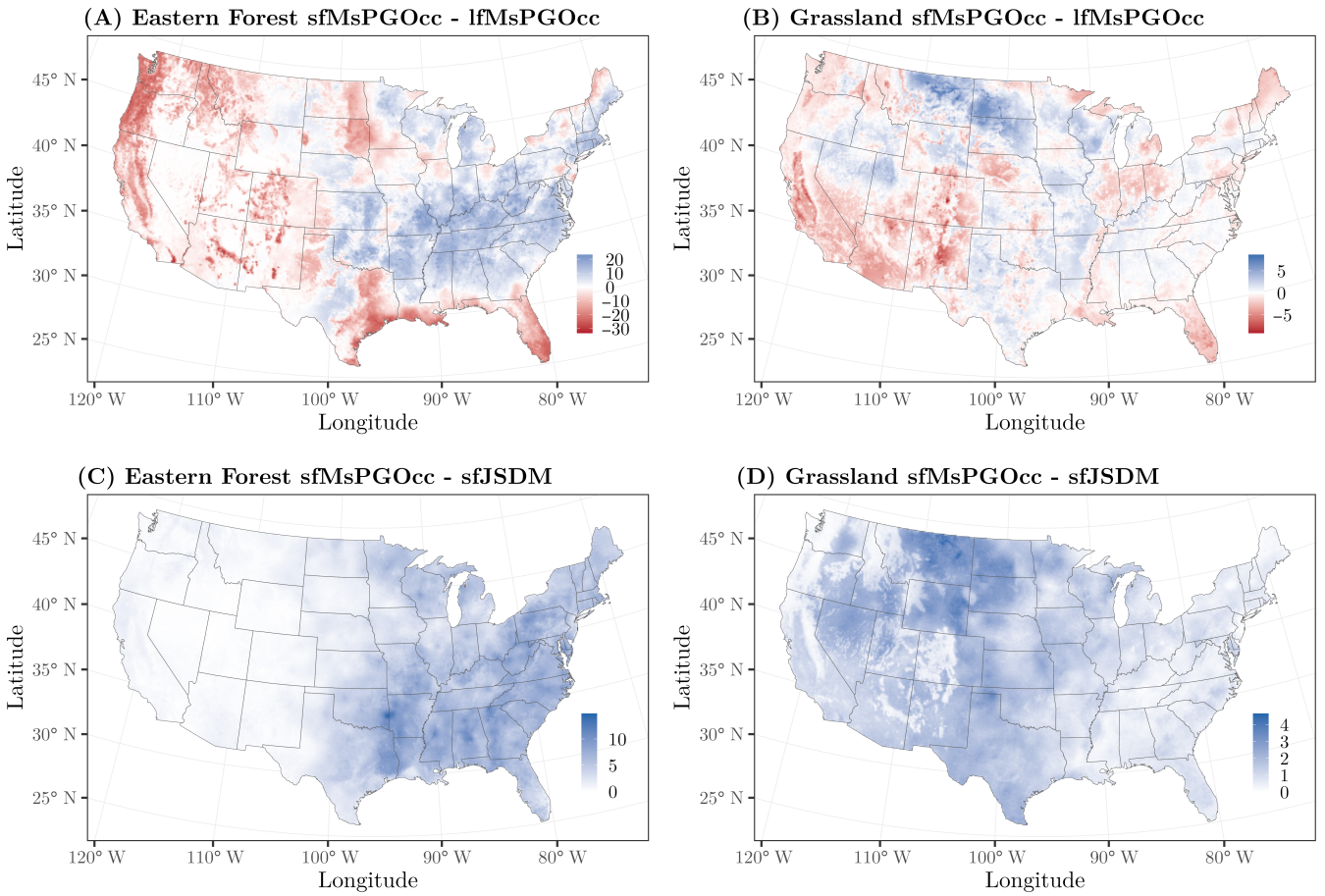

FIGURE 2 Difference in predicted mean richness between a spatial latent factor multi-species occupancy model (sfMsPGOcc) and two simpler alternative models. Panels (A) and (B) show differences between the nonspatial latent factor multi-species occupancy model (lfMsPGOcc) for the eastern forest and grassland bird communities, respectively, while panels (C) and (D) show differences with the spatial factor joint species distribution model without imperfect detection (sfJSDM).

outperformed sfJSDM (Appendix S2: Table S3), suggesting that accounting for imperfect detection provides improved predictive performance of the latent ecological process. Further, estimates of species richness from sfJSDM were substantially lower across the US for both the eastern forest and grassland bird community (Figure 2C,D) compared with estimates from sfMsPGOcc.

DISCUSSION

Multi-species detection–nondetection data are often complicated by residual correlations among species detections (Ovaskainen et al., 2010), imperfect detection of species (MacKenzie et al., 2002), and spatial autocorrelation (Latimer et al., 2009). While many methods exist to accommodate a subset of these complexities (e.g., Tikhonov et al., 2020; Tobler et al., 2019), no approaches exist that simultaneously incorporate all three complexities, despite the well recognized impacts of ignoring them. Here, we developed a spatial factor multi-species occupancy model that simultaneously accounts for residual species correlations, imperfect detection, and spatial autocorrelation in a computationally efficient framework. We showed using simulations that ignoring these three complexities when present leads to inferior inference and prediction. Further, the spatial factor multi-species occupancy model improved predictive performance compared to models that failed to address the three complexities in an empirical case study of 98 bird species across the continental US.

In our simulation study, failing to account for residual species correlations, imperfect detection, and/or spatial autocorrelation when present led to increased bias and low coverage rates. We found that the standard multi-species occupancy model (msPGOcc) had high bias and low coverage rates for both the latent occurrence and occurrence covariate effects for all scenarios except when data were simulated without

species correlations and spatial autocorrelation (Table 2, Appendix S2: Tables S1 and S2), clearly indicating the importance of accommodating these data complexities if they exist. Similarly, estimates from JSDMs that failed to account for imperfect detection resulted in increased bias and low coverage rates, although these findings were less prominent under ideal scenarios of constant, high detection probability. Interestingly, Table 2 suggests that if it is not possible to accommodate all three complexities (e.g., because of limited resources and small sample sizes) determining which complexities to ignore will depend on the study objectives. For example, when data were simulated with imperfect detection and species correlations, coverage rates were better for lfJSDM than msPGOcc for the occurrence probability estimates, but coverage rates from msPGOcc were better than lfJSDM for the occurrence covariate effect. This suggests that under these scenarios, lfJSDM would be better for prediction, while msPGOcc would be better for inference. Our simulation study did not consider all potential complexities when comparing the performance of JSDMs, such as differing degrees of residual species correlations versus spatial autocorrelation or assessment of the sensitivity of model performance to more complex forms of spatial dependence (e.g., Mohankumar & Hefley, 2022). However, our results do illustrate that specific data characteristics and research questions will determine whether it is necessary to account for residual species correlations, imperfect detection, and/or spatial autocorrelation. Our findings, as well as additional simulation studies geared toward specific ecological scenarios, could have important implications for designing detection–nondetection surveys to meet specific objectives. We include code to fit all six alternative models (Table 1) in the *spOccupancy* R package, as well as functions for data simulation and model comparison to enable ecologists and conservation practitioners to accommodate these three complexities using accessible and well documented software. See Appendix S4 for a detailed vignette on fitting these models in *spOccupancy*.

In the breeding bird case study, accounting for species correlations, imperfect detection, and spatial autocorrelation in the spatial factor multi-species occupancy model resulted in improved predictive performance compared with models that failed to address all three complexities. Accounting for species correlations in lfMsPGOcc improved model fit over the standard multi-species occupancy model (msPGOcc) according to WAIC but did not improve predictive performance for the out-of-sample deviance metric using the raw data (Appendix S2: Table S3). This is likely a result of treating the latent factors as independent standard normal random variables, which results in predictions that are not able to use the estimated values of the latent variables at nearby sampled locations to improve prediction at nonsampled locations. Alternatively, the spatial factor multi-species occupancy model (sfMsPGOcc) had the smallest WAIC and the best predictive performance for both deviance metrics among the three occupancy models. Further, sfJSDM substantially outperformed lfJSDM according to all criteria. These results demonstrate how assigning spatial structure to the latent factors in a model that accounts for species correlations can yield large improvements in model predictive performance. We thus recommend using sfMsPGOcc when there is a desire to account for species correlations and the primary goal of the analysis is prediction.

The spatial factor multi-species occupancy model leverages a spatial factor dimension reduction approach (Hogan & Tchernis, 2004; Ren & Banerjee, 2013; Zhang & Banerjee, 2021) and NNGPs (Datta et al., 2016) to ensure computational efficiency when modeling data sets with a large number of species (e.g., >100) and/or spatial locations (e.g., 100,000). Our proposed model requires the specification of the number of latent spatial factors (q) as well as the number of neighbors to use in the NNGP. When choosing the number of nearest neighbors for the NNGP, Datta et al. (2016) showed 15 neighbors is sufficient for most data sets, with as few as five neighbors providing adequate performance for certain data sets. Determining the optimal number of factors for a given data set is not straightforward and will vary depending on the characteristics of the specific community of species (e.g., species rarity, variability among species). See Appendix S3 for recommendations and considerations for making this decision, as well as a discussion on assessing the convergence of these high-dimensional models.

The use of spatial replicates in the BBS case study instead of the more traditional temporal replicates used in an occupancy modeling framework may lead to an upward bias in the estimated occupancy probabilities (Kendall & White, 2009). Additionally, the large spatial scale of the BBS data (each route is ~39.2 km in length) likely influences the estimates of the residual species co-occurrence patterns. Data collected at a smaller spatial scale using temporal replicates may provide more accurate estimates of occupancy and species co-occurrence patterns. Regardless of how the data are collected, we caution against the interpretation of the residual co-occurrences as true biological interactions, as co-occurrence does not imply an interaction (Poggiato et al., 2021).

The latent spatial factors and the species-specific factor loadings can provide insight into the additional processes that govern the distributions of species in the modeled community. In our case study, we found

the spatial factors showed clear distinctions between the two bird communities. See Appendix S2 for additional discussion on interpreting the latent factors and Appendices S3 and S4 for practical information on how to troubleshoot MCMC convergence problems with the factor loadings.

As both the number and size of multi-species detection–nondetection data sets increase, we require computationally efficient models and software to address common data complexities. Our spatial factor multi-species occupancy model extends previous approaches (Tikhonov et al., 2020; Tobler et al., 2019) to efficiently model species-specific and community-level occurrence patterns while accounting for residual species correlations, imperfect detection, and spatial autocorrelation. Our proposed framework, together with its user-friendly implementation in the *spOccupancy* R package (Doser et al., 2022), will enable ecologists to study spatial variation in species occurrence and co-occurrence patterns, develop spatially explicit maps of individual species distributions and biodiversity metrics, and explicitly account for common complexities in multi-species detection–nondetection data.

ACKNOWLEDGMENTS

We thank Viviana Ruiz-Gutierrez, Narmadha Mohankumar, and two anonymous reviewers for insightful comments that improved the manuscript. This work was supported by National Science Foundation (NSF) grants DEB-2213565, EF-1253225, and DMS-1916395.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The package *spOccupancy* is available on the Comprehensive R Archive Network (CRAN; https://cran.r-project.org/web/packages/spOccupancy/index.html). Data and code used in the manuscript (Doser et al., 2023) are available on Zenodo: https://zenodo.org/record/8037367.

ORCID

Jeffrey W. Doser https://orcid.org/0000-0002-8950-9895

REFERENCES

Banerjee, S., B. P. Carlin, and A. E. Gelfand. 2014. *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton: CRC Press.

Bateman, B. L., C. Wilsey, L. Taylor, J. Wu, G. S. LeBaron, and G. Langham. 2020. "North American Birds Require Mitigation and Adaptation to Reduce Vulnerability to Climate Change." *Conservation Science and Practice* 2(8): e242.

Brooks, S. P., and A. Gelman. 1998. "General Methods for Monitoring Convergence of Iterative Simulations." *Journal of Computational and Graphical Statistics* 7(4): 434–455.

Datta, A., S. Banerjee, A. O. Finley, and A. E. Gelfand. 2016. "Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets." *Journal of the American Statistical Association* 111(514): 800–812.

Dorazio, R. M., and J. A. Royle. 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., A. O. Finley, and S. Banerjee. 2023. "Code and Data for Joint Species Distribution Models with Imperfect Detection for High-Dimensional Spatial Data." Zenodo. https://zenodo.org/record/8037367.

Doser, J. W., A. O. Finley, M. Kéry, and E. F. Zipkin. 2022. "spOccupancy: An R Package for Single-Species, Multi-Species, and Integrated Spatial Occupancy Models." *Methods in Ecology and Evolution* 13(8): 1670–78.

Finley, A. O., A. Datta, and S. Banerjee. 2022. "spNNGP R Package for Nearest Neighbor Gaussian Process Models." *Journal of Statistical Software* 103(5): 1–40.

Gelfand, A. E., A. M. Schmidt, S. Banerjee, and C. F. Sirmans. 2004. "Nonstationary Multivariate Process Modeling through Spatially Varying Coregionalization." *Test* 13(2): 263–312.

Gelfand, A. E., A. M. Schmidt, S. Wu, J. A. Silander, Jr., A. Latimer, and A. G. Rebelo. 2005. "Modelling Species Diversity through Species Level Hierarchical Modelling." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54(1): 1–20.

Guisan, A., and N. E. Zimmermann. 2000. "Predictive Habitat Distribution Models in Ecology." *Ecological Modelling* 135(2-3): 147–186.

Hogan, J. W., and R. Tchernis. 2004. "Bayesian Factor Analysis for Spatially Correlated Data, with Application to Summarizing Area-Level Material Deprivation from Census Data." *Journal of the American Statistical Association* 99(466): 314–324.

Hogg, S. E., Y. Wang, and L. Stone. 2021. "Effectiveness of Joint Species Distribution Models in the Presence of Imperfect Detection." *Methods in Ecology and Evolution* 12(8): 1458–74.

Kendall, W. L., and G. C. White. 2009. "A Cautionary Note on Substituting Spatial Subunits for Repeated Temporal Sampling in Studies of Site Occupancy." *Journal of Applied Ecology* 46(6): 1182–88.

Latimer, A., S. Banerjee, H. Sang, Jr., E. Mosher, and J. Silander, Jr. 2009. "Hierarchical Models Facilitate Spatial Analysis of Large Data Sets: A Case Study on Invasive Plant Species in the Northeastern United States." *Ecology Letters* 12(2): 144–154.

MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. 2002. "Estimating Site Occupancy Rates when Detection Probabilities Are Less than One." *Ecology* 83(8): 2248–55.

Mohankumar, N. M., and T. J. Hefley. 2022. "Using Machine Learning to Model Nontraditional Spatial Dependence in Occupancy Data." *Ecology* 103(2): e03563.

Ovaskainen, O., J. Hottola, and J. Siitonen. 2010. "Modeling Species Co-occurrence by Multivariate Logistic Regression Generates New Hypotheses on Fungal Interactions." *Ecology* 91(9): 2514–21.

Ovaskainen, O., G. Tikhonov, A. Norberg, F. Guillaume Blanchet, L. Duan, D. Dunson, T. Roslin, and N. Abrego. 2017. "How to Make More out of Community Data? A Conceptual Framework and its Implementation as Models and Software." *Ecology Letters* 20(5): 561–576.

Pardieck, K., D. Ziolkowski, Jr., M. Lutmerding, V. Aponte, and M.-A. Hudson. 2020. "North American Breeding Bird Survey Dataset 1966–2019." U.S. Geological Survey Data Release. https://doi.org/10.5066/P9J6QUF6.

Plummer, M., N. Best, K. Cowles, and K. Vines. 2006. "CODA: Convergence Diagnosis and Output Analysis for MCMC." *R News* 6(1): 7–11.

Poggiato, G., T. Münkemüller, D. Bystrova, J. Arbel, J. S. Clark, and W. Thuiller. 2021. "On the Interpretations of Joint Modeling in Community Ecology." *Trends in Ecology & Evolution* 36(5): 391–401.

Polson, N. G., J. G. Scott, and J. Windle. 2013. "Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables." *Journal of the American Statistical Association* 108(504): 1339–49.

Ren, Q., and S. Banerjee. 2013. "Hierarchical Factor Models for Large Spatially Misaligned Data: A Low-Rank Predictive Process Approach." *Biometrics* 69(1): 19–30.

Shirota, S., A. Gelfand, and S. Banerjee. 2019. "Spatial Joint Species Distribution Modeling Using Dirichlet Processes." *Statistica Sinica* 29: 1127–54.

Taylor-Rodriguez, D., A. O. Finley, A. Datta, C. Babcock, H.-E. Andersen, B. D. Cook, D. C. Morton, and S. Banerjee. 2019. "Spatial Factor Models for High-Dimensional and Large Spatial Data: An Application in Forest Variable Mapping." *Statistica Sinica* 29: 1155–80.

Thorson, J. T., M. D. Scheuerell, A. O. Shelton, K. E. See, H. J. Skaug, and K. Kristensen. 2015. "Spatial Factor Analysis: A New Tool for Estimating Joint Species Distributions and Correlations in Species Range." *Methods in Ecology and Evolution* 6(6): 627–637.

Tikhonov, G., L. Duan, N. Abrego, G. Newell, M. White, D. Dunson, and O. Ovaskainen. 2020. "Computationally Efficient Joint Species Distribution Modeling of Big Spatial Data." *Ecology* 101(2): e02929.

Tobler, M. W., M. Kéry, F. K. Hui, G. Guillera-Arroita, P. Knaus, and T. Sattler. 2019. "Joint Species Distribution Models with Species Correlations and Imperfect Detection." *Ecology* 100(8): e02754.

Watanabe, S. 2010. "Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory." *Journal of Machine Learning Research* 11(12): 3571–94.

Wilkinson, D. P., N. Golding, G. Guillera-Arroita, R. Tingley, and M. A. McCarthy. 2019. "A Comparison of Joint Species Distribution Models for Presence–Absence Data." *Methods in Ecology and Evolution* 10(2): 198–211.

Zhang, L., and S. Banerjee. 2021. "Spatial Factor Modeling: A Bayesian Matrix-Normal Approach for Misaligned Data." *Biometrics* 78(2): 560–573.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Doser, Jeffrey W., Andrew O. Finley, and Sudipto Banerjee. 2023. "Joint Species Distribution Models with Imperfect Detection for High-Dimensional Spatial Data." *Ecology* e4137. https://doi.org/10.1002/ecy.4137