www.nature.com/scientificreports

scientific reports

OPEN

Community confounding in joint species distribution models

Justin J. Van Ee[1], Jacob S. Ivan[2] & Mevin B. Hooten[3]

Joint species distribution models have become ubiquitous for studying species-environment relationships and dependence among species. Accounting for community structure often improves predictive power, but can also affect inference on species-environment relationships. Specifically, some parameterizations of joint species distribution models allow interspecies dependence and environmental effects to explain the same sources of variability in species distributions, a phenomenon we call community confounding. We present a method for measuring community confounding and show how to orthogonalize the environmental and random species effects in suite of joint species distribution models. In a simulation study, we show that community confounding can lead to computational difficulties and that orthogonalizing the environmental and random species effects can alleviate these difficulties. We also discuss the inferential implications of community confounding and orthogonalizing the environmental and random species effects in a case study of mammalian responses to the Colorado bark beetle epidemic in the subalpine forest by comparing the outputs from occupancy models that treat species independently or account for interspecies dependence. We illustrate how joint species distribution models that restrict the random species effects to be orthogonal to the fixed effects can have computational benefits and still recover the inference provided by an unrestricted joint species distribution model.

Historically, species distributions have been modeled independently from each other due to unavailability of multispecies datasets and computational restraints. However, ecological datasets that provide insights about collections of organisms have become prevalent over the last decade thanks to efforts like Long Term Ecological Research Network (LTER), National Ecological Observatory Network (NEON), and citizen science surveys[1]. In addition, technology has improved our ability to fit modern statistical models to these datasets that account for both species environmental preferences and interspecies dependence. These advancements have allowed for the development of joint species distribution models (JSDM)[2–4] that can model dependence among species simultaneously with environmental drivers of occurrence and/or abundance.

Species distributions are shaped by both interspecies dynamics and environmental preferences[5–8]. JSDMs integrate both sources of variability and adjust uncertainty to reflect that multiple confounded factors can contribute to similar patterns in species distributions. Some have proposed that JSDMs not only account for biotic interactions but also correct estimates of association between species distributions and environmental drivers[3,9], while others claim JSDMs cannot disentangle the roles of interspecies dependence and environmental drivers[5]. We address why JSDMs can provide inference distinct from their concomitant independent SDMs, how certain parameterizations of a JSDM induce confounding between the environmental and random species effects, and when deconfounding these effects may be appealing for computation and interpretation.

Because of the prevalence of occupancy data for biomonitoring in ecology, we focus our discussion of community confounding in JSDMs on occupancy models, although we also consider a JSDM for species density data in the simulation study. The individual species occupancy model was first formulated by MacKenzie et al.[10] and has several joint species extensions[4,11–16]. We chose to investigate the impacts of community confounding on the probit model since it has been widely used in the analysis of occupancy data[4,13,17]. We also developed a joint species extension to the Royle-Nichols model[18] and consider community confounding in that model.

We use the probit and Royle-Nichols occupancy models to improve our understanding of montaine mammal communities in what follows. We show that including unstructured random species effects in either occupancy model induces confounding between the fixed environmental and random species effects. We demonstrate how to orthogonalize these effects in the model and compare the resulting inference compared to models where species are treated independently.

[1]Department of Statistics, Colorado State University, Fort Collins 80523, USA. [2]Colorado Parks and Wildlife, Fort Collins 80526, USA. [3]Department of Statistics and Data Sciences, The University of Texas at Austin, Austin 78712, USA. email: vanee002@colostate.edu

Unlike previous approaches that have applied restricted regression techniques similar to ours, we use it in the context of well-known ecological models for species occupancy and intensity. While such approaches have been discussed in spatial statistics and environmental science, they have not been adopted in settings involving the multivariate analysis of community data. We draw parallels between restricted spatial regression and restricted JSDMs but also highlight where the methods differ in goals and outcomes. We find that the computational benefits conferred by performing restricted spatial regression also hold for some joint species distribution models.

Royle-Nichols joint species distribution model.

We present a JSDM extension to the Royle-Nichols model[18]. The Royle-Nichols model accounts for heterogeneity in detection induced by the species' latent intensity, a surrogate related to true species abundance. Abundance, density, and occupancy estimation often requires an explicit spatial region that is closed to emmigration and immigration. In our model, the unobservable intensity variable helps us explain heterogeneity in the frequencies we observe a species at different sites without making assumptions about population closure. In the "Model" section, we further discuss the distinctions between abundance and intensity in the Royle-Nichols model.

The Royle-Nichols model utilizes occupancy survey data but provides inference distinct from the basic occupancy model[10]. In the Royle-Nichols model, we estimate individual detection probability for homogeneous members of the population, whereas in an occupancy model, we estimate probability of observing at least one member of the population given that the site is occupied. Furthermore, the Royle-Nichols model allows us to relate environmental covariates to the latent intensity associated with a species at a site, while in an occupancy model, environmental covariates are associated with the species latent probability of occupancy at a site. Species intensity and occupancy may be governed by different mechanisms, and inference from an intensity model can be distinct from that provided by an occupancy model[19–21]. Cingolani et al.[20] proposed that, in plant communities, certain environmental filters preclude species from occupying a site and an additional set of filters may regulate if a species can flourish. Hence, certain covariates that were unimportant in an occupancy model may improve predictive power in an intensity model.

Community confounding.

Species distributions are shaped by environment as well as competition and mutualism within the community[8,22,23]. Community confounding occurs when species distributions are explained by a convolution of environmental and interspecies effects and can lead to inferential differences between a joint and single species distribution model as well as create difficulties for fitting JSDMs. Former studies have incorporated interspecies dependence into an occupancy model[4,11–16], and others have addressed spatial confounding[1,17,24,25], but none of these explicitly addressed community confounding. However, all Bayesian joint occupancy models naturally attenuate the effects of community confounding due to the prior on the regression coefficients. The prior, assuming it is proper, induces regularization on the regression coefficients[26] that can lessen the inferential and computational impacts of confounding[27]. Furthermore, latent factor models like that described by Tobler et al.[4] restrict the dimensionality of the random species effect which should also reduce confounding with the environmental effects.

We address community confounding by formulating a version of our model that orthogonalizes the environmental effects and random species effects. Orthogonalizing the fixed and random effects is common practice in spatial statistics and often referred to as restricted spatial regression[27–31]. Restricted regression has been applied to spatial generalized linear mixed models (SGLMM) for observations $\boldsymbol{y}$, which can be expressed as

$$\boldsymbol{y} \sim [\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\psi}], \tag{1}$$

$$g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta}, \tag{2}$$

$$\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{3}$$

where $g(\cdot)$ is a link function, $\boldsymbol{\psi}$ are additional parameters for the data model, and $\boldsymbol{\Sigma}$ is the covariance matrix of the spatial random effect. In the SGLMM, prior information facilitates the estimation of $\boldsymbol{\eta}$, which would not be estimable otherwise due to its shared column space with $\boldsymbol{\beta}$[30]. This is analogous to applying a ridge penalty to $\boldsymbol{\eta}$, which stabilizes the likelihood. Another method for fitting the confounded SGLMM is to specify a restricted version:

$$\boldsymbol{y} \sim [\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\psi}], \tag{4}$$

$$g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\delta} + (\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{\eta}, \tag{5}$$

$$\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{6}$$

where $\boldsymbol{P}_X = \boldsymbol{X}(\boldsymbol{X}\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the projection matrix onto the column space of $\boldsymbol{X}$. In the unrestricted SGLMM, the regression coefficients $\boldsymbol{\beta}$ and random effect $\boldsymbol{\eta}$ in (1) compete to explain variability in the latent mean $\boldsymbol{\mu}$ in the direction of $\boldsymbol{X}$[27]. In the restricted model, however, all variability in the direction of $\boldsymbol{X}$ is explained solely by the regression coefficients $\boldsymbol{\delta}$ in (4)[31], and $\boldsymbol{\eta}$ explains residual variation that is orthogonal to $\boldsymbol{X}$. We refer to $\boldsymbol{\beta}$ as the conditional effects because they depend on $\boldsymbol{\eta}$, and $\boldsymbol{\delta}$ as the unconditional effects.

Restricted regression, as specified in (4), was proposed by Reich et al.[28]. Reich et al.[28] described a disease-mapping example in which the inclusion of a spatial random effect rendered one covariate effect unimportant that was important in the non-spatial model. Spatial maps indicated an association between the covariate and response, making inference from the spatial model appear untenable. Reich et al.[28] proposed restricted spatial regression as a method for recovering the posterior expectations of the non-spatial model and shrinking the posterior variances which tend to be inflated for the unrestricted SGLMM.

Several modifications of restricted spatial regression have been proposed[30,32–35]. All restricted spatial regression methods seek to provide posterior means $\mathrm{E}\left(\delta_j|\boldsymbol{y}\right)$ and marginal posterior variances $\mathrm{Var}\left(\delta_j|\boldsymbol{y}\right)$, $j = 1, ..., p$ that satisfy the following two conditions[36]:

1. $\mathrm{E}\left(\boldsymbol{\delta}|\boldsymbol{y}\right) = \mathrm{E}\left(\boldsymbol{\beta}_{\mathrm{NS}}|\boldsymbol{y}\right)$ and,
2. $\mathrm{Var}\left(\beta_{\mathrm{NS},j}|\boldsymbol{y}\right) \leq \mathrm{Var}\left(\delta_j|\boldsymbol{y}\right) \leq \mathrm{Var}\left(\beta_{\mathrm{Spatial},j}|\boldsymbol{y}\right)$ for $j = 1, ..., p$,

where $\boldsymbol{\beta}_{NS}$ and $\boldsymbol{\beta}_{Spatial}$ are the regression coefficients corresponding to the non-spatial and unrestricted spatial models, respectively.

The inferential impacts of spatial confounding on the regression coefficients has been debated. Hodges and Reich[29] outlined five viewpoints on spatial confounding and restricted regression in the literature and refuted the two following views:

1. Adding the random effect $\boldsymbol{\eta}$ corrects for bias in $\boldsymbol{\beta}$ resulting from missing covariates.
2. Estimates of $\boldsymbol{\beta}$ in a SGLMM are shrunk by the random effect and hence conservative.

The random effect $\boldsymbol{\eta}$ can increase or decrease the magnitude of $\boldsymbol{\beta}$, and the change may be galvanized by mechanisms not related to missing covariates. Therefore, we cannot assume the regression coefficients in the SGLMM will exceed those of the restricted model, nor should we regard the estimates in either model as biased due to misspecification. Confounding in the SGLMM causes $\mathrm{Var}\left(\beta_j|\boldsymbol{y}\right) \geq \mathrm{Var}\left(\delta_j|\boldsymbol{y}\right)$, $j = 1, ..., p$, because of the shared column space of the fixed and random effects. Thus, we refer to the conditional coefficients as conservative with regard to their credible intervals, not their posterior expectations.

Reich et al.[28] argued that restricted spatial regression should always be applied because the spatial random effect is generally added to improve predictions and/or correct the fixed effect variance estimate. While it may be inappropriate to orthogonalize a set of fixed effects in an ordinary linear model, orthogonalizing the fixed and random effect in a spatial model is permissible because the random effect is generally not of inferential interest. Paciorek[37] provided the alternative perspective that, if confounding exists, it is inappropriate to attribute all contested variability in $\boldsymbol{y}$ to the fixed effects. Hanks et al.[31] discussed factors for deciding between the unrestricted and restricted SGLMM on a continuous spatial support. The restricted SGLMM leads to improved computational stability, but the unconditional effects are less conservative under model misspecification and more prone to type-S errors: The Bayesian analogue of Type I error. Fitting the unrestricted SGLMM when the fixed and random effects are truly orthogonal does not introduce bias, but it will increase the fixed effect variance. Given these considerations, Hanks et al.[31] suggested a hybrid approach where the conditional effects, $\boldsymbol{\beta}$, are extracted from the restricted SGLMM. This is possible because the restricted SGLMM is a reparameterization of the unrestricted SGLMM. This hybrid approach leads to improved computational stability but yields the more conservative parameter estimates. We describe how to implement this hybrid approach for joint species distribution models in the “Community confounding” section.

Restricted regression has also been applied in time series applications. Dominici et al.[38] debiased estimates of fixed effects confounded by time using restricted smoothing splines. Without the temporal random effect, Dominici et al.[38] asserted all temporal variation in the response would be wrongly attributed to temporally correlated fixed effects. Houseman et al.[39] used restricted regression to ensure identifiability of a nonparametric temporal effect and highlighted certain covariate effects that were more evident in the restricted model (i.e., the unconditional effects' magnitude was greater). Furthermore, restricted regression is implicit in restricted maximum likelihood estimation (REML). REML is often employed for debiasing the estimate of the variance of $\boldsymbol{y}$ in linear regression and fitting linear mixed models that are not estimable in their unrestricted format[40]. Because REML is generally applied in the context of variance and covariance estimation, considerations regarding the effects of REML on inference for the fixed effects are lacking in the literature.

In ecological science, JSDMs often include an unstructured random effect like $\boldsymbol{\eta}$ in (1) to account for interspecies dependence, and hence can also experience community confounding between $\boldsymbol{X}$ and $\boldsymbol{\eta}$ analogous to spatial confounding. Unlike a spatial or temporal random effect, we consider random species effects to be inferentially important, rather than a tool solely for improving predictions or catch-all for missing covariates. An orthogonalization approach in a JSDM attributes contested variation between the fixed effects (environmental information) and random effect (community information) to the fixed effect.

We describe how to orthogonalize the fixed and random species effects in a suite of JSDMs and present a method for detecting community confounding. In the simulation study, we test the efficacy of our method for detecting confounding, show that community confounding can lead to computational difficulties similar to those caused by spatial confounding[31], and highlight that, for some models, restricted regression can improve model fitting. We also investigate the inferential implications of community confouding and restricted regression in JSDMs by comparing outputs from the SDM, unrestricted JSDM, and restricted JSDM of the Royle-Nichols and probit occupancy models fit to mammalian camera trap data. Lastly, we discuss other inferential and computational methods for confounded models and consider their appropriateness for joint species distribution modeling.

Model

Data model. The probit and Royle-Nichols occupancy models were developed for analyzing multispecies binary detection data, y_{ijk}, arising from a zero-inflated Bernoulli process with probability of success p_{ijk}, where $i = 1, \ldots, n$, $j = 1, \ldots, J_i$, and $k = 1, \ldots, K$ correspond to sites, occasions, and species, respectively. Occupancy data of this form have traditionally been analyzed in a latent variable framework[10,41,42]. In what follows, we let $z_{ik} \sim \text{Bern}(\psi_{ik})$ be an indicator on whether species k occupies site i. Given a site is occupied, we detect species k on occasion j with some probability p_{ijk}, such that $(y_{ijk}|z_{ik} = 1) \sim \text{Bern}(p_{ijk})$, but if species k is absent from the site, we have zero probability of detecting it, $P(y_{ijk} = 0|z_{ik} = 0) = 1$.

The probit occupancy model is so named because it links ψ_{ik} and p_{ijk} to occupancy and detection covariates $\boldsymbol{x}_{ik}$ and $\boldsymbol{w}_{ijk}$, respectively, with the standard normal CDF Φ. The probit link function can be paired with data augmentation[17,43–45] to yield efficient Gibbs samplers for the occupancy and detection regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, respectively.

Royle and Nichols[18] introduced a method for analyzing occupancy data that explicitly modeled the probability of detecting species k at a site as a function of a surrogate related to the true species abundance. Assuming there are N_{ik} individuals of species k in sample region i and that all individuals in species k on the sample unit have identical detection probabilities and are detected independently of other individuals, the probability of detecting at least one of these individuals can be expressed as

$$\rho_{ijk} = 1 - (1 - r_{jk})^{N_{ik}}, \tag{7}$$

where r_{jk} is a binomial sampling probability that a particular individual of species k is detected on occasion j. While the Royle-Nichols model facilitates inference on number of individuals of species k, N_{ik}, at each site when all the assumptions are met, we do not interpret them as such because sites are not necessarily closed in camera trap studies due to mobile species with home ranges larger than the sampling radius of the camera. Note that ρ_{ijk} in (7) corresponds to the species probability of detection conditional on an intensity process. This is distinct from p_{ijk} in the probit model that is conditional on an occupancy process.

The nonlinear function of r_{jk} and N_{ik} in (7) involves more parameters than would be identifiable in a typical occupancy model, especially when the individual detection probability is heterogeneous across occasions (e.g., r_{jk} are heterogeneous). In the heterogeneous case, r_{jk} is connected to covariates with the logit link function:

$$\text{logit}(r_{jk}) = f(\boldsymbol{w}_{jk}, \boldsymbol{\alpha}_k), \tag{8}$$

where $f(\boldsymbol{w}_{ijk}, \boldsymbol{\alpha}_k)$ is a linear function of the detection covariates $\boldsymbol{w}_{ijk}$ and regression parameters $\boldsymbol{\alpha}_k$.

Modeling interspecies dependence. We extend both occupancy models to account for interspecies dependence by including random species effects in their process models. Following Royle and Nichols[18], we assume $N_{ik} \sim \text{Pois}(\lambda_{ik})$, where λ_{ik} is mean intensity of species k at site i. We let $\boldsymbol{\lambda}$ denote the vector of site specific intensities stacked across the K species in the community. To model interspecies dependence, we specify the conditional multivariate normal distribution:

$$\log(\boldsymbol{\lambda}) \sim \mathcal{N}(\boldsymbol{X\beta} + \boldsymbol{\eta}, \boldsymbol{\Sigma}_{\lambda}), \tag{9}$$

$$\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{spp} \otimes \boldsymbol{I}_n), \tag{10}$$

where $\boldsymbol{X}$ is a block-diagonal matrix of the K species design matrices, $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \ldots, \boldsymbol{\beta}'_K)'$ is a stacked vector of species specific regression coefficients, $\boldsymbol{\eta}$ represents the random species effects, and $\boldsymbol{\Sigma}_{spp}$ is a species covariance matrix, and $\boldsymbol{\Sigma}_{\lambda}$ is a matrix that allows for additional covariance structures such as spatial dependence. For our purposes of comparing the SDM, unrestricted JSDM, and restricted JSDM for differences galvanized by community confounding, we assumed a simple independent structure for $\log(\boldsymbol{\lambda})$ and set $\boldsymbol{\Sigma}_{\lambda} = \tau \boldsymbol{I}$.

In the probit model, we include a random species effect in the latent probability of occupancy: $\Phi(\boldsymbol{\psi}) = \boldsymbol{X\beta} + \boldsymbol{\eta}$, where $\boldsymbol{\psi}$ is a vector of site specific occupancy probabilities stacked across the K species in the community and $\boldsymbol{X}$, $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \ldots, \boldsymbol{\beta}'_K)'$, and $\boldsymbol{\eta}$ are defined as above.

In both occupancy models, $\boldsymbol{\eta}$ allows for dependence between all K species in the community at each site. In the probit model, $\boldsymbol{\eta}$ characterizes interspecies dependence in the probability of occupancy, whereas in the Royle-Nichols model interspecies dependence is characterized in the species latent intensities. Just as certain environmental features may not preclude species occupancy but can curb intensity, some species may coexist in a region but not be able to jointly flourish[46]. Hence, interspecies dependence on latent intensity is conceptually distinct from interspecies dependence on probability of occupancy and may lead to inferential differences in $\boldsymbol{\eta}$ in the two occupancy models.

Tobler et al.[4] developed a joint occupancy model that accounts for community structure using a latent variable approach. They express the latent probability of occupancy of species k at site i as

$$\Phi(\psi_{ik}) = \boldsymbol{x}'_i \boldsymbol{\beta}_k + \boldsymbol{l}'_i \boldsymbol{\theta}_k, \tag{11}$$

where $\boldsymbol{l}'_i$ is a vector of length T of latent variables, and $\boldsymbol{\theta}_k$ are species specific regression coefficients. The latent variable model (LVM) is a computationally efficient and implicitly accounts for community structure. Other occupancy models have included interspecies dependence in the structure of the regression coefficients. Known as multispecies models, these models assume the species specific regression coefficients $\boldsymbol{\beta}_k$ stem from a common multivariate normal distribution $\boldsymbol{\beta}_k \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\beta})$ where $\boldsymbol{\mu}$ is the typical response of a species to covariates $\boldsymbol{x}$ and $\boldsymbol{\Sigma}_{\beta}$ allows for dependence in different species response to the same covariates[47]. In our study of mammalian camera trap data, each species is modeled with unique covariates, and we do not consider shared environmental responses.

Scheffe[48] stipulated that the levels of a random effect are draws from a population, and the draws are not of interest in themselves but only as samples from the larger population, which is of interest. In more recent literature, the term random effect is used more broadly. Hodges and Clayton[49] categorized modern definitions of a random effect into three different varieties. The definition commonly used in spatial statistics is, the levels of the effect arise from a meaningful population, but they are the whole population and these particular levels are of interest. We adopt this definition for the random species effects in (9). In practice, some levels of the population will likely not be included in the random species effects. For example, in Ivan et al.[50], cameras were baited and arranged to capture all members of the mammalian community, but several species were excluded from the random species effects due to a lack of detections.

Priors. We specified normal priors for the regression coefficients, $\boldsymbol{\beta}$, in the intensity and occupancy processes of the Royle-Nichols and probit models, respectively to facilitate comparison with the occupancy and spatial confounding literature. We also specified normal priors for the detection coefficients, $\boldsymbol{\alpha}$, in the observation model and the conjugate Inverse-Wishart prior for the species covariance matrix $\boldsymbol{\Sigma}_{spp}$. A more general alternative to the Inverse-Wishart prior is to apply a Cholesky decomposition, $\boldsymbol{\Sigma}_{spp} = \boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{L}'$, where $\boldsymbol{L}$ is lower diagonal with ones along the diagonal and $\boldsymbol{D}$ is diagonal with positive diagonal elements, and specify priors for the lower diagonal elements of $\boldsymbol{L}$ and diagonal elements of $\boldsymbol{D}$[51]. We found the Inverse-Wishart prior suitable for our inferential goals, but see Chan and Jeliazkov[51] for alternative covariance matrix priors.

The joint posterior distribution associated with our model is

$$[\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{N}, \boldsymbol{\Sigma}_{spp}|\boldsymbol{y}] \propto \prod_{k=1}^{K}\left(\prod_{i=1}^{n}\left(\prod_{j=1}^{J_i}\left([y_{ijk}|N_{ik}, \boldsymbol{\alpha}_k]\right)[N_{ik}|\lambda_{ik}]\right)[\boldsymbol{\alpha}_k][\boldsymbol{\beta}_k]\right)[\boldsymbol{\lambda}|\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_K, \boldsymbol{\Sigma}_{spp}][\boldsymbol{\Sigma}_{spp}]. \tag{12}$$

See Appendix A for the full statements of both the joint probit and Royle-Nichols occupancy models.

Community confounding

Restricted regression approach. We fit a restricted version of the each JSDM that orthogonalizes the fixed and random species effects. In the Royle-Nichols model, we express the species latent intensity and occupancy process conditionally as

$$\log(\boldsymbol{\lambda}) \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\delta} + (\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{\eta}, \tau^2\boldsymbol{I}), \tag{13}$$

$$\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{spp} \otimes \boldsymbol{I}_n), \tag{14}$$

where $\boldsymbol{P}_X$ is the projection matrix onto the column space of $\boldsymbol{X}$. Likewise, in the probit model we specify $\Phi(\boldsymbol{\psi}) = \boldsymbol{X}\boldsymbol{\delta} + (\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{\eta}$ and retain the same prior for $\boldsymbol{\eta}$ as in (14). This specification forces the random species effects to explain patterns in the community that are orthogonal to the fixed effects. The latent variables and fixed effects in the LVM can also be orthogonalized. Writing (11) in matrix form, we have

$$\Phi(\boldsymbol{\psi}_k) = \boldsymbol{X}'\boldsymbol{\delta}_k + \boldsymbol{L}\boldsymbol{\theta}_k, \tag{15}$$

where $\boldsymbol{X}$ and $\boldsymbol{L}$ are the matrices of covariates and latent variables vertically stacked across sites, respectively. If we assume common covariates across all K species, we can specify a restricted LVM as follows:

$$\Phi(\boldsymbol{\psi}_k) = \boldsymbol{X}'\boldsymbol{\delta}_k + (\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{L}\boldsymbol{\theta}_k. \tag{16}$$

However, if covariates differ by species, i.e., $\boldsymbol{X} = \boldsymbol{X}_k$, then the posterior distribution of latent variables will differ by species. To retain a common posterior distribution of latent variables across all species, the latent variables need to be orthogonalized against all covariates among the k species,

$$\boldsymbol{L}^R = \prod_{k=1}^{K}(\boldsymbol{I} - \boldsymbol{P}_{X_k})\boldsymbol{L}. \tag{17}$$

The specification of (17) is more restrictive than the orthogalization in the Royle-Nichols and probit model, and so we omit the LVM from our case study.

Hanks et al.[31] showed that the restricted (13) and unrestricted (9) generalized linear mixed models GLMM are reparameterizations of the same model and derived the following relationship between the unconditional $\boldsymbol{\delta}$ and conditional $\boldsymbol{\beta}$ fixed effects:

$$\boldsymbol{\delta} \equiv \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\eta}. \tag{18}$$

Using (18), one can easily sample both sets of fixed effects by fitting either the restricted or unrestricted parameterization. We can also sample the covariance structure of the restricted random species effect from either model fit by drawing samples from the distribution

Model	Data	β ESS Ratio	β Mean ESS	η ESS Ratio	η Mean ESS	$\mathrm{E}\big(R^2(\boldsymbol{x}_1)\vert \boldsymbol{Y}\big)$	Rejection rate
Normal	Unconfounded	18.69	5143	6.20	5670	0.04	0.01
Normal	Confounded	8.67	4219	5.79	4800	0.51	0.87
Probit	Unconfounded	1.73	959	1.08	534	0.04	0.00
Probit	Confounded	1.96	444	1.23	293	0.19	0.63
Royle-Nichols	Unconfounded	0.81	232	0.98	307	0.04	0.00
Royle-Nichols	Confounded	0.80	186	0.98	301	0.18	0.51

Table 1. Summary of simulations results. All results are averaged across 3 magnitudes of random species effects and 50 simulated datasets. ESS Ratio is the effective sample size of the restricted parameterizations over the unrestricted and the mean ESS is the average of the two. $\mathrm{E}\big(R^2(\boldsymbol{x}_1)|\boldsymbol{Y}\big)$ is the posterior mean R^2 of confounding for species 1 continuous habitat covariate. Rejection rate is the portion of times the the posterior mean p-value from overall F-test of a linear relationship between $\boldsymbol{x}_1$ and $\boldsymbol{\Delta}$ was below 0.05.

$$\boldsymbol{\Sigma}_{spp,R}^{-1} \sim \mathrm{Wishart}(\boldsymbol{S}\nu + \boldsymbol{\eta}'(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{\eta}, \nu + n). \tag{19}$$

Hence, regardless of which model is fit, we can obtain both the unconditional and conditional habitat effects as well as the unrestricted and restricted species covariance matrices.

Measuring confounding. Hefley et al.[27] showed how to assess confounding in SGLMM models by computing the Pearson correlation coefficient between each pair of covariates and eigenvectors from the spectral decomposition of the spatial covariance matrix. Likewise, Prates et al.[35] proposed a test for spatial confounding that can be calculated prior to model fitting. We propose another approach relevant to our method that aids in interpretation. We compute the coefficient of determination of each covariate for species k regressed on the estimated random species effects. Because the latent intensities are unknown, the coefficents of determination of all covariates are derived quantities and can be computed at each iteration of the MCMC algorithm:

$$R^{2^{(l)}}(\boldsymbol{x}_k) = \frac{SSR^{(l)}(\boldsymbol{x}_k)}{SST(\boldsymbol{x}_k)} = \frac{\left(\boldsymbol{\Delta}^{(l)}\hat{\boldsymbol{\theta}}^{(l)} - \bar{\boldsymbol{x}}_k\right)'\left(\boldsymbol{\Delta}^{(l)}\hat{\boldsymbol{\theta}}^{(l)} - \bar{\boldsymbol{x}}_k\right)}{(\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k)'(\boldsymbol{x}_k - \bar{\boldsymbol{x}}_k)}, \tag{20}$$

where $\bar{\boldsymbol{x}}_k = (\bar{x}_k, \ldots, \bar{x}_k)'$ is the mean of the covariate $\boldsymbol{x}_k$ for species k repeated n times, $\boldsymbol{\Delta}^{(l)} = \left(\boldsymbol{\eta}_1^{(l)}, \ldots, \boldsymbol{\eta}_K^{(l)}\right)$ is a matrix of the random species effects sampled for MCMC iteration l, and $\hat{\boldsymbol{\theta}}^{(l)}$ are estimated regression coefficients relating the estimated species intensities at iteration l to $\boldsymbol{x}_k$. The posterior mean $\mathrm{E}\big(R^2(\boldsymbol{x}_k)|\boldsymbol{Y}\big)$ provides a measure of community confounding for the covariate $\boldsymbol{x}_k$ and can help identify which fixed effects will vary between the unrestricted and restricted models. Furthermore, we can use the global F-test of the linear relationship between $\boldsymbol{x}_k$ and $\boldsymbol{\Delta}$ to determine if confounding exists.

Simulation study

We performed a simulation study to investigate the effects of community confounding and orthogonalization of the fixed and random species effects on model fitting. Specifically, we compared the effective sample sizes of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ for three different models for confounded and unconfounded data with unrestricted and restricted parameterizations. The effective sample size (ESS) is the number of independent MCMC samples of a quantity and is a metric for measuring the sampling efficiency of an MCMC algorithm. Higher ESS are preferable as posterior distributions of quantities of interest can be obtained in fewer iterations.

We considered three models: The joint probit occupancy model, joint Royle-Nichols model, and joint normal model, which is derived from the scenario where $\boldsymbol{\lambda}$ in the Royle-Nichols is known (e.g., species density data). For each model, 150 datasets were generated with the fixed and random species effects independent and another 150 datasets were generated with confounding between the fixed and random species effects. To induce confounding between the fixed and random species effect, we expressed one covariate of the first species as a linear combination of the random species effects (i.e., $\boldsymbol{x}_1 = \boldsymbol{\Delta\theta}$).

Because the ratio of the random effects and random error magnitude is known to affect the severity of confounding in the spatial context[29,31,35], we varied the magnitude of the random species effect in each model while holding the random error magnitude constant. Specifically, each dataset was subdivided into thirds with 50 datasets simulated to have small, medium, and large random species effects relative to the random error.

All 900 simulated datasets across models and confounding levels were for $K = 2$ species across $n = 50$ sites with $J = 10$ occasions per site for the occupancy models. The correlation between the two species was allowed to vary for each dataset. Each habitat design matrix included an intercept and one continuous covariate. Each MCMC algorithm was run for a burn-in period of $L = 10000$ to ensure convergence. The next $L = 10000$ iterations were used to calculate the posterior quantities in Table 1. Code for performing the simulation study in R are available in the supplementary electronic files.

For both $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$, ESS was lower on average for the confounded data than the unconfounded data for all three models demonstrating the negative impacts confounding can have on model fitting. For all three models, the computational impact of fitting the restricted parameterization did not differ depending on whether confounding

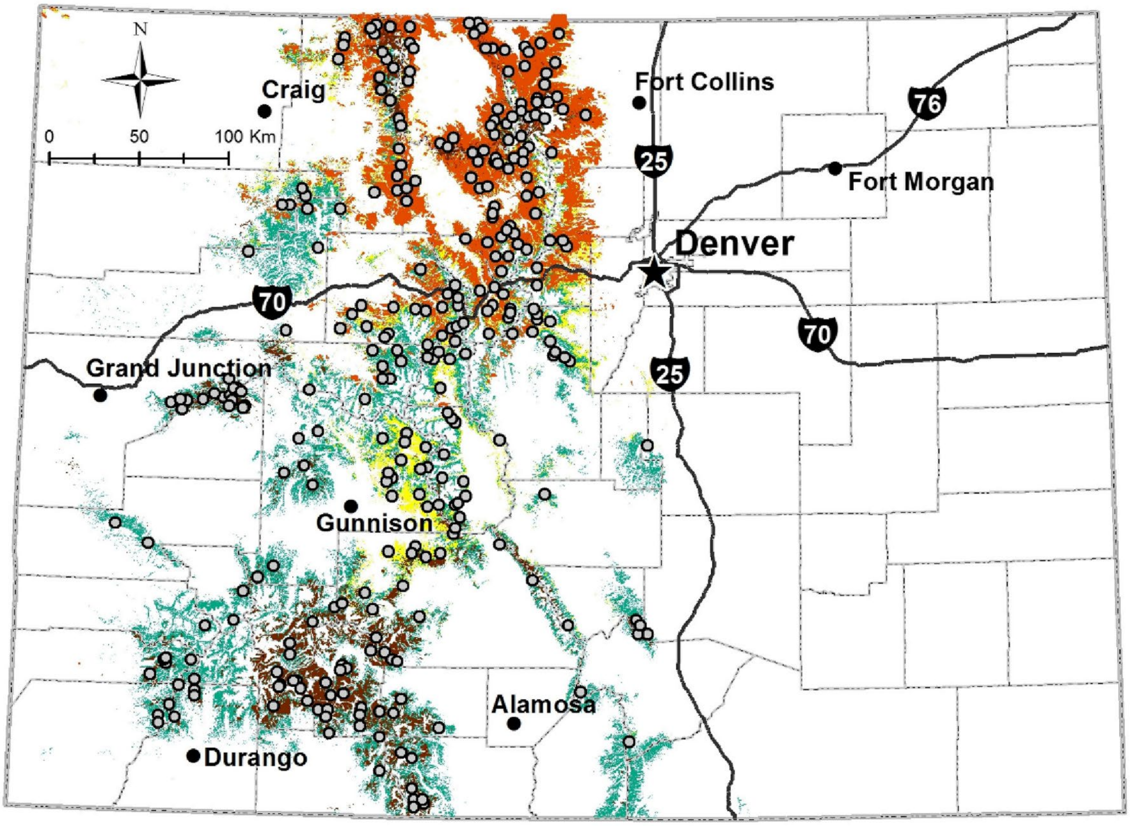

Figure 1. Randomly selected sampling sites (gray circles) where passive infrared game cameras were deployed in spruce-fir (green) and lodgepole pine (yellow) forests in Colorado, USA, 2013-2014. Brown and orange are the approximate extents of spruce beetle and mountain pine beetle impacts in spruce-fir and lodgepole pine forests, respectively, as of 2014. Reprinted from "Mammalian responses to changed forest conditions resulting from bark beetle outbreaks in the southern Rocky Mountains," by J. S. Ivan, 2018, Ecosphere, 9(8), e02369. Figure produced with ArcMap 10.7 available at: https://desktop.arcgis.com/en/arcmap/10.7/get-started/setup/arcgis-desktop-quick-start-guide.html.

exists or not. In the case of the normal and probit models, fitting the restricted parameterization improved ESS for both $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$, although the gains were much greater for the normal model. On the other hand, the restricted parameterization of the Royle-Nichols model did not improve ESS for $\boldsymbol{\beta}$ or $\boldsymbol{\eta}$. The success of our method for detecting community confounding differed across models. The method was most powerful for the normal model followed by the probit and Royle-Nichols models.

Camera trap survey

Study area. We analyzed data arising from a study area comprised of subalpine forests in the state of Colorado between 2590 and 3660 m elevation (Fig. 1). Sites were restricted to public lands managed by the United States Forest Service, National Park Service, Bureau of Land Management, and Colorado State Forest Service. Forests in our study area were primarily composed of Lodgepole pine (*Pinus contorta*), Engelmann spruce (*Picea engelmannii*), and subalpine fir (*Abies lasiocarpa*). Lodgepole pine was dominant at lower elevations as well as higher elevations that were drier and/or on south-facing slopes; high elevation regions that had cool north-facing slopes were co-dominated by Engelmann spruce and subalpine fir. Lodgepole pine is restricted to the northern two-thirds of Colorado, so all sites in the southern region of the study area were Engelmann spruce, subalpine fir co-dominated. Quaking aspen (*Populus tremuloides*), Douglas-fir (*Pseudotsuga menziesii*), bristlecone pine (*Pinus aristata*), limber pine (*Pinus flexilis*), and blue spruce (*Picea pungens*) were also present at some sites. Mean July and January temperature across the study area were 14 and − 6.1 °C respectively. All camera data were collected during summers 2013–2014.

Sampling design. The primary goal of Ivan et al.[50] was to assess mammalian responses to bark beetle outbreaks, thus sites were randomly selected to facilitate inference on the beetle outbreak covariates. Beetle outbreak covariates included the number of years since the initial outbreak (YSO) and the severity of the outbreak measured by mean overstory mortality (severity). The sample of $n = 300$, 1 km^2 sites was evenly split across the two dominant forest types, spruce-fir and lodgepole pine. Additional environmental covariates were collected at each site, and a description of these is included in Appendix B.

Passive infrared camera traps (Reconyx PC800, Holmen, Wisconsin, USA) were deployed near the center of each site. Cameras were approximately 0.5 m above the ground and pointed toward a lure tree 4–5 m away[52]. The setup was designed to maximize detections of both large and small-bodied mammals in the local community while minimizing attraction of individuals from outside the sampling region of the site. The sampling regions were likely not closed to immigration/emigration; thus, we interpret elevated detections at a site as more individuals using, as opposed to occupying, that site[53]. For additional details regarding the sampling design and study area see Ivan et al.[50].

Model fitting. We fit both the Royle-Nichols and probit occupancy models to the camera trap data binned into 20 two-day occasions because simulations showed this was the number of replications needed to identify a quadratic effect of occasion on individual detection probability. Not all cameras were operational for the entire 40 day sampling period, and thus the number of occasions varied from 7-20. We discarded four sites at which the camera was operational for less than one occasion. We also discarded another 12 sites that had been infested by bark beetles for more than 10 years. Ivan et al.[50] truncated the bark beetle infestation covariate at 10 years because estimates of response curves beyond 10 years would be unreliable with so few sites. The final sample size was $n = 284$ sites. We built distribution models for the 13 species for which Ivan et al.[50] performed a single species analysis; several rare species were excluded from analysis due to insufficient detections. We note, however, that these rare species parameters may be identifiable in the joint model as has been the case in previous studies[2,47,54–57]. Our final dataset then included 3692 unique encounter histories at $n = 284$ sites, stacked across $K = 13$ species.

Ivan et al.[50] used a sequential procedure similar to that described in Lebreton et al.[58] to select the covariates in the occupancy and detection processes for each species. We adopted their detection model and used the same covariates but a different set of basis functions for YSO. Ivan et al.[50] treated YSO as a grouping variable and considered probability of use response curves that allowed for cubic associations and delayed responses to bark beetle infestation. Multiple response curves were model averaged to produce predictive YSO response curves for each species. We used orthogonal polynomial basis functions for the YSO variable in the species intensity models. The basis functions included a linear (YSO1) and quadratic (YSO2) effect. Appendix E provides a full description of the intensity and detection models. All continuous covariates were scaled to have mean 0 and variance 1.

We fit all models using Markov chain Monte Carlo (MCMC). To improve mixing and predictive ability, we regularized the coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ with informative priors: $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$[26]. We specified a vague prior of $\boldsymbol{\Sigma}_{spp}^{-1} \sim \text{Wishart}(15, (15\boldsymbol{I})^{-1})$ for the species variance-covariance matrix[59]. For the Royle-Nichols model, we used Gibbs sampling based on conjugate priors for parameters $\boldsymbol{\Sigma}_{spp}$, $\boldsymbol{\eta}$, and $\boldsymbol{\beta}$ and Metropolis-Hastings updates for $\boldsymbol{N}$, $\boldsymbol{\lambda}$, and $\boldsymbol{\alpha}$. Derivations of the conjugate full-conditional distributions are provided in Appendix C with details about the Metropolis-Hastings updates. We tuned the Metropolis-Hastings updates so that acceptance rates varied between 20 and 40% for $\boldsymbol{\alpha}$, $\boldsymbol{N}$, and $\boldsymbol{\lambda}$. Using data augmentation[17,43–45], all the parameters of the probit model can be sampled with Gibbs updates.

We set $\tau^2 = 2.25$ in both (9) and (13). This choice was supported by the asymptotic equivalence between Poisson and logistic regression. In a generalized occupancy model, the latent probability of occupancy is specified as $\text{logit}(\psi_i) \sim \mathcal{N}(\boldsymbol{x}_i'\boldsymbol{\beta}, \tau^2)$. Hanson et al.[60] investigated the relationship between the prior on $\boldsymbol{\beta}$ and induced prior on the latent probability of success ψ_i in logistic regression; their work showed that specifying an uninformative normal prior on $\boldsymbol{\beta}$ (i.e., setting τ^2 large) induces a U-shaped prior for ψ_i with most of the density concentrated near 0 and 1. Broms et al.[14] recommended setting $\tau^2 = 2.25$ in occupancy models, which results in a relatively flat prior for ψ. For rare species, λ_i in (9) and (13) is analogous to ψ_i, and specifying a variance of $\tau^2 = 2.25$ is minimally informative.

Baddeley[61] motivated the asymptotic equivalence of Poisson and logistic regression in a spatial context where counts of points from a non-homogeneous Poisson process are recorded in a lattice; they showed that, as the grid cells of the lattice become infinitesimally small, the inference yielded from Poisson and logistic regression are equivalent. This result can be applied more generally to any dataset where there is a high proportion of zero counts. We demonstrate the asymptotic equivalence between Poisson and logistic regression in the Royle-Nichols model in Appendix D.

We ran the MCMC algorithm for $L = 50000$ iterations, and discarded the first 12500 iterations as burn-in. We fit an SDM, unrestricted JSDM, and restricted JSDM of both the Royle-Nichols and probit occupancy models. The "Results" section presents inference for regression coefficients for all six model fits.

Results. Ivan et al.[50] fit SDMs to infer changes in mammalian use of stands impacted by the bark beetle epidemic. The impact of bark beetle damage was measured by years since initial infestation (YSO) and severity of outbreak quantified by mean overstory mortality (DeafConif). The posterior distributions of the regression coefficients varied between the probit SDM and unrestricted JSDM, although the magnitude of difference differed by species (Fig. 2). The posterior variances of the SDM regression coefficients were smaller than the unrestricted JSDM. Posterior variances and means of the restricted probit JSDM regression coefficients quite similar to those from unrestricted JSDM. The only noticeable difference between the unrestricted and restricted regression coefficients was that the restricted coefficients had slightly smaller posterior variances on average.

As with the probit modeling results, posterior distributions of the regression coefficients in the Royle-Nichols SDM were more concentrated near zero than those of the JSDM (Fig. 3). Also, posterior distributions of the restricted JSDM regression coefficients were slightly tighter and centered closer to zero.

We calculated the unrestricted and restricted posterior correlation matrices for both the probit and Royle-Nichols models. Pairwise differences between each entry of the posterior mean of the four correlation matrices were bounded between $(-0.2, 0.2)$, so only the correlation matrix of the unrestricted Royle-Nichols model

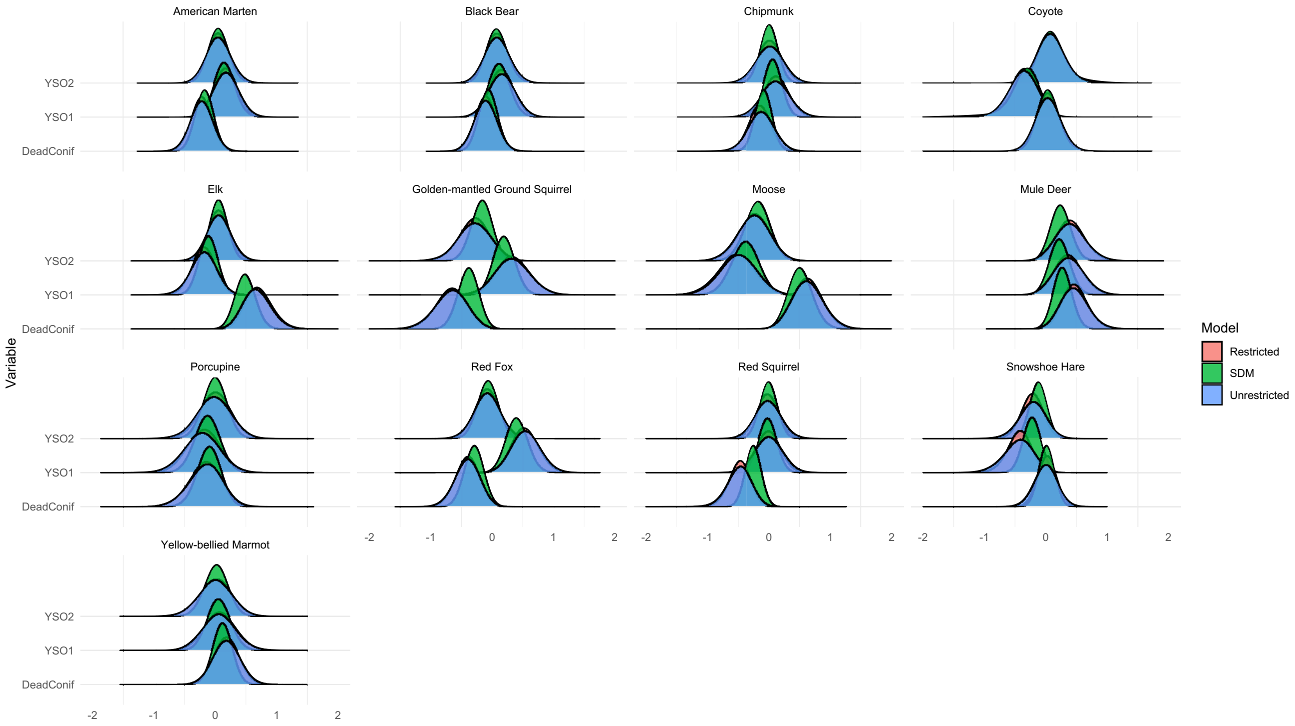

Figure 2. Marginal posterior distributions of infestation regression parameters. Posterior distributions shown are from the probit SDM, unrestricted JSDM, and restricted JSDM. DeadConif is the overstory mortality percentage, a proxy for severity of bark beetle infestation. YSO1 is the linear effect of the number of years since a site was infested with bark beetles. YSO2 is the quadratic effect. Figure created in R 4.1.2[62].

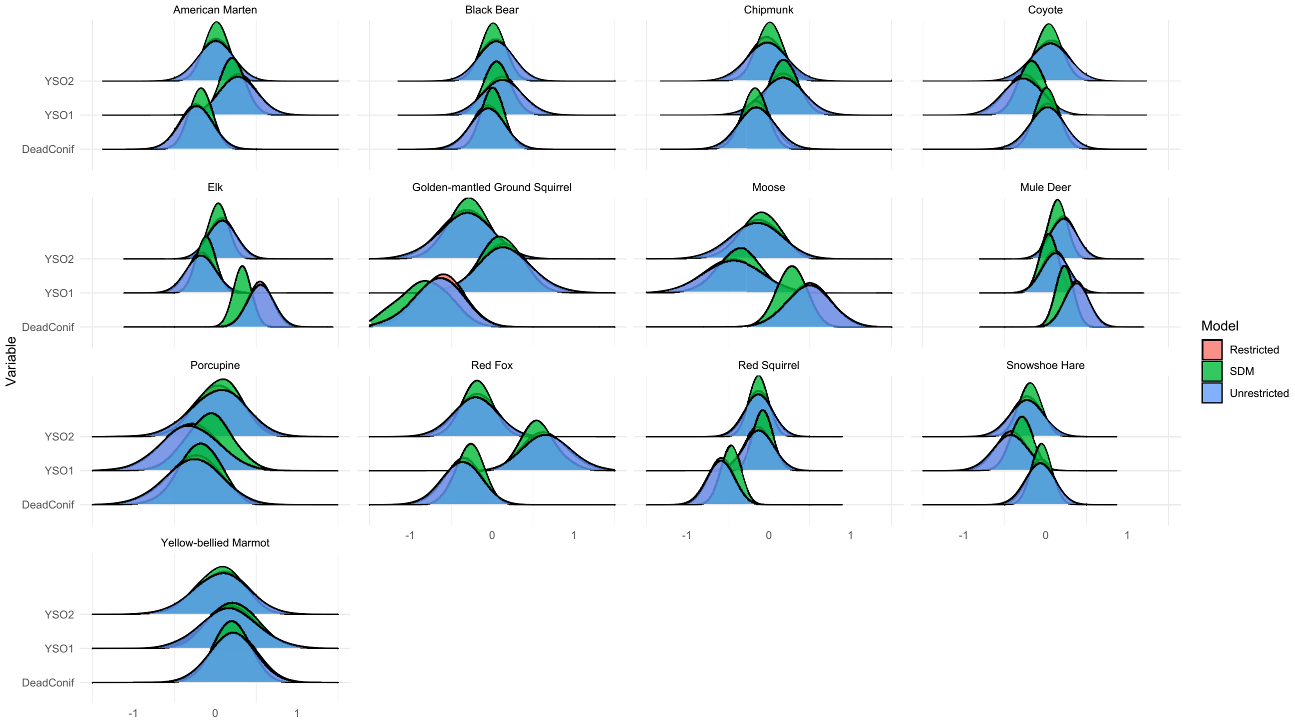

Figure 3. Marginal posterior distributions of infestation regression parameters. Posterior distributions shown are from the Royle-Nichols SDM, unrestricted JSDM, and restricted JSDM. DeadConif is the overstory mortality percentage, a proxy for severity of bark beetle infestation. YSO1 is the linear effect of the number of years since a site was infested with bark beetles. YSO2 is the quadratic effect. Figure created in R 4.1.2[62].

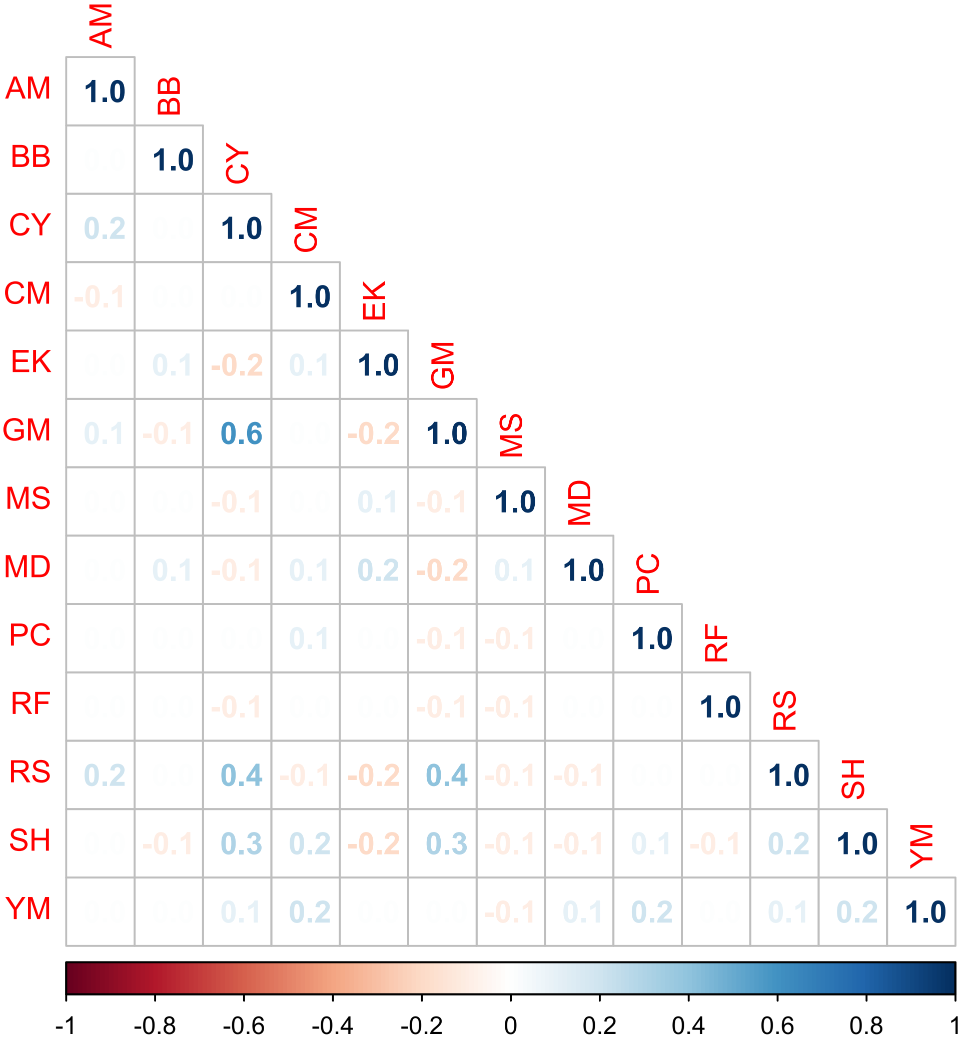

Figure 4. Posterior mean of species correlation matrix. Estimates are from the Royle-Nichols unrestricted joint species distribution model. AM = American Marten, BB = Black Bear, CY = Coyote, CM = Chipmunk spp., Ek = Elk, GM = Golden-mantled Ground Squirrel, MS = Moose, MD = Mule Deer, PC = Porcupine, RF = Red Fox, RS = Red Squirrel, SH = Snowshoe Hare, YM = Yellow-bellied Marmot. Figure created in R 4.1.2[62].

is shown (Fig. 4). The posterior distributions of the pairwise correlations all overlapped zero except for the pairwise correlations between coyotes and golden-mantled ground squirrels, coyotes and red squirrels, and golden-mantled ground squirrels and red squirrels. In the restricted probit JSDM, the correlations between coyotes and snowshoe hares, and snowshoe hares and red squirrels also did not overlap zero.

We calculated the posterior R^2 of confounding for each covariate in each species specific model as described in (20). All posterior R^2 were below 0.05 for both the Royle-Nichols and probit models giving no indication of community confounding for all covariates considered.

Discussion

We found that confounding between the fixed and random species effects can reduce sampling efficiency in MCMC algorithms and that orthogonalizing the fixed and random species effects can alleviate this problem when fitting some joint species distribution models. In the simulation study, we discovered that, even when the data were not confounded, orthogonalizing the fixed and random species effects still conferred a computational benefit for the normal and probit model. This was also true for our case study where the mean effective sample size of the conditional habitat effects $\boldsymbol{\beta}$ in the probit model was 32% larger when fit with the restricted parameterization. The effective sample size of $\boldsymbol{\eta}$ in the probit model was 3% greater for the restricted parameterization.

The case study indicated that inference on species-environment associations in occupancy models can change based on whether the distribution model accounts for community structure. Orthogonalizing the fixed and random species effects in the probit and Royle-Nichols model slightly reduced but did not nullify the differences as in the case for normal data. The similarity between the restricted and unrestricted JSDM coupled with the lack of evidence for community confounding suggests additional mechanisms lead inference in SDMs and JSDMs to differ, a finding consistent with Caradima et al.[63]. Overall, there was still large agreement in posterior inference produced by the SDM and JSDMs for both occupancy models. In additional simulation studies on the probit and Rolye-Nichols occopancy models, we found that community confounding can lead to larger differences between the SDM and unrestricted JSDM and that the restricted JSDM again mitigates but rarely nullifies these differences.

We were also interested in whether the Royle-Nichols model could identify additional associations compared with the probit model. The Royle-Nichols model measures associations conditional on an intensity process rather than an occupancy process, and intensity is likely a function of additional factors beyond those influencing occupancy[19–21]. For the camera trap data, the opposite was true, in that the probit model identified more environmental-species and species-species associations. One possible explanation for this is that the probit model is more parsimonious which sharpens posterior distributions.

A related method to restricted regression, which orthogonalizes the fixed and random effects, is principal components regression, which performs an orthogonalization procedure solely among the fixed effects. To motivate their similarities, consider a simpler case where the latent intensities, $\boldsymbol{\lambda}$, of the K species in our community were known. We could construct K regression models for predicting each species intensity as follows:

$$\boldsymbol{\lambda}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{\Lambda}_{-k}\boldsymbol{\eta}_k + \boldsymbol{\varepsilon}, \tag{21}$$

where $\boldsymbol{\Lambda}_{-k} = \left(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_{k-1}, \boldsymbol{\lambda}_{k+1}, \ldots, \boldsymbol{\lambda}_K\right)$ is a matrix of the $K-1$ other species intensities. If $\boldsymbol{X}_k$ and $\boldsymbol{\Lambda}_{-k}$ were highly collinear, principal component regression might be applied. Principal components regression is so named because it decomposes the variation explained by $\boldsymbol{X}_k$ and $\boldsymbol{\Lambda}_{-k}$ into $p = p_1 + p_2$ principal components, $\boldsymbol{\Gamma}_k = \left(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p\right)$, where p_1 and p_2 are the number of columns of $\boldsymbol{X}_k$ and $\boldsymbol{\Lambda}_{-k}$ respectively. The p principal components retain all the information explained by $\boldsymbol{X}_k$ and $\boldsymbol{\Lambda}_{-k}$ but are orthogonal. The regression model

$$\boldsymbol{\lambda}_k = \boldsymbol{W}_k\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{22}$$

$$\boldsymbol{W}_k = \left(\boldsymbol{X}_k, \boldsymbol{\Lambda}_{-k}\right)\boldsymbol{\Gamma}_k, \tag{23}$$

often improves sampling efficiency and can recover the posterior means and variances of $\boldsymbol{\beta}_k$ and $\boldsymbol{\eta}_k$ in (21). However, inference on $\boldsymbol{\beta}_k$ and $\boldsymbol{\eta}_k$ is often adjusted by truncating off the last $p-r$, for $r < p$, eigenvectors of $\boldsymbol{\Gamma}_k$ and employing the new design matrix

$$\boldsymbol{W}_k^\star = \left(\boldsymbol{X}_k, \boldsymbol{\Lambda}_{-k}\right)\boldsymbol{\Gamma}_k^\star, \tag{24}$$

$$\boldsymbol{\Gamma}_k^\star = \left(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_r\right). \tag{25}$$

By retaining only the first r principal components, the smallest sources of variation are ignored in the estimation of $\boldsymbol{\beta}_k$ and $\boldsymbol{\eta}_k$. Jeffers[64] implemented this approach truncating off the last 7 of 13 principal components to adjust the estimates of regression coefficients relating various tree characteristics to maximum compressive strength. Other studies have selected a subset of principal components based on their strength of association with the response variable[65–68]. In some cases, the coefficient estimates from these reduced rank approaches appeared more tenable than those from the full rank specifications based on known physical relationships between the predictors and response. Thus, like restricted regression, principal components regression can be used for solely computation purposes or to adjust inference.

Recently, concerns regarding the coverage properties of the fixed effects estimator under restricted regression have been expressed[36,69]. For example, Zimmerman and Ver Hoef[69] showed that applying any restricted regression method to a SGLMM leads to frequentest coverage of the fixed effects that is lower than the corresponding non-spatial model. Similarly, Khan and Calder[36] found that when fitting a restricted version of the SGLMM with an intrinsic conditional autoregressive prior, credible intervals of the fixed effects from the restricted model were generally nested inside those yielded by the non-spatial model. Given these results, both Zimmerman and Ver Hoef[69] and Khan and Calder[36] recommended reverting to inference from the non-spatial model, rather than that of the restricted SGLMM, when inference from the unrestricted SGLMM appears untenable.

We did not observe the same pattern in our restricted JSDM but found the length of credible intervals of the restricted regression coefficients to generally be between that of the SDM and unrestricted JSDM. Nonetheless, if higher coverage is desired, one can always extract the conditional coefficients from the restricted JSDM while still benefiting from the increased stability that results from orthogonalizing the fixed and random effets. When deciding between inference from the restricted and unrestricted JSDM, one should also consider the random species effects $\boldsymbol{\eta}$. Because the random effect $\boldsymbol{\eta}$ is rarely of interest in spatial applications, there has been little investigation on the inferential impacts of restricted regression on $\boldsymbol{\eta}$. Such investigation, however, may be helpful in determining the appropriateness of restricted regression for JSDMs.

There are several conceptual facets to consider regarding the application of restricted regression in joint species distribution modeling. Frequently, JSDMs are described as accounting for residual correlations between species that cannot be explained by the environmental covariates[4,5,63]. We have shown, however, that in some JSDMs, the random species effect can explain variation that is collinear with environmental covariates. Only in the restricted JSDM, does the random species effect explain variation that is residual to the environmental covariates attributing all contested sources of variation to the fixed effect. Yet, given that species environmental requirements can fluctuate based on their symbiotic relationships, one might argue that interplay between the environmental effects and interspecies dependence is ecology warranted. Therefore, any method that removes the conditional nature of these effects like restricted regression is inappropriate.

JSDMs have been described as correcting our knowledge of species-environment relationship by accounting for interspecies dependence[5]. Poggiato et al.[5] argued that JSDMs help us better quantify uncertainty regarding species-environment relationships, but they cannot explain discrepancies in a species theoretical and realized niche. We agree that phenomenological JSDMs should not be used to disentangle the marginal effects of environment and interspecies dependence on species distributions and would recommend the development of mechanistic models to investigate interspecies-environment associations.

Experimental methods and modeling techniques for alleviating confounding have been proposed in ecology. Hefley et al.[70] showed that replicate populations can help disentangle confounded fixed and random effects. In the context of joint species distribution modeling, replication involves analyzing several communities simultaneously, which is often infeasible. Hefley et al.[70] also recommended explicit population models rather than phenomenological regression-based models for analysis of temporally confounded count data. Similarly, Fieberg et al.[71] advocated for mechanistic models guided by causal diagrams for analyzing temporally confounded animal movement data. An avenue of future research for joint species distribution modeling is to compare inference from phenomenological regression-based models, such as the one proposed here, with that of models that explicitly include ecological mechanisms such as competitive exclusion, mutualism, and predation. Because community and temporal confounding have the same mathematical framework, mechanistic models are a promising solution for confounded multispecies data.

In summary, we specified a JSDM that accounts for interspecies dependence at the intensity level, and examined how inference from the joint model differed from the joint probit model. We performed a simulation study on three JSDMs to examine the computational difficulties associated with community confounding and investigated whether orthogonalizing the fixed and random species effect could alleviate these difficulties. Further, we considered how inference in both occupancy models differed depending on the assumed community structure. Lastly, we discussed how joint species distribution modeling is distinct from spatial and time series applications in that the random effect is almost always of inferential interest, and hence, adjustments to the regression coefficients, $\boldsymbol{\beta}$, and random effects, $\boldsymbol{\eta}$, should both be considered. Our main conclusion is that, even for researchers who desire inference solely on the conditional relationship between the fixed species-environment and random species effects, fitting the JSDM with a restricted parameterization can give computational benefits.

Data availability

The data are available in the Supplementary Information files.

Code availability

All algorithms and code for fitting and analyzing results from the six model variants are available in the Supplementary Information files. All MCMC algorithms and analyses were coded in R 4.1.2[62].

Received: 24 November 2021; Accepted: 28 June 2022
Published online: 18 July 2022

References

1. Altwegg, R. & Nichols, J. D. Occupancy models for citizen-science data. *Methods Ecol. Evol.* **10**, 8–21 (2019).
2. Hui, F., Warton, D., Foster, S. & Dunstan, P. To mix or not to mix: Comparing the predictive performance of mixture models vs. separate species distribution models. *Ecology* **94**, 1913–1919. https://doi.org/10.1890/12-1322.1 (2013).
3. Warton, D. *et al.* So many variables: Joint modeling in community ecology. *Trends Ecol. Evol.* **30**, 766–779 (2015).
4. Tobler, M. W. *et al.* Joint species distribution models with species correlations and imperfect detection. *Ecology* **100**, e02754. https://doi.org/10.1002/ecy.2754 (2019).
5. Poggiato, G. *et al.* On the interpretations of joint modeling in community ecology. *Trends Ecol. Evol.* **36**, 391–401 (2021).
6. Estevo, C. A., Nagy-Reis, M. B. & Nichols, J. D. When habitat matters: Habitat preferences can modulate co-occurrence patterns of similar sympatric species. *PLoS One* **12**, e0179489 (2017).
7. Steen, D. A. *et al.* Snake co-occurrence patterns are best explained by habitat and hypothesized effects of interspecific interactions. *J. Anim. Ecol.* **83**, 286–295 (2014).
8. Wisz, M. S. *et al.* The role of biotic interactions in shaping distributions and realised assemblages of species: Implications for species distribution modelling. *Biol. Rev.* **88**, 15–30 (2013).
9. Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R. & McCarthy, M. A. A comparison of joint species distribution models for presence-absence data. *Methods Ecol. Evol.* **10**, 198–211 (2019).
10. MacKenzie, D. I. *et al.* Estimating site occupancy rates when detection probabilities are less than one. *Ecology* **83**, 2248–2255 (2002).
11. Dorazio, R. M. & Royle, J. A. Estimating size and composition of biological communities by modeling the occurrence of species. *J. Am. Stat. Assoc.* **100**, 389–398 (2005).
12. Dorazio, R. M., Royle, J. A., Söderström, B. & Glimskär, A. Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology* **87**, 842–854 (2006).
13. Tobler, M. W., Zúñiga Hartley, A., Carrillo-Percastegui, S. E. & Powell, G. V. N. Spatiotemporal hierarchical modelling of species richness and occupancy using camera trap data. *J. Appl. Ecol.* **52**, 413–421. https://doi.org/10.1111/1365-2664.12399 (2015).
14. Broms, K. M., Hooten, M. B. & Fitzpatrick, R. M. Model selection and assessment for multi-species occupancy models. *Ecology* **97**, 1759–1770. https://doi.org/10.1890/15-1471.1 (2016).
15. Rota, C. T. *et al.* A multispecies occupancy model for two or more interacting species. *Methods Ecol. Evol.* **7**, 1164–1173. https://doi.org/10.1111/2041-210X.12587 (2016).
16. Maphisa, D. H., Smit-Robinson, H. & Altwegg, R. Dynamic multi-species occupancy models reveal individualistic habitat preferences in a high-altitude grassland bird community. *PeerJ* **7**, e6276 (2019).
17. Johnson, D. S., Conn, P. B., Hooten, M. B., Ray, J. C. & Pond, B. A. Spatial occupancy models for large data sets. *Ecology* **94**, 801–808 (2013).
18. Royle, J. & Nichols, J. Estimating abundance from repeated presence-absence data or point counts. *Ecology* **84**, 777–790 (2003).
19. Orrock, J. L., Pagels, J. F., McShea, W. J. & Harper, E. K. Predicting presence and abundance of a small mammal species: The effect of scale and resolution. *Ecol. Appl.* **10**, 1356–1366 (2000).

20. Cingolani, A. M., Cabido, M., Gurvich, D. E., Renison, D. & Díaz, S. Filtering processes in the assembly of plant communities: Are species presence and abundance driven by the same traits?. *J. Veg. Sci.* **18**, 911–920 (2007).
21. Dibner, R. R., Doak, D. F. & Murphy, M. Discrepancies in occupancy and abundance approaches to identifying and protecting habitat for an at-risk species. *Ecol. Evol.* **7**, 5692–5702. https://doi.org/10.1002/ece3.3131 (2017).
22. Bascompte, J. Mutualistic networks. *Front. Ecol. Environ.* **7**, 429–436 (2009).
23. Van Dam, N. How plants cope with biotic interactions. *Plant Biol.* **11**, 1–5 (2009).
24. Clark, A. E. & Altwegg, R. Efficient Bayesian analysis of occupancy models with logit link functions. *Ecol. Evol.* **9**, 756–768 (2019).
25. Broms, K. M., Johnson, D. S., Altwegg, R. & Conquest, L. L. Spatial occupancy models applied to atlas data show Southern Ground Hornbills strongly depend on protected areas. *Ecol. Appl.* **24**, 363–374 (2014).
26. Hooten, M. B. & Hobbs, N. T. A guide to Bayesian model selection for ecologists. *Ecol. Monogr.* **85**, 3–28. https://doi.org/10.1890/14-0661.1 (2015).
27. Hefley, T. J., Hooten, M. B., Hanks, E. M., Russell, R. E. & Walsh, D. P. The Bayesian group lasso for confounded spatial data. *J. Agric. Biol. Environ. Stat.* **22**, 42–59 (2017).
28. Reich, B. J., Hodges, J. S. & Zadnik, V. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics* **62**, 1197–1206 (2006).
29. Hodges, J. S. & Reich, B. J. Adding spatially-correlated errors can mess up the fixed effect you love. *Am. Stat.* **64**, 325–334. https://doi.org/10.1198/tast.2010.10052 (2010).
30. Hughes, J. & Haran, M. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *J. R. Stat. Soc. Ser. B (Statistical Methodology)* **75**, 139–159. https://doi.org/10.1111/j.1467-9868.2012.01041.x (2013).
31. Hanks, E. M., Schliep, E. M., Hooten, M. B. & Hoeting, J. A. Restricted spatial regression in practice: Geostatistical models, confounding, and robustness under model misspecification. *Environmetrics* **26**, 243–254. https://doi.org/10.1002/env.2331 (2015).
32. Bradley, J. R. *et al.* Multivariate spatio-temporal models for high-dimensional areal data with application to longitudinal employer-household dynamics. *Ann. Appl. Stat.* **9**, 1761–1791 (2015).
33. Murakami, D. & Griffith, D. A. Random effects specifications in eigenvector spatial filtering: A simulation study. *J. Geogr. Syst.* **17**, 311–331 (2015).
34. Thaden, H. & Kneib, T. Structural equation models for dealing with spatial confounding. *Am. Stat.* **72**, 239–252 (2018).
35. Prates, M. O. *et al.* Alleviating spatial confounding for areal data problems by displacing the geographical centroids. *Bayesian Anal.* **14**, 623–647 (2019).
36. Khan, K. & Calder, C. A. Restricted spatial regression methods: Implications for inference. *J. Am. Stat. Assoc* 1–13 (2020).
37. Paciorek, C. The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Stat. Sci. A Rev. J. Inst. Math. Stat.* **25**, 107–125. https://doi.org/10.1214/10-STS326 (2010).
38. Dominici, F., McDermott, A. & Hastie, T. J. Improved semiparametric time series models of air pollution and mortality. *J. Am. Stat. Assoc.* **99**, 938–948 (2004).
39. Houseman, E. A., Coull, B. A. & Shine, J. P. A nonstationary negative binomial time series with time-dependent covariates: Enterococcus counts in Boston Harbor. *J. Am. Stat. Assoc.* **101**, 1365–1376 (2006).
40. Corbeil, R. R. & Searle, S. R. Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics* **18**, 31–38 (1976).
41. Hoeting, J. A., Leecaster, M. & Bowden, D. An improved model for spatially correlated binary responses. *J. Agric. Biol. Environ. Stat.* **5**, 102–114 (2000).
42. Tyre, A. J. *et al.* Improving precision and reducing bias in biological surveys: Estimating false-negative error rates. *Ecol. Appl.* **13**, 1790–1801. https://doi.org/10.1890/02-5078 (2003).
43. Albert, J. H. & Chib, S. Bayesian analysis of binary and polychotomous response data. *J. Am. Stat. Assoc.* **88**, 669–679 (1993).
44. Hooten, M. B., Larsen, D. R. & Wikle, C. K. Predicting the spatial distribution of ground flora on large domains using a hierarchical Bayesian model. *Landsc. Ecol.* **18**, 487–502 (2003).
45. Dorazio, R. M. & Rodriguez, D. T. A Gibbs sampler for Bayesian analysis of site-occupancy data. *Methods Ecol. Evol.* **3**, 1093–1098 (2012).
46. Clark, J. S. *et al.* High-dimensional coexistence based on individual variation: A synthesis of evidence. *Ecol. Monogr.* **80**, 569–608 (2010).
47. Ovaskainen, O. & Soininen, J. Making more out of sparse data: Hierarchical modeling of species communities. *Ecology* **92**, 289–295 (2011).
48. Scheffe, H. *The Analysis of Variance* Vol. 72 (John Wiley & Sons, New Jersey, 1959).
49. Hodges, J. S. & Clayton, M. K. Random effects old and new. *Stat. Sci.* (2011).
50. Ivan, J., Seglund, A., Truex, R. & Newkirk, E. Mammalian responses to changed forest conditions resulting from bark beetle outbreaks in the southern Rocky Mountains. *Ecosphere*https://doi.org/10.1002/ecs2.2369 *(2018).*
51. Chan, J.C.-C. & Jeliazkov, I. MCMC estimation of restricted covariance matrices. *J. Comput. Graph. Stat.* **18**, 457–480 (2009).
52. Blecha, K. A. Risk-reward tradeoffs in the foraging strategy of cougar (puma concolor): Prey distribution, anthropogenic development, and patch selection. Thesis, Colo. State Univ. Fort Collins, Color. USA (2015).
53. MacKenzie, D. I. *et al. Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence* (Academic Press, USA, 2006).
54. Guisan, A., Weiss, S. & Weiss, A. GLM versus CCA spatial modeling of plant species distribution. *Plant Ecol.* **143**, 107–122. https://doi.org/10.1023/A:1009841519580 (1999).
55. Madon, B., Warton, D. I. & Araújo, M. B. Community-level vs species-specific approaches to model selection. *Ecography* **36**, 1291–1298. https://doi.org/10.1111/j.1600-0587.2013.00127.x (2013).
56. Ovaskainen, O., Abrego, N., Halme, P. & Dunson, D. Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods Ecol. Evol.* **7**, 549–555. https://doi.org/10.1111/2041-210X.12501 (2015).
57. Tikhonov, G., Abrego, N., Dunson, D. & Ovaskainen, O. Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. *Methods Ecol. Evol.* **8**, 443–452. https://doi.org/10.1111/2041-210X.12723 (2017).
58. Lebreton, J.-D., Burnham, K. P., Clobert, J. & Anderson, D. R. Modeling survival and testing biological hypotheses using marked animals: A unified approach with case studies. *Ecol. Monogr.* **62**, 67–118. https://doi.org/10.2307/2937171 (1992).
59. Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J. & Dorie, V. Weakly informative prior for point estimation of covariance matrices in hierarchical models. *J. Educ. Behav. Stat.* **40**, 136–157. https://doi.org/10.3102/1076998615570945 (2015).
60. Hanson, T. E. *et al.* Informative *g*-priors for logistic regression. *Bayesian Anal.* **9**, 597–612 (2014).
61. Baddeley, A. *et al.* Spatial logistic regression and change-of-support in poisson point processes. *Electron. J. Stat.* **4**, 1151–1201. https://doi.org/10.1214/10-EJS581 (2010).
62. R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria (2021).
63. Caradima, B., Schuwirth, N. & Reichert, P. From individual to joint species distribution models: A comparison of model complexity and predictive performance. *J. Biogeogr.* **46**, 2260–2274 (2019).
64. Jeffers, J. N. Two case studies in the application of principal component analysis. *J. R. Stat. Soc. Ser. C (Applied Statistics)* **16**, 225–236 (1967).

65. Hill, C. R., Fomby, T. B. & Johnson, S. R. Component selection norms for principal components regression. *Commun. Stat. Methods* **6**, 309–334 (1977).
66. Kung, E. C. & Sharif, T. A. Regression forecasting of the onset of the Indian summer monsoon with antecedent upper air conditions. *J. Appl. Meteorol. Climatol.* **19**, 370–380 (1980).
67. Smith, G. & Campbell, F. A critique of some ridge regression methods. *J. Am. Stat. Assoc.* **75**, 74–81 (1980).
68. Jolliffe, I. T. A note on the use of principal components in regression. *J. R. Stat. Soc. Ser. C (Applied Statistics)* **31**, 300–303 (1982).
69. Zimmerman, D. L. & Ver Hoef, J. M. On deconfounding spatial confounding in linear models. *Am. Stat.* 1–9 (2021).
70. Hefley, T. J., Hooten, M. B., Drake, J. M., Russell, R. E. & Walsh, D. P. When can the cause of a population decline be determined?. *Ecol. Lett.* **19**, 1353–1362 (2016).
71. Fieberg, J., Ditmer, M. & Freckleton, R. Understanding the causes and consequences of animal movement: A cautionary note on fitting and interpreting regression models with time-dependent covariates. *Methods Ecol. Evol.*https://doi.org/10.1111/j.2041-210X.2012.00239.x *(2012).*

Acknowledgements

This research was funded by Colorado Parks and Wildlife.

Author contributions

J.V. and M.H. wrote the manuscript and designed the model. J.I. acquired the data. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-15694-6.

Correspondence and requests for materials should be addressed to J.J.V.E.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.