APPLICATION

Methods in Ecology and Evolution | BRITISH ECOLOGICAL SOCIETY

`spOccupancy`: An R package for single-species, multi-species, and integrated spatial occupancy models

Jeffrey W. Doser[1,2] | Andrew O. Finley[1,2] | Marc Kéry[3] | Elise F. Zipkin[2,4]

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

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

[3]Swiss Ornithological Institute, Sempach, Switzerland

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

Correspondence
Jeffrey W. Doser

Email: doserjef@msu.edu

Abstract

1. Occupancy modelling is a common approach to assess species distribution patterns, while explicitly accounting for false absences in detection–nondetection data. Numerous extensions of the basic single-species occupancy model exist to model multiple species, spatial autocorrelation and to integrate multiple data types. However, development of specialized and computationally efficient software to incorporate such extensions, especially for large datasets, is scarce or absent.

2. We introduce the `spOccupancy` R package designed to fit single-species and multi-species spatially explicit occupancy models. We fit all models within a Bayesian framework using Pólya-Gamma data augmentation, which results in fast and efficient inference. `spOccupancy` provides functionality for data integration of multiple single-species detection–nondetection datasets via a joint likelihood framework. The package leverages Nearest Neighbour Gaussian Processes to account for spatial autocorrelation, which enables spatially explicit occupancy modelling for potentially massive datasets (e.g. 1,000s–100,000s of sites).

3. `spOccupancy` provides user-friendly functions for data simulation, model fitting, model validation (by posterior predictive checks), model comparison (using information criteria and k-fold cross-validation) and out-of-sample prediction. We illustrate the package's functionality via a vignette, simulated data analysis and two bird case studies.

4. The `spOccupancy` package provides a user-friendly platform to fit a variety of single and multi-species occupancy models, making it straightforward to address detection biases and spatial autocorrelation in species distribution models even for large datasets.

KEYWORDS
Bayesian, data fusion, data integration, hierarchical model, imperfect detection, MCMC, occupancy model, spatial autocorrelation

1 | INTRODUCTION

Understanding the processes that drive species distributions across space and time is a fundamental objective in ecology (Pulliam, 2000). Species distribution models (SDMs) are the primary tool used to study the spatial distributions of both individual species and entire communities (Guisan & Zimmermann, 2000). Imperfect detection, or the failure to observe a species during sampling where it is in fact present, is a ubiquitous complication that must be addressed when modelling species distributions (MacKenzie et al., 2002; Tyre et al., 2003). Single-species occupancy models (SSOMs), a specialized type of SDM, explicitly incorporate the detection process separately from the latent species occurrence process using replicated detection–nondetection data (MacKenzie et al., 2002; Tyre et al., 2003). Multi-species occupancy models (MSOMs) are an extension to SSOMs that leverage detection–nondetection data from multiple species (Dorazio & Royle, 2005; Gelfand et al., 2005). This approach views species-specific parameters as random effects arising from a common community-level distribution, which enables inferences at multiple scales (species, community) and leads to greater precision of species-specific effects and biodiversity metrics, all with fully propagated uncertainty (Dorazio & Royle, 2005).

While SSOMs and MSOMs were developed for use with only a single detection–nondetection dataset, there is increasing interest in combining multiple data sources within a single statistical model to improve species distribution inferences (Isaac et al., 2020; Chapter 10 in Kéry & Royle, 2021). Data integration (also referred to as data fusion) is a model-based approach that combines multiple data sources and accommodates different sampling processes among data sources (Miller et al., 2019). Integrated occupancy models (IOMs) simultaneously analyse multiple detection–nondetection data sources to better estimate the latent process of interest and sources of uncertainty. These models are particularly attractive as there are numerous possible protocols to obtain detection–nondetection data sources across potentially vast spatial regions, such as autonomous recording units (Doser et al., 2021).

As detection–nondetection data sources increase in both spatial extent and number of observed locations, accounting for spatial autocorrelation becomes increasingly more important (Guélat & Kéry, 2018). Accommodating sources of spatial dependency among observations is key to delivering valid inferences about species distributions and has led to the development of spatial occupancy models (Johnson et al., 2013). Modelling spatial dependence via spatially structured random effects can improve predictive performance in occurrence probabilities across a region of interest (Wright et al., 2021). Including such random effects, however, is notoriously computationally expensive and can easily lead to intolerable software run times when the number of locations used to define the spatial random effects becomes even moderately large (e.g. 100s–1,000s of locations). This is particularly true when modelling detection–nondetection data as point locations in continuous space (i.e. in point-referenced spatial regression models) rather than as discrete units on a gridded study area (i.e. in areal spatial regression

models; Banerjee et al., 2003), as computational complexity increases in cubic order with the number of spatial locations. This so-called 'big N' problem (Banerjee & Fuentes, 2012) renders even moderately large datasets computationally infeasible using common Bayesian software packages such as `Stan` (Carpenter et al., 2017), `JAGS` (Plummer, 2003) and `NIMBLE` (de Valpine et al., 2017).

A paucity of user-friendly and computationally efficient software has so far limited adoption of spatial occupancy models in practice, as models may take weeks to run or simply not be possible to fit at all. In the context of SSOMs, `MARK` (White & Burnham, 1999), `PRESENCE` (Hines, 2006) and the `R` package `unmarked` (Fiske & Chandler, 2011) fit a variety of models for wildlife data using likelihood inference, but lack functionality to account for spatial autocorrelation. The `R` packages `stocc` (Johnson et al., 2013), `hSDM` (Vieilledent, 2019), `Rcppocc` (Clark & Altwegg, 2019) and `ubms` (Kellner et al., 2021) all fit spatial SSOMs using areal spatial models. However, these packages cannot fit MSOMs and may not be adequate for large datasets (e.g. thousands of locations). In a multi-species framework, the `R` package `HMSC` (Tikhonov et al., 2020) fits a wide range of spatially explicit joint species distribution models allowing for estimation of complex correlations among species occurrence patterns, but this package fails to account for imperfect detection. Additionally, none of these packages facilitate model-based data integration, which together leads many practitioners to use Bayesian programming languages such as `JAGS`, `NIMBLE` and `Stan` to fit SSOMs, MSOMs or IOMs. While these programming languages are incredibly flexible for defining specialized models, they are often not optimized for efficient computation involving dense covariance matrices and may pose a steep learning curve for novice users.

Here we present `spOccupancy`, an `R` package that fits SSOMs, MSOMs and IOMs that can accommodate spatial autocorrelation in potentially massive datasets. We fit spatial SSOMs, MSOMs and IOMs in a point-referenced framework using either Gaussian processes or Nearest Neighbour Gaussian Processes (NNGPs; Datta et al., 2016; Finley et al., 2019). NNGPs use local information from a small set of nearest neighbours to closely approximate a full Gaussian process while drastically reducing computational complexity (Finley et al., 2019), which allows `spOccupancy` to fit computationally efficient models even for large spatial datasets. In this paper, we review SSOMs, MSOMs and IOMs, describe the functionality of `spOccupancy`, and illustrate its features using two case studies on breeding birds in the United States.

2 | MODELS

`spOccupancy` fits nonspatial and spatial SSOMs, MSOMs and IOMs (six different model structures; see Table 1 for function names and descriptions). All models are fit in a Bayesian framework with priors and numerical algorithms implemented to maximize computational efficiency. See Supplemental Information S1.3 for a detailed discussion on the prior distributions and their default values in `spOccupancy`.

Functionality	Description
Data simulation	
simOcc	Simulate single-species occupancy data
simMsOcc	Simulate multi-species occupancy data
simIntOcc	Simulate single-species occupancy data from multiple data sources
Model fitting	
PGOcc	Single-species occupancy model
spPGOcc	Single-species spatial occupancy model
intPGOcc	Single-species occupancy model with multiple data sources
spIntPGOcc	Single-species spatial occupancy model with multiple data sources
msPGOcc	Multi-species occupancy model
spMsPGOcc	Multi-species spatial occupancy model
Model assessment	
ppcOcc	Posterior predictive check using Bayesian p-values
waicOcc	Compute Widely Applicable Information Criterion

2.1 | Single-species occupancy models (SSOMs)

Let z_j denote the true presence (1) or absence (0) of a species at site $j = 1, \ldots, J$. The SSOM assumes z_j arises from a Bernoulli process following

$$z_j \sim \text{Bernoulli}(\psi_j), \tag{1}$$

where ψ_j is the probability of occurrence at site j (MacKenzie et al., 2002; Tyre et al., 2003). We model ψ_j using a logit link following

$$\text{logit}(\psi_j) = \mathbf{x}_j^\top \boldsymbol{\beta}, \tag{2}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients (including an intercept) that describe the effect of covariates $\mathbf{x}_j$ and the $\top$ denotes transposition of column vector $\mathbf{x}_j$.

To estimate ψ_j while accounting for imperfect detection, $k = 1, \ldots, K_j$ sampling replicates are obtained at each site j. We model the observed detection (1) or nondetection (0) of a study species during replicate visit k at site j, denoted $y_{j,k}$, conditional on the true occupancy process, z_j, following

$$y_{j,k} \sim \text{Bernoulli}(p_{j,k} z_j), \tag{3}$$

where $p_{j,k}$ is the probability of detecting the species at site j during visit k. Detection probability can vary by site and/or sampling covariates following

$$\text{logit}(p_{j,k}) = \mathbf{v}_{j,k}^\top \boldsymbol{\alpha}, \tag{4}$$

where $\boldsymbol{\alpha}$ is a vector of regression coefficients (including an intercept) that describe the effect of site and/or observation covariates $\mathbf{v}_{j,k}$ on detection.

We complete the Bayesian specification of the model by assigning Gaussian priors to the occurrence ($\boldsymbol{\beta}$) and detection ($\boldsymbol{\alpha}$) regression coefficients (including the intercepts). See Supplemental Information S2.2 for further model details. The `spOccupancy` function `PGOcc` fits SSOMs.

2.2 | Spatial single-species occupancy models (SSOMs)

We extend the previous SSOM to account for residual spatial variation in species occurrence. Let $\mathbf{s}_j$ denote the geographical coordinates of site j for $j = 1, \ldots, J$. In the spatial SSOM, we include $\mathbf{s}_j$ directly in the notation of spatially indexed variables to indicate the model is spatially explicit. More specifically, the occurrence probability at site j with coordinates $\mathbf{s}_j$, $\psi(\mathbf{s}_j)$, now takes the form

$$\text{logit}(\psi(\mathbf{s}_j)) = \mathbf{x}(\mathbf{s}_j)^\top \boldsymbol{\beta} + \mathbf{w}(\mathbf{s}_j), \tag{5}$$

where $\mathbf{w}(\mathbf{s}_j)$ is a realization from a zero-mean spatial Gaussian process. In particular, we assume that

$$\mathbf{w}(\mathbf{s}) \sim N(\mathbf{0}, \Sigma(\mathbf{s}, \mathbf{s}', \theta)), \tag{6}$$

where $\Sigma(\mathbf{s}, \mathbf{s}', \theta)$ is a $J \times J$ covariance matrix that is a function of the distances between any pair of site coordinates $\mathbf{s}$ and $\mathbf{s}'$ and a set of parameters (θ) that govern the spatial process according to a spatial correlation function. `spOccupancy` supports four spatial correlation functions: exponential, spherical, Gaussian and Matérn (see Chapters 1 and 3 in Banerjee et al. (2003) for correlation function details). For the exponential, spherical and Gaussian functions, $\theta = \{\sigma^2, \phi\}$, where σ^2 is the spatial variance parameter and ϕ is a spatial decay parameter, while the Matérn specification additionally includes a spatial smoothness parameter, ν. As a result of the additional smoothness parameter, the Matérn correlation function is the most flexible out of the four, but using this function may require comparatively more data. Generally, these functions can all adequately accommodate spatial autocorrelation, and we recommend using the widely applicable information criterion (WAIC; Watanabe, 2010) and k-fold cross-validation to select among different correlation functions (see Section 4). We assign an inverse-Gamma prior to the spatial variance parameter (σ^2) and uniform priors to the spatial decay (ϕ) and smoothness (ν) parameters. The remainder of the model follows the nonspatial SSOM (Supplemental Information S2.3). The `spOccupancy` function `spPGOcc` fits spatial SSOMs.

2.3 | Multi-species occupancy models (MSOMs)

As an extension to the SSOM, consider a dataset $y_{i,j,k}$ denoting the detection or nondetection of species i at site j during replicate k for

$i = 1, \ldots, N$ species. The ecological process of interest is $z_{i,j}$, the true presence or absence status of species i at site j. Following Dorazio and Royle (2005), $z_{i,j}$ and $y_{i,j,k}$ are modelled according to Equations 1 and 3 with all detection and occupancy parameters now varying by species. Species-specific regression coefficients for occupancy (β_i) and detection (α_i) are treated as random effects arising from community-level normal distributions, which leads to greater precision of species-specific effects (particularly for rare species) and facilitates estimation of biodiversity metrics (Zipkin et al., 2009). For example, the species-specific occurrence intercept, $\beta 0_i$, is modelled according to

$$\beta 0_i \sim \text{Normal}\left(\mu_{\beta 0}, \tau_{\beta 0}^2\right), \tag{7}$$

where $\mu_{\beta 0}$ is the community-level occurrence intercept and $\tau_{\beta 0}^2$ is the variance of the intercept among species in the community. Models for all parameters in β_i and α_i are defined analogously. We assign normal priors to the community-level occurrence (μ_β) and detection (μ_α) regression coefficients, and inverse-Gamma priors to each of the occurrence (τ_β^2) and detection (τ_α^2) variance parameters (Supplemental Information S2.4). The spOccupancy function msPGOcc fits MSOMs.

2.4 | Spatial multi-species occupancy models (MSOMs)

The spatial MSOM is identical to the MSOM except the occurrence probability for each species also includes a species-specific spatial Gaussian process ($\mathbf{w}_i(\mathbf{s})$), along with associated parameters (θ_i), analogous to the spatial SSOM (Equation 5). The spatial parameters are estimated individually for each species using inverse-Gamma priors for the spatial variance parameters and uniform priors for the spatial range and smoothness parameters (Supplemental Information S2.5). The function spMsPGOcc fits spatial MSOMs.

2.5 | Integrated occupancy models (IOMs)

Model-based integration of multiple data types has become common as the number of available data sources has increased (Miller et al., 2019). When integrating multiple detection–nondetection data sources in an IOM, each data source has its own unique detection model defined by Equations 3 and 4 that are conditional on a shared latent occupancy process defined by Equations 1 and 2. This joint-likelihood approach enables explicit estimation of different covariate effects (and intercepts) on the detection processes of each data source (Supplemental Information S2.6). The spOccupancy function intPGOcc fits IOMs.

2.6 | Spatial integrated occupancy models (IOMs)

The spatial IOM is identical to the IOM except the latent occurrence probability includes a spatial random intercept following Equation 5

(Supplemental Information S2.7). The function spIntPGOcc fits spatial IOMs.

3 | COMPUTATIONAL ADVANCES IN spOccupancy

Bayesian occupancy models using a logit link function are often slow and inefficient in standard Bayesian software packages (Clark & Altwegg, 2019), which arises from a need to use inefficient algorithms to estimate the occurrence and detection regression parameters. We avoid this computational burden by using Pólya-Gamma data augmentation (Polson et al., 2013), a statistical approach in which we introduce two sets of latent auxiliary variables that follow a Pólya-Gamma distribution, which results in an efficient Gibbs update for the occurrence and detection regression parameters in occupancy models with a logit link that is orders of magnitude faster than traditional algorithms (Clark & Altwegg, 2019). See Supplemental Information S1.1 for details.

Spatial models are a notorious computational bottleneck once the number of sites (J) becomes even moderately large. We provide users the option to fit all spatial occupancy models in spOccupancy using a NNGP (Datta et al., 2016), which enables spatially explicit occupancy modelling of data sources comprising locations in the tens to hundreds of thousands. The NNGP uses local information from a reduced set of nearest neighbours to provide inferences that are nearly indistinguishable from the full Gaussian process. Fifteen neighbours is often sufficient, although as few as five neighbours may be adequate for certain datasets with long-range spatial dependence (Datta et al., 2016). See Supplemental Information S1.2 for additional details.

We wrote all occupancy models fit by spOccupancy in C/C++ using R's foreign language interface. Our code draws heavily from the computational technology employed in the spNNGP package (Finley et al., 2020). See Supplemental Information S2 for complete details on the Gibbs samplers used in spOccupancy.

4 | IMPLEMENTATION AND USAGE OF spOccupancy

Here we briefly describe functionality for the five main tasks performed by spOccupancy (Table 1). See the package vignette (Supplemental Information S3), the package website (https://www.jeffdoser.com/files/spoccupancy-web/), and the R package documentation for additional details and examples.

1. *Data simulation.* The functions simOcc, simMsOcc and simIntOcc simulate data under the SSOM, MSOM and IOM frameworks. All simulation functions include arguments to optionally simulate data with spatial random effects in the occurrence portion of the model.
2. *Model fitting.* Each of the model fitting functions was described previously (Section 2). Functions for SSOMs and MSOMs allow

for the inclusion of random intercepts in the occurrence and detection portion of the occupancy model (e.g. random effects for some habitat classification, random observer effects). Users can specify each parameter's prior distribution to yield vague or informative priors as desired, with the default being weakly informative priors (See Supplemental Information S1.3).

3. *Model validation and comparison*. The `spOccupancy` function `ppcOcc` performs posterior predictive checks on all `spOccupancy` model objects, with options to calculate Bayesian *p*-values as a simple assessment of model fit. For model selection and assessment, the function `waicOcc` computes the WAIC (Watanabe, 2010). Alternatively, users can perform k-fold cross-validation using the `k.fold` argument in all `spOccupancy` model fitting functions. We use the model deviance as a scoring rule for k-fold cross-validation (Hooten & Hobbs, 2015).

4. *Posterior summaries*. All posterior samples are returned as `coda::mcmc` objects (Plummer et al., 2006). We include `summary` functions for all `spOccupancy` model objects, which print concise summaries (e.g. posterior quantiles, posterior means) of the posterior distributions for estimated parameters as well as the Gelman–Rubin diagnostic (Rhat; Brooks & Gelman, 1998) and effective sample size for convergence diagnostics.

5. *Prediction*. We implement a `predict` function for all `spOccupancy` model objects to yield predictions of latent occurrences and occurrence probabilities across a set of locations (given covariate values and spatial coordinates), which may or may not include a subset (or all) of the surveyed locations. The resulting object consists of posterior predictive distributions which can be used to provide maps of occurrence probability (i.e. species distribution maps) and fully propagated estimation uncertainty. We

additionally allow users to predict detection probability across a user-specified range of covariate values.

5 | CASE STUDIES

We demonstrate `spOccupancy` functionality with two case studies on forest breeding birds in the eastern United States. See Supplemental Information S1 for full case study details and an analysis of three simulated datasets across $J = 40,000$ locations using a spatial IOM.
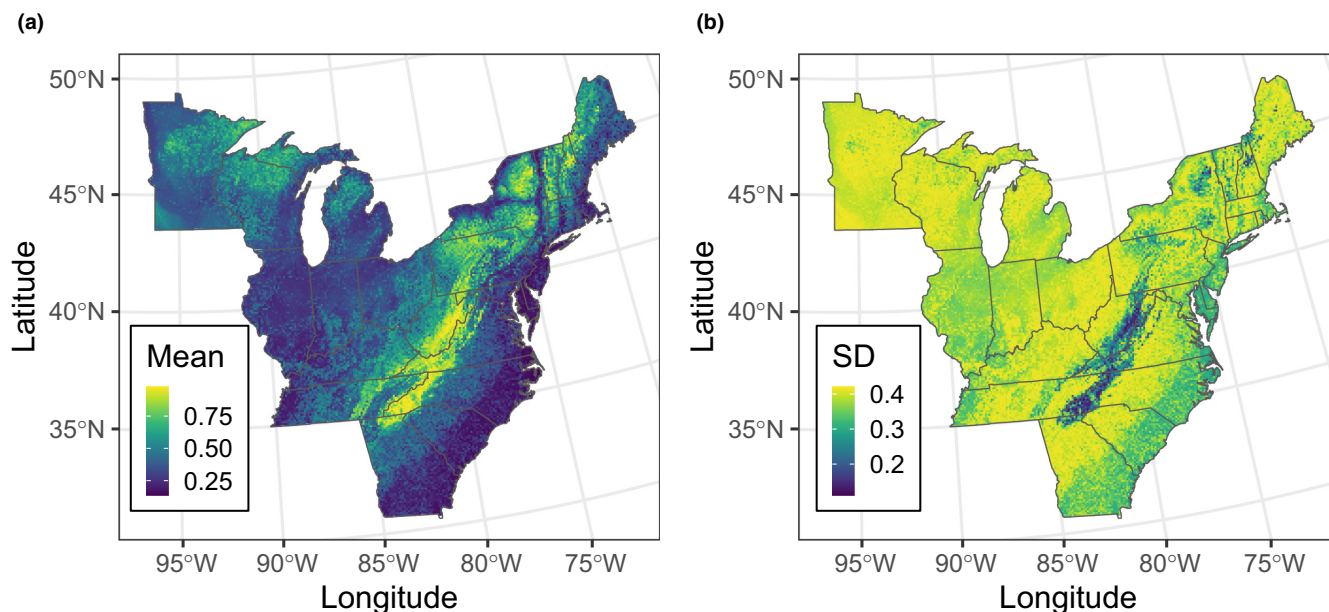
5.1 | Black-throated Green Warbler in eastern United States (SSOM)

For our first example, we estimated occurrence of Black-throated Green Warbler *Setophaga virens* across the eastern United States in 2018. We used data from the North American Breeding Bird Survey (BBS; Pardieck et al., 2020), where observers perform roadside surveys at 50 stops along ~3,000 routes each year distributed across the United States and Canada. We modelled route-level occurrence as a function of local forest cover (linear) and elevation (linear and quadratic) and modelled detection as a function of day of survey (linear and quadratic), time of day (linear) and a random observer effect. All variables were standardized to have mean 0 and standard deviation 1. We fit a non-spatial and a spatial SSOM using `PGOcc` and `spPGOcc` respectively. We fit the spatial SSOM using a full Gaussian process and also with a NNGP with 15 neighbours to assess the computational benefits

TABLE 2 Posterior median estimates (95% credible intervals), WAIC, 10-fold cross-validation (CV) deviance and run time for a SSOM and spatial SSOMs using a NNGP and a full Gaussian process (GP) for the Black-throated Green Warbler case study. Run times are times to complete 50,000 MCMC samples

	Nonspatial	Spatial (NNGP)	Spatial (GP)
Occurrence			
Intercept	−1.43 (−1.74, −1.13)	−1.28 (−3.98, 0.96)	−1.61 (−3.94, 0.94)
Linear elevation	1.38 (1.06, 1.73)	2.13 (1.22, 3.27)	1.95 (1.05, 3.00)
Quadratic elevation	0.26 (−0.08, 0.61)	0.16 (−0.32, 0.82)	0.21 (−0.30, 0.90)
Forest cover	1.15 (0.89, 1.44)	1.11 (0.58, 1.84)	1.06 (0.52, 1.68)
Spatial variance	—	23.57 (9.77, 51.89)	21.44 (8.53, 45.46)
Spatial range	—	0.0021 (0.0013, 0.0039)	0.0022 (0.0013, 0.0041)
Detection			
Intercept	−0.46 (−0.90, −0.08)	−0.14 (−0.44, 0.15)	−0.14 (−0.45, 0.16)
Linear Day	−0.15 (−0.40, 0.10)	−0.14 (−0.37, 0.08)	−0.14 (−0.34, 0.08)
Quadratic Day	0.062 (−0.17, 0.30)	0.033 (−0.17, 0.24)	0.040 (−0.16, 0.24)
Time of Day	−0.01 (−0.32, 0.29)	−0.010 (−0.25, 0.24)	−0.019 (−0.27, 0.24)
Observer variance	2.49 (1.55, 3.70)	1.60 (1.01, 2.34)	1.58 (0.99, 2.37)
WAIC	2,401.16	2,117.01	2,118.95
CV deviance	3,317.39	2,504.03	2,537.67
Run time (min)	3.17	7.16	1,670.73

FIGURE 1 Predicted occurrence probability mean (a) and standard deviation (b) for Black-throated Green Warbler across the eastern United States in 2018 as estimated using a spatial SSOM
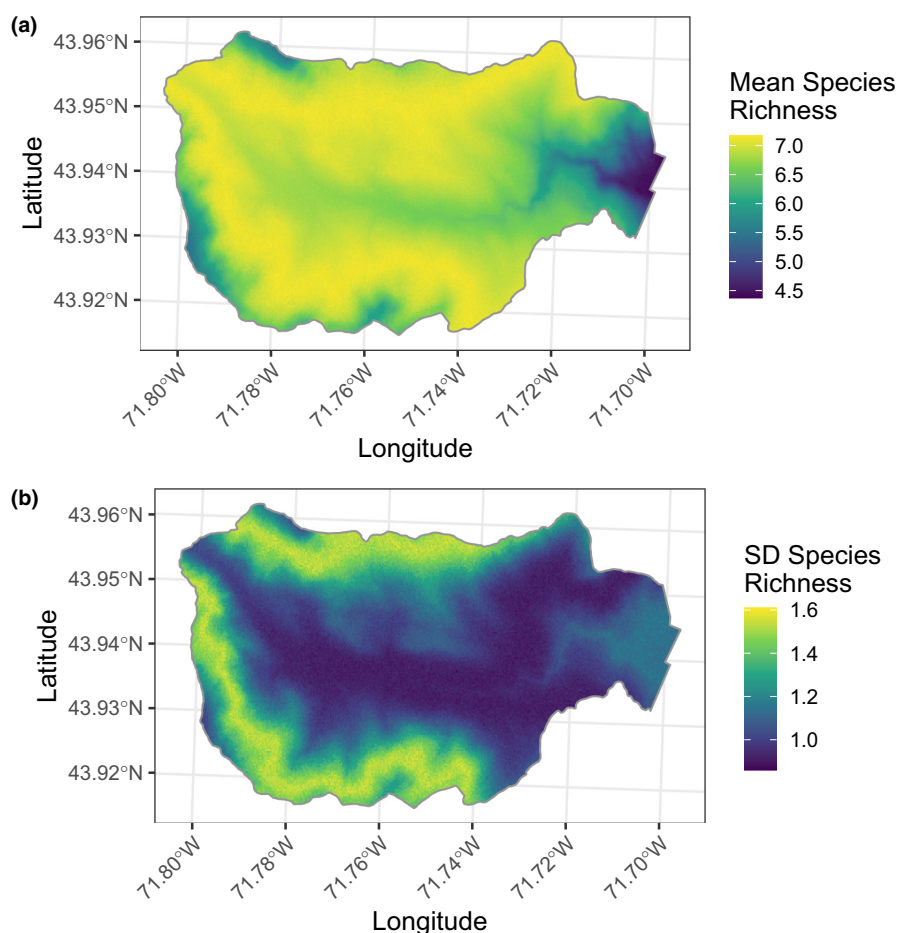
FIGURE 2 Estimate of species richness and the associated uncertainty of the community of 12 foliage-gleaning birds across the Hubbard Brook Experimental Forest from a non-spatial MSOM. Panel (a) shows posterior means and panel (b) shows posterior standard deviations

provided by the NNGP. For both spatial models we used an exponential correlation function.

The spatial SSOMs outperformed the nonspatial SSOM according to both the WAIC and 10-fold cross-validation (Table 2). The

NNGP spatial model provided massive increases in computing efficiency compared to the full Gaussian process spatial model (Table 2) with only small differences in parameter estimates and effective sample sizes. Predicted occurrence probabilities from the NNGP

TABLE 3 Candidate model community-level posterior median estimates (95% credible intervals), WAIC and four-fold cross-validation (CV) deviance in the foliage-gleaning bird case study in the Hubbard Brook Experimental Forest. μ_{β_0} is the community-level intercept, μ_{β_1} and μ_{β_2} are the linear and quadratic effects of elevation, respectively, and $\tau^2_{\beta,0}$, $\tau^2_{\beta,1}$ and $\tau^2_{\beta,2}$ are community-level variances for the intercept, linear and quadratic elevation effects across species respectively. Run times are minutes to complete 150,000 MCMC iterations

	Intercept	Intercept + Spatial	Elevation	Elevation + Spatial
μ_{β_0}	0.37 (−0.82, 1.55)	0.38 (−1.44, 2.19)	0.43 (−1.28, 2.23)	0.42 (−1.59, 2.36)
μ_{β_1}	—	—	0.26 (−0.79, 1.24)	0.29 (−1.08, 1.60)
μ_{β_2}	—	—	−0.21 (−0.68, 0.30)	−0.26 (−0.90, 0.43)
$\tau^2_{\beta,0}$	4.51 (1.39, 13.10)	12.99 (3.61, 39.51)	11.67 (4.21, 30.83)	18.10 (6.60, 45.84)
$\tau^2_{\beta,1}$	—	—	3.21 (1.02, 8.56)	5.70 (1.66, 15.18)
$\tau^2_{\beta,2}$	—	—	0.60 (0.15, 1.76)	1.00 (0.22, 3.14)
WAIC	9,714.31	9,057.95	9,195.61	9,043.74
CV deviance	10,137.57	10,091.36	9,989.48	10,063.4
Run time	20.56	42.69	20.90	42.74

spatial SSOM indicated high occurrence of Black-throated Green Warbler across the Appalachian mountains, the northeastern states and a portion of the northern Midwest (Figure 1).

5.2 | Foliage-gleaning birds in Hubbard brook (MSOM)

For our second case study, we estimated species richness for a community of 12 foliage-gleaning birds in 2015 in the Hubbard Brook Experimental Forest in New Hampshire, USA. Data were collected using standard point count surveys at 373 sites three times during the breeding season. We included time of day (linear) and day of year (linear and quadratic) as fixed effects in the detection portion of the model, and specified linear and quadratic effects of elevation as occurrence predictors. All variables were standardized to have mean 0 and standard deviation 1. We fit a series of nonspatial and spatial MSOMs using `msPGOcc` and `spMsPGOcc` where we compared the benefits of including elevation as an occurrence predictor and spatial random effects using an exponential correlation function. We fit all spatial models using an NNGP with five neighbours.

The spatial MSOM that included linear and quadratic elevation covariates on occurrence was the best performing model according to the WAIC, but the nonspatial MSOM with elevation performed best according to four-fold cross-validation (Table 3). The spatially explicit intercept-only model outperformed the nonspatial intercept-only model according to both criteria, which altogether indicates there is spatial variation in foliage-gleaning bird occurrence across the study region, but that this variation is fairly well-explained by the important elevational gradient in the study area. We predicted species richness across Hubbard Brook using the nonspatial MSOM with the elevation covariate, which revealed moderate variation across Hubbard Brook with areas of low and high elevation having on average lower richness than moderate elevations (Figure 2a) and high uncertainty along the edges of the forest, which closely corresponds

to high elevation areas (Figure 2b, Supplemental Information S1.S4 Figure S1).

6 | CONCLUSIONS AND FUTURE DIRECTIONS

Our `spOccupancy` R package fits spatially explicit single-species, multi-species and integrated occupancy models for potentially massive datasets. The package includes functions for data simulation, model fitting, model validation, model comparison and out-of-sample prediction. The package vignette (Supplemental Information S3) and website (https://www.jeffdoser.com/files/spoccupancy-web/) contain full details and examples on all `spOccupancy` model functions. We are currently working on including the following extensions within the package: (a) dynamic occupancy models (MacKenzie et al., 2003); (b) spatially varying coefficients (SVCs; Finley, 2011) in the occurrence model; (c) multi-species integrated occupancy models (Doser, Leuenberger, et al., 2022). We expect `spOccupancy` will serve as a user-friendly tool for ecologists and conservation practitioners to account for detection biases and spatial autocorrelation using large datasets (e.g. hundreds of thousands of locations) in assessments of species distributions and community patterns across broad spatial regions, an increasingly important objective in species distribution modelling applications.

CONFLICT OF INTEREST

We declare no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13897.

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 examples are available at https://doi.org/10.5281/zenodo.6514011 (Doser, Finley, et al., 2022).

ORCID

Jeffrey W. Doser (ID) https://orcid.org/0000-0002-8950-9895
Andrew O. Finley (ID) https://orcid.org/0000-0002-2277-2912
Elise F. Zipkin (ID) https://orcid.org/0000-0003-4155-6139

REFERENCES

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2003). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.

Banerjee, S., & Fuentes, M. (2012). Bayesian modeling for large spatial datasets. *WIREs Computational Statistics*, 4(1), 59–66.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4), 434–455.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 1–32.

Clark, A. E., & Altwegg, R. (2019). Efficient Bayesian analysis of occupancy models with logit link functions. *Ecology and Evolution*, 9(2), 756–768.

Datta, A., Banerjee, S., Finley, A. O., & 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 Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman, C., Temple Lang, D., & Bodik, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, 26, 403–413.

Dorazio, R. M., & 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., Kéry, M., & Zipkin, E. F. (2022). Code and data for spOccupancy: An R package for single-species, multi-species, and integrated spatial occupancy models. *Zenodo*. https://doi.org/10.5281/zenodo.6514011

Doser, J. W., Finley, A. O., Weed, A. S., & Zipkin, E. F. (2021). Integrating automated acoustic vocalization data and point count surveys for estimation of bird abundance. *Methods in Ecology and Evolution*, 12(6), 1040–1049.

Doser, J. W., Leuenberger, W., Sillett, T. S., Hallworth, M. T., & Zipkin, E. F. (2022). Integrated community occupancy models: A framework to assess occurrence and biodiversity dynamics using multiple data sources. *Methods in Ecology and Evolution*, 13(4), 919–932.

Finley, A. O. (2011). Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods in Ecology and Evolution*, 2(2), 143–154.

Finley, A. O., Datta, A., & Banerjee, S. (2020). spNNGP R package for nearest neighbor Gaussian process models. *arXiv preprint arXiv*, 2001, 09111.

Finley, A. O., Datta, A., Cook, B. D., Morton, D. C., Andersen, H. E., & Banerjee, S. (2019). Efficient algorithms for Bayesian nearest neighbor Gaussian processes. *Journal of Computational and Graphical Statistics*, 28(2), 401–414.

Fiske, I., & Chandler, R. (2011). Unmarked: An R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software*, 43(10), 1–23.

Gelfand, A. E., Schmidt, A. M., Wu, S., Silander, J. A., Jr., Latimer, A., & Rebelo, A. G. (2005). Modelling species diversity through species level hierarchical modelling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1), 1–20.

Guélat, J., & Kéry, M. (2018). Effects of spatial autocorrelation and imperfect detection on species distribution models. *Methods in Ecology and Evolution*, 9(6), 1614–1625.

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, 135(2–3), 147–186.

Hines, J. E. (2006). *Presence 3.1 software to estimate patch occupancy and related parameters*. http://www.mbr-pwrc.usgs.gov/software/presence.html

Hooten, M. B., & Hobbs, N. T. (2015). A guide to Bayesian model selection for ecologists. *Ecological Monographs*, 85(1), 3–28.

Isaac, N. J., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., Freeman, S. N., Golding, N., Guillera-Arroita, G., Henrys, P. A., et al. (2020). Data integration for large-scale models of species distributions. *Trends in Ecology & Evolution*, 35(1), 56–67.

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

Kellner, K. F., Fowler, N. L., Petroelje, T. R., Kautz, T. M., Beyer, D. E., Jr., & Belant, J. L. (2021). Ubms: An R package for fitting hierarchical occupancy and N-mixture abundance models in a Bayesian framework. *Methods in Ecology and Evolution*, 13(3), 577–584.

Kéry, M., & Royle, J. A. (2021). *Applied hierarchical modeling in ecology: Analysis of distribution, abundance, and species richness in R and BUGS: Volume 2: Dynamic and advanced models*. Academic Press.

MacKenzie, D. I., Nichols, J. D., Hines, J. E., Knutson, M. G., & Franklin, A. B. (2003). Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology*, 84(8), 2200–2207.

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

Miller, D. A., Pacifici, K., Sanderlin, J. S., & Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, 10(1), 22–37.

Pardieck, K., Ziolkowski, D., Jr., Lutmerding, M., Aponte, V., & Hudson, M.-A. (2020). *North american breeding bird survey dataset 1966–2019*. U.S. Geological Survey data release. https://doi.org/10.5066/P9J6QUF6

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1), 7–11.

Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504), 1339–1349.

Pulliam, H. R. (2000). On the relationship between niche and distribution. *Ecology Letters*, 3(4), 349–361.

Tikhonov, G., Opedal, Ø. H., Abrego, N., Lehikoinen, A., de Jonge, M. M., Oksanen, J., & Ovaskainen, O. (2020). Joint species distribution modelling with the r-package Hmsc. *Methods in Ecology and Evolution*, 11(3), 442–447.

Tyre, A. J., Tenhumberg, B., Field, S. A., Niejalke, D., Parris, K., & Possingham, H. P. (2003). Improving precision and reducing bias in biological surveys: Estimating false-negative error rates. *Ecological Applications*, 13(6), 1790–1801.

Vieilledent, G. (2019). *hSDM: Hierarchical Bayesian species distribution models*. R package version 1.4.1.

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), 3571–3594.

White, G. C., & Burnham, K. P. (1999). Program MARK: Survival estimation from populations of marked animals. *Bird Study*, 46(suppl. 1), S120–S139.

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

Zipkin, E. F., DeWan, A., & Andrew Royle, J. (2009). Impacts of forest fragmentation on species richness: A hierarchical approach to community modelling. *Journal of Applied Ecology*, 46(4), 815–822.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.