Received: 12 December 2025 | Accepted: 13 May 2026

DOI: 10.1111/2041-210x.70329

Methods in Ecology and Evolution

RESEARCH ARTICLE

Geometric overdispersion facilitates the integration of ecological data

Justin J. Van Ee[1,2] | Yoichiro Kanno[2] | Xinyi Lu[3] | Jacob M. Rash[4] | Mevin B. Hooten[1]

[1]Department of Statistics and Data Sciences, The University of Texas at Austin, Austin, Texas, USA

[2]Wildlife Biology Program and Department of Ecosystem Conservation and Sciences, University of Montana, Missoula, Montana, USA

[3]Mathematics and Statistics Department, Utah State University, Logan, Utah, USA

[4]North Carolina Wildlife Resources Commission, Marion, North Carolina, USA

Correspondence
Justin J. Van Ee
Email: justin.vanee@gmail.com

Funding information
U.S. Fish and Wildlife Service, Grant/Award Number: F25AP00123

Handling Editor: Eric Ward

Abstract

1. Statistical data integration facilitates inference based on the variety of data prevalent in ecology. In particular, integrated distribution models (IDMs) have been proposed for inferring spatial patterns in abundance using combinations of noisy count, presence–absence and presence–only data. Obtaining posterior inference from these models is challenging because most optimized software only pertains to a single data type or perfectly observed counts and more general approaches are computationally burdensome.
2. We propose an efficient modelling framework for the joint analysis of multiple datasets. Our framework assumes that each data type arises from the same latent spatial process and allows for unique observational bias and uncertainty. Our implementation is computationally tractable because we model abundance as geometric random variables, which facilitates Markov chain Monte Carlo algorithms with Gibbs updates using Pólya-Gamma augmentation.
3. In a simulation study, we show that our geometric model can provide a useful approximation to Poisson-based IDMs and improves posterior inference for spatial variation in abundance, especially for overdispersed count data. Even for moderately sized datasets (i.e. $n \geq 500$ sites), our approach is nearly 100 times more computationally efficient than Poisson IDMs implemented with `Stan`. Our approach has computational efficiency similar to variational inference, but results in less bias. We demonstrate the flexibility and scalability of the geometric-based IDM in case studies of *Turdus migratorius* and *Salvelinus fontinalis* abundance.
4. Our contribution complements ongoing developments in integrated ecological models by providing a fast and flexible framework for joint inference on high-dimensional latent effects. We further evaluate the relative performance of Poisson, negative binomial and geometric models for estimating covariate associations and spatial variation in abundance.

KEYWORDS
abundance, data integration, Gibbs sampling, integrated distribution models, Pólya-Gamma augmentation, presence–absence, presence–only, species distribution models

1 | INTRODUCTION

Interest in data integration approaches has grown substantially (Han & Si, 2025; Hector et al., 2024), with especially strong uptake in ecology (Zipkin et al., 2021). Collaboration in ecology often produces several data types for the same phenomenon. For example, researchers from different institutions study the same processes, multiple agencies manage transregional populations, and two entities study different characteristics of the same organism. Likewise, long-term, range-wide data are often complemented by more restricted (but also targeted) datasets or opportunistic data (Di Febbraro et al., 2023; Johnston et al., 2023). A joint analysis of these data has many advantages (Schaub & Kéry, 2021), including increased parameter precision (Farr et al., 2021; Frost et al., 2023; Robinson et al., 2020; Schindler et al., 2022), inference about variation in processes across scales (Zipkin et al., 2021), improved predictive performance (Fletcher et al., 2019; Grüss et al., 2023) and correction of bias in the individual data types (Hanks et al., 2011; Pagel et al., 2014; Stillman et al., 2023; Zhao et al., 2024).

Methods in ecology that accommodate multiple data types are predominantly Bayesian (Clark, 2005; Schaub & Kéry, 2021), and wide-scale adoption has been facilitated by Bayesian software development (Brooks et al., 2004). Popular Bayesian software in ecology includes `JAGS` (Plummer et al., 2003), `BUGS` (Lunn, Jackson, et al., 2013), `Stan` (Carpenter et al., 2017) and `NIMBLE` (de Valpine et al., 2017). The aforementioned software can be used to make posterior inference from multiple data types, but may be an order of magnitude slower than more optimized statistical software (Strebel et al., 2022). Furthermore, the need to marginalize discrete parameters to use `Stan` and slow mixing for high-dimensional parameters using `JAGS` and `BUGS` (Monnahan et al., 2017; Turek et al., 2017) can create additional barriers to implementation.

Integrated nested Laplace approximation (INLA) as implemented in software `INLA` (Lindgren & Rue, 2015) has become a popular way to analyse presence–only and presence–absence data (Isaac et al., 2020; Morera-Pujol et al., 2023; Simmonds et al., 2020; Watson et al., 2021), and Mostert and O'Hara (2023) developed an `R` package to facilitate its application in the context of integrated distribution models (IDMs). One limitation of the INLA approach is that it is only amenable to latent Gaussian models where each data type is assumed to be observed perfectly (Rue et al., 2009). This limitation can be circumvented for particular forms of uncertainty or bias (e.g. preferential sampling; Gelfand & Shirota, 2019; Simmonds et al., 2020; and simplified N-mixture models as described in Meehan et al., 2020), but INLA may not be applicable when one or more of the data types has a complex observation model (e.g. site-level abundance inferred from capture–recapture or distance sampling designs). Furthermore, `INLA` reports posterior univariate marginals as opposed to traditional Markov chain Monte Carlo (MCMC) approaches that can be used to make joint inference based on the full posterior distribution.

A variety of maximum likelihood methods have also been proposed for accelerating inference from IDMs. Renner et al. (2019) developed a penalized likelihood approach for presence–only and detection/non-detection data. Dovers et al. (2024) proposed approximating a latent Gaussian random field shared across presence–absence and presence–only data with bi-square basis functions and fixed-rank kriging. Vu et al. (2025) used Template Model Builder (Osgood-Zimmerman & Wakefield, 2023) for modelling multiple types of abundance data. Like INLA, these approaches are not directly applicable to latent count models. Maximum likelihood methods have been developed for fitting latent count models (Zhang et al., 2019) independently but not jointly with presence–only and presence–absence data.

Variational Bayesian inference provides a more general framework for reducing computational burden (Kucukelbir et al., 2017; Zhang et al., 2022). Such approaches approximate the posterior distribution by finding a distribution that minimizes the Kullback–Leibler divergence from the posterior. Because the approach relies on a single optimization, variational inference (like INLA) can be one to three orders of magnitude faster than MCMC approaches (Blei et al., 2017). Variational inference approaches are not common in ecology but a few have been proposed for certain modelling frameworks (Clark et al., 2016; Dennis et al., 2024; Hui et al., 2017; Niku et al., 2019; Tirronen & Kuparinen, 2024). Limited adoption may be related to the tendency of variational approaches to underestimate posterior uncertainty (Duan et al., 2018; Grimmer, 2011), which is particularly unappealing in ecology where Bayesian approaches have been popularized for their ability to robustly quantify multiple sources of variation (Dorazio, 2016). Moreover, variational approaches can suffer from convergence issues for high-dimensional effects (Dhaka et al., 2021) that are often used to account for dependence in space and time.

Given the limitations of INLA and maximum likelihood and drawbacks of variational inference, a fully Bayesian implementation of models for multiple data types is often unavoidable. We propose an efficient, tuning-free and hierarchical implementation for IDMs. The IDM is an adaptation of data integration used for analysing spatial patterns in habitat suitability (Hirzel et al., 2002), species presence (Phillips et al., 2006) or species abundance (Warton & Shepherd, 2010) using multiple data types. In particular, we consider IDMs for inferring species abundance from count, presence–absence and summarized presence–only data. Abundance-based species distribution models (SDMs) provide several advantages over typical occurrence-based SDMs. Stable abundance of an indicator species provides stronger evidence for ecosystem health than mere presence (Waldock et al., 2022) because abundance is more sensitive to ecosystem degradation. For example, a review by Maxwell et al. (2019) reported abundance declines following extreme weather for over 100 studies whereas local extinction was only recorded for 31. Likewise, ecosystem services are more strongly associated with abundance than species richness (i.e. the presence of many species). Winfree et al. (2015) found that variation in crop pollination was primarily explained by

the abundance of dominant bee species and that species richness was only weakly associated with ecosystem function. Finally, the management of harvested species involves abundance more than presence–absence.

We consider settings in which data are observed imperfectly (e.g. detection/non-detection, mark–recapture, distance sampling, spatially replicated counts or depletion survey designs). These latent count models often lead to likelihoods that cannot be marginalized analytically, limiting the applicability of optimized approaches such as `INLA` or standard maximum likelihood methods (Renner et al., 2019). Bayesian approaches that account for the sampling design of each data type have been successfully implemented to learn fixed covariate effects (Dorazio, 2014; Fithian et al., 2015; Kéry et al., 2024a; Koshkina et al., 2017; Twining et al., 2024), but joint inference for high-dimensional structured random effects is more challenging (Dovers et al., 2024). Pacifici et al. (2017) incorporated a latent spatial field within an IDM for count and detection/non-detection data, but their implementation relied on independent Metropolis updates for the spatial random effect that are prone to poor mixing in high dimensions (Andrieu et al., 2010; Duan et al., 2018; Turek et al., 2017). Various multistage MCMC approaches have been successfully applied for modelling high-dimensional ecological data (Johnson et al., 2022; Van Ee et al., 2023, 2024), but recursive Bayesian techniques can be challenging to implement (Hooten et al., 2021; Lunn, Barrett, et al., 2013). Our approach fills this gap by providing fast and flexible posterior inference for high-dimensional structured random effects shared across multiple, noisy data types.

Our method assumes geometric distributions for the true, generally unobserved, site-level abundance. We accommodate the spatial and temporal heterogeneity in site-level abundance by modelling the expected value at each site as a log-linear function of covariates and random effects analogous to a Poisson regression. However, unlike a Poisson model, which induces a 'cloglog' link function, our geometric IDM implies a logistic link between the linear predictor and probability of species presence. We show that the joint likelihood for species abundance and presence can be written as a product of binomial likelihoods parameterized in terms of their log-odds, enabling Gibbs sampling via Pólya-Gamma augmentation (Polson et al., 2013). Under the conditionally conjugate formulation, mixing for the fixed and latent spatial random effect is efficient because posterior samples are drawn directly from their full-conditional distribution each iteration. In contrast, typically applied Metropolis updates are prone to low acceptance rates for high-dimensional parameters, and Hamiltonian dynamics as implemented in `Stan` are burdensome to simulate per iteration.

Pólya-Gamma augmentation frameworks have only recently been adopted in the context of SDM. Clark and Altwegg (2019) proposed Pólya-Gamma augmentation for efficient sampling of covariate effects in single species occupancy models, and various extensions have been proposed that highlight the computational efficiency and flexibility of the approach. Pólya-Gamma augmentation can substantially improve computation for occupancy models fit to large spatial datasets with spatio-temporal random effects (Diana et al., 2023; Doser et al., 2023). It is also straightforward to modify the Pólya-Gamma augmentation approach to accommodate spike-and-slab regression (Griffin et al., 2020), multiple types of detection/non-detection data (Doser et al., 2022), multiple species (Clark & Altwegg, 2023) and spatially varying coefficients (Doser et al., 2025). We present the first Pólya-Gamma augmentation approach for the joint analysis of count, presence–absence and presence–only data.

In a simulation study, we assess the computational efficiency and inferential implications of this approach relative to Poisson-based IDMs implemented in `Stan` using Hamilitonian Monte Carlo and variational inference. We also compare with negative binomial-based IDMs that can accommodate variable levels of overdispersion but are not amenable to Pólya-Gamma augmentation. We use our geometric IDM for joint modelling of distance sampling surveys with presence–absence data in a case study of *Turdus migratorius* (American robins) abundance. In a case study of *Salvelinus fontinalis* (brook trout) abundance, we demonstrate efficient mixing for high-dimensional structured random effects. We also explore the relative strengths of each integrated modelling approach for inferring species distributions.

2 | METHODS

2.1 | Poisson integrated distribution model

Species distribution are often modelled using count, presence–absence and presence–only data. The connections between these data types can be motivated from a point process model (Fletcher et al., 2016, 2019; Hefley & Hooten, 2016). In particular, it is assumed that all data types share a latent species intensity (i.e. the 'joint likelihood' approach; Miller et al., 2019) over the spatial domain, $\mathcal{S}$, although IDMs that relax this assumption have been proposed (Pacifici et al., 2017, 2019). The intensity of abundance is modelled as a log-linear combination of covariates and a latent Gaussian process,

$$\log(\lambda(\boldsymbol{s})) = \boldsymbol{x}'(\boldsymbol{s})\boldsymbol{\beta} + \eta(\boldsymbol{s}), \tag{1}$$

$$\boldsymbol{\eta} \sim \text{GP}(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{2}$$

where $\boldsymbol{s} = (s_1, s_2)'$ denotes position in geographic space, $\boldsymbol{\eta} = (\eta(\boldsymbol{s}_1), \ldots, \eta(\boldsymbol{s}_n))'$, and we use lower case light, lower case bold and upper case bold symbols to represent scalars, vectors and matrices, respectively. For simplicity, we describe a distribution model at a single time point, but this approach can be used to model species distributions with space–time dynamics.

The Gaussian process in (2) accounts for spatial dependence related to missing environmental covariates as well as social behaviours (e.g. schooling, flocking and territoriality), and has been shown to improve inference in IDMs (Gelfand & Shirota, 2019; Isaac et al., 2020; Pacifici et al., 2019; Renner et al., 2019; Simmonds et al., 2020).

Assuming the number and locations of individuals follows an inhomogeneous Poisson point process, a natural choice in SDMs (Warton & Shepherd, 2010), the distribution for the abundance of individuals N_i at site i is.

$$N_i \mid \boldsymbol{\beta}, \boldsymbol{\eta} \sim \text{Poisson}\left(\int_{\mathcal{S}_i} \lambda(\boldsymbol{s}) d\boldsymbol{s}\right), \tag{3}$$

where $\mathcal{S}_i$ denotes the region surveyed at site i. The presence of a species can be viewed as a censoring of the latent abundance ($z_i = \mathbb{I}(N_i > 0)$) and has probability distribution:

$$z_i \mid \boldsymbol{\beta}, \boldsymbol{\eta} \sim \text{Bernoulli}\left(1 - e^{-\int_{\mathcal{S}_i} \lambda(\boldsymbol{s}) d\boldsymbol{s}}\right). \tag{4}$$

For presence–only data, sites are not naturally defined, and the spatial domain $\mathcal{S}$ could be partitioned into survey regions $\mathcal{S}_i$ such that $\mathcal{S} = \bigcup_{i=1}^{n} \mathcal{S}_i$ to aggregate point-level records into site-level counts (Aarts et al., 2012; Baddeley et al., 2010; Renner et al., 2015). Presence–only data may also only be unobserved for a subset of sampling regions.

Without loss of generality, we assume a collection of site-level abundances either observed directly or aggregated from presence–only data. At a subset of sites, the site-level abundances are censored, resulting in presence–absence data. We denote the set of locations where we have observed count and presence–absence data by $\mathcal{S}_N$ and $\mathcal{S}_z$, which have n_N and n_z elements (sites), respectively. To accommodate measurement error, both the true abundance, N, and presence, z, of the species can be treated as latent variables and learned using the observed count y and binary v data, respectively. We use the bracket notation:

$$y_i \sim [y_i \mid N_i, \theta_N], \tag{5}$$

$$v_i \sim [v_i \mid z_i, \theta_z], \tag{6}$$

for $i = 1, \ldots, n$, to represent probability distributions (Gelfand & Smith, 1990) describing observational uncertainty in the sampling design for abundance and presence–absence. Likewise, θ_N and θ_z denote data-level parameters related to the sampling designs (e.g. detection probabilities, capture rates or correction factors).

In practice, IDMs are often fit to areal data where both the Gaussian process and covariates are assumed to be approximately constant over the sampling region $\mathcal{S}_i$. In this case, we write the expected abundance at site i as:

$$\int_{\mathcal{S}_i} \lambda(\boldsymbol{s}) d\boldsymbol{s} \approx a(\boldsymbol{s}_i)\lambda(\boldsymbol{s}_i), \tag{7}$$

where $a(\boldsymbol{s}_i)$ denotes the area surveyed at site i. The relevance of Equation (7) is that we can express the intensity of abundance at site i as a log-linear function of fixed and spatial random effects:

$$\int_{\mathcal{S}_i} \lambda(\boldsymbol{s}) d\boldsymbol{s} \approx a(\boldsymbol{s}_i) \exp\{\boldsymbol{x}'(\boldsymbol{s}_i)\boldsymbol{\beta} + \eta(\boldsymbol{s}_i)\}, \tag{8}$$

which also implies a log-linear connection between expected abundance and the predictors such that:

$$\mathbb{E}(N_i \mid \boldsymbol{\beta}, \boldsymbol{\eta}) = a(\boldsymbol{s}_i) \exp\{\boldsymbol{x}'(\boldsymbol{s}_i)\boldsymbol{\beta} + \eta(\boldsymbol{s}_i)\}. \tag{9}$$

The probability of species presence at site i is approximately:

$$\text{cloglog}^{-1}\left(\log\left(\int_{\mathcal{S}_i} \lambda(\boldsymbol{s}) d\boldsymbol{s}\right)\right) \equiv 1 - e^{-a(\boldsymbol{s}_i)\exp\{\boldsymbol{x}'(\boldsymbol{s}_i)\boldsymbol{\beta} + \eta(\boldsymbol{s}_i)\}}. \tag{10}$$

For contexts in which the intensity is thought to vary considerably within the survey region of a site $\mathcal{S}_i$ we can still write $\int_{\mathcal{S}_i} \lambda(\boldsymbol{s}) d\boldsymbol{s}$ as a log-linear function of $\boldsymbol{\beta}$ and $\eta(\boldsymbol{s}_i)$ using a fine-scale grid approximation.

2.2 | Geometric integrated distribution model

The crux in fitting the Poisson IDM is efficiently sampling the spatial random effect, $\boldsymbol{\eta}$, in a MCMC algorithm. The full-conditional distribution for $\boldsymbol{\eta}$ contains both Poisson and binomial likelihoods and is intractable. Independent Metropolis updates for the spatial random effect have been implemented in the context of IDMs (Pacifici et al., 2017), but this sampling approach is known to have poor mixing in high dimensions (Andrieu et al., 2010; Duan et al., 2018). Hamiltonian Monte Carlo (HMC) and its extensions (Hoffman & Gelman, 2014), as implemented in `Stan`, can be used to obtain less correlated samples in high dimensions, but the computational burden may be prohibitive for large datasets. Variational inference approaches can substantially reduce computational burden but may be unstable and underestimate posterior uncertainty.

We adapted the IDM to enable efficient joint updates for fixed and spatial random effects using Pólya-Gamma data augmentation (Polson et al., 2013). Instead of modelling the true (i.e. in most cases unobserved) abundance of individuals N_i at site i as a Poisson random variable, we adopted a geometric model,

$$N_i \sim \text{geometric}(\psi_i), \tag{11}$$

Figure 1 shows the probability mass function for a Poisson distribution alongside negative binomial distributions with varying levels of overdispersion. A geometric distribution is equivalent to a negative binomial distribution with overdispersion parameter fixed to $\nu = 1$. The geometric distribution implies more variation in site-level abundances than the Poisson. Because the geometric distribution is a simplification of the negative binomial, it can also be motivated from a Poisson–Gamma mixture with latent spatial process, which we show in the Supporting Information.

To match conditional means in the Poisson and geometric distributions, we let:

$$\psi_i = \text{logit}^{-1}\left(\log\left(\int_{\mathcal{S}_i} \lambda(\boldsymbol{s}) d\boldsymbol{s}\right)\right) \equiv \frac{\int_{\mathcal{S}_i} \lambda(\boldsymbol{s}) d\boldsymbol{s}}{1 + \int_{\mathcal{S}_i} \lambda(\boldsymbol{s}) d\boldsymbol{s}}, \tag{12}$$

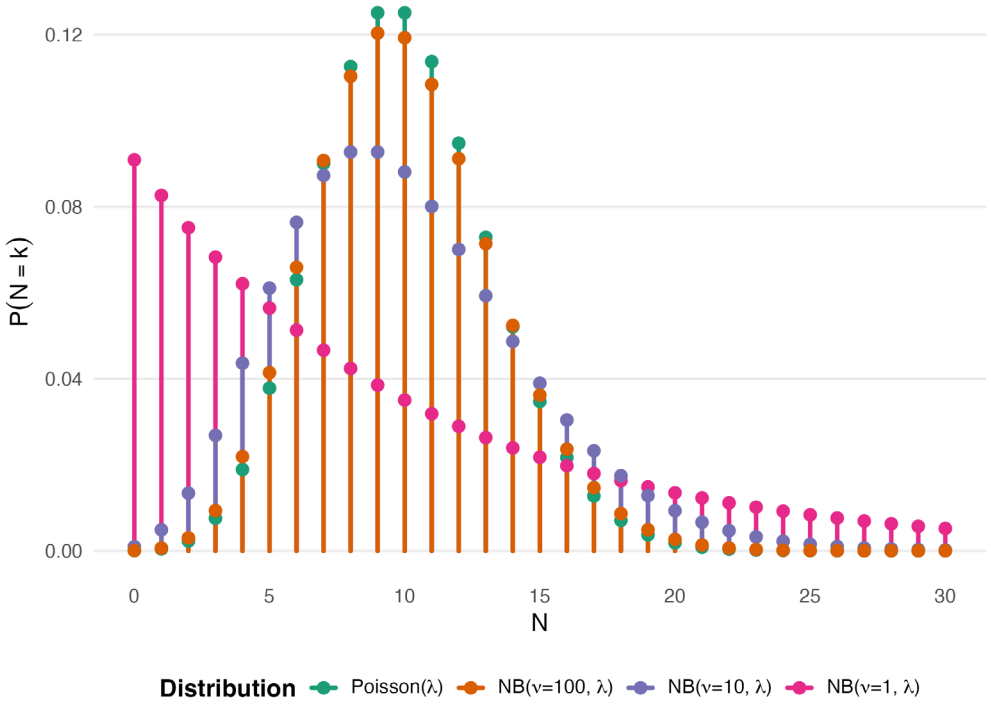

FIGURE 1 Poisson and negative binomial distributions (Equation 27) with varying levels of overdispersion. All distributions have a mean of $\lambda = 10$. Note that the NB $(\lambda, \nu = 1)$ coincides with a geometric distribution.

which implied:

$$\mathbb{E}(N_i|\,\psi_i) = \frac{\psi_i}{1-\psi_i} = \int_{S_i} \lambda(\boldsymbol{s})d\boldsymbol{s}, \tag{13}$$

$$\mathrm{Var}(N_i|\,\psi_i) = \frac{\psi_i}{(1-\psi_i)^2} = \int_{S_i} \lambda(\boldsymbol{s})d\boldsymbol{s}\left(1 + \int_{S_i} \lambda(\boldsymbol{s})d\boldsymbol{s}\right). \tag{14}$$

If we substitute the approximation for $\int_{S_i} \lambda(\boldsymbol{s})d\boldsymbol{s}$ from Equation (8), then:

$$\mathbb{E}(N_i|\,\psi_i) = a(\boldsymbol{s}_i)\exp\{\boldsymbol{x}'(\boldsymbol{s}_i)\boldsymbol{\beta} + \eta(\boldsymbol{s}_i)\} \equiv \lambda_i, \tag{15}$$

which is equivalent to Equation (9). Hence, the interpretation of fixed and spatial random effect is the same as in a Poisson regression, the expected change in the log-intensity per unit of area for a unit change in the predictor. The resulting probability of presence is as follows:

$$\Pr(z_i = 1) = \Pr(N_i > 0) = 1 - \Pr(N_i = 0) = 1 - (1-\psi_i) = \psi_i, \tag{16}$$

which implies $z_i \sim \mathrm{Bernoulli}(\psi_i)$ and that the predictors also model the change in log-odds of species presence at site i.

Under this specification, the joint likelihood for abundance and presence–absence is a combination of binomial likelihoods parameterized by log-odds and, therefore, is amenable to Pólya-Gamma augmentation (Polson et al., 2013). Specifically, with a normal prior for the fixed effect, $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$, samples of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ can be drawn directly from their full-conditional distribution:

$$[\boldsymbol{\beta}, \boldsymbol{\eta}|\,\boldsymbol{\Omega}_z, \boldsymbol{\Omega}_N, \boldsymbol{\Sigma}] = \mathcal{N}(\boldsymbol{m}, \boldsymbol{V}), \tag{17}$$

where:

$$\boldsymbol{V} = \left(\boldsymbol{B}_z'\boldsymbol{\Omega}_z\boldsymbol{B}_z + \boldsymbol{B}_N'\boldsymbol{\Omega}_N\boldsymbol{B}_N + \begin{pmatrix} \boldsymbol{\Sigma}_\beta & \boldsymbol{0}_{n\times p} \\ \boldsymbol{0}_{p\times n} & \boldsymbol{\Sigma} \end{pmatrix}^{-1}\right)^{-1}, \tag{18}$$

$$\boldsymbol{m} = \boldsymbol{V}\left(\boldsymbol{B}_z'\boldsymbol{\kappa}_z + \boldsymbol{B}_N'\boldsymbol{\kappa}_N + \left(\left(\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta\right)', \boldsymbol{0}_{n\times 1}'\right)'\right), \tag{19}$$

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{X} & \boldsymbol{I}_{n\times n} \end{pmatrix}. \tag{20}$$

We use $\boldsymbol{\Omega}_z = \mathrm{diag}\left\{\omega_{z_1}, \ldots, \omega_{z_{n_z}}\right\}$ and $\boldsymbol{\Omega}_N = \mathrm{diag}\left\{\omega_{N_1}, \ldots, \omega_{N_{n_N}}\right\}$ to denote diagonal matrices of Pólya-Gamma random variables which have full-conditional distributions:

$$\omega_{z_i}\,|\,\boldsymbol{\beta}, \boldsymbol{\eta} \sim \mathrm{PG}(1, \lambda_i), \quad \text{for } i = 1, \ldots, n_z, \tag{21}$$

$$\omega_{N_i}\,|\,\boldsymbol{\beta}, \boldsymbol{\eta} \sim \mathrm{PG}(N_i + 1, \lambda_i), \quad \text{for } i = n_z + 1, \ldots, n_z + n_N, \tag{22}$$

where we have used the definition of λ_i from Equation (15). The submatrices $\boldsymbol{B}_z$ and $\boldsymbol{B}_N$ are obtained by selecting the rows of $\boldsymbol{B}$ that belong to sites with binary and count observations, respectively (i.e. $\boldsymbol{B}_z = \{\boldsymbol{B}_i\}$ for $i \in \mathcal{S}_z$). We define $\boldsymbol{\kappa}_z = (z_1 - 1/2, \ldots, z_{n_z} - 1/2)'$, $\boldsymbol{\kappa}_N = ((N_1 - 1)/2, \ldots, (N_{n_N} - 1)/2)'$ and use $\boldsymbol{0}_{p\times n}$ to denote zero matrices with p rows and n columns. The form of the full-conditional distribution in Equation (17) is intuitive. The first, second and third terms in the sums for the posterior variance $\boldsymbol{V}$ and mean $\boldsymbol{m}$ represent the contributions of presence–absence, abundance and the prior.

To summarize, we have modified the typical framework for modelling species abundance by using a geometric distribution rather than Poisson distribution. Under these distributional assumptions, the interpretation of the fixed effect is unified across the two most commonly used ecological modelling frameworks for abundance and presence–absence data. In particular, $\boldsymbol{\beta}$ represents both the expected change in log-intensity and log-odds of species presence for a unit change in the predictors. This convenient relation between the fixed effect and the log-odds of species presence does not apply to the Poisson model with 'cloglog' link function nor more general negative binomial models. The conditional probability of species presence for commonly applied distribution models is provided in Table 1.

Our framework also accommodates overdispersion by specifying a variance for the latent, site-level abundance from Equation (14) that is $\left(1+\int_{S_i}\lambda(\boldsymbol{s})d\boldsymbol{s}\right)$ times greater relative to the Poisson. Overdispersion in abundance data is common in ecology (Miller et al., 2019; Richards, 2008) and assuming a Poisson distribution can result in overly narrow posterior predictive distributions (Hooten & Hefley, 2019). Distribution models systematically over-predict low observed abundances and under-predict high observed abundances (Fukaya et al., 2020; Ploton et al., 2020). The level of overdispersion also scales with true abundance such that greater overdispersion is associated with more numerous species (e.g. greater overdispersion for abundance of insects than for mammals).

Overdispersion arises more generally in ecological count data because of sampling errors, environmental stochasticity, missing predictors and flocking (Lindén & Mäntyniemi, 2011). Our distribution model accounts for these additional sources of variability through its hierarchical formulation and inclusion of covariates and spatial dependence. While the spatial random effect may partially absorb variation related to flocking or species aggregation, the influence of this factor is scale dependent. In particular, the influence of flocking is likely more pronounced at smaller sites, where a few to no groups are surveyed, versus at larger sites, where multiple groups are encountered. Our geometric model assumes a fixed level of overdispersion and therefore cannot fully account for differences induced by the spatial resolution of sampling. We empirically assess its robustness to misspecification in the following simulation study.

Our geometric model is also less flexible than negative binomial and Conway–Maxwell–Poisson (Conway & Maxwell, 1961) models that can accommodate varying levels of underdispersion and overdispersion in abundance (Lynch et al., 2014). While overdispersion is prevalent in site-level abundances for many species, the level of overdispersion is rarely known a priori. These models, however, do not imply a binomial model for species presence parameterized by log-odds (Table 1) and therefore are not compatible with the proposed dual Pólya–Gamma framework. In the next section, we investigate the inferential consequences of assuming a geometric distribution for abundance when the data are generated from a Poisson IDM. We also assess the performance of our geometric IDM relative to Poisson and negative binomial models for overdispersed abundance data.

TABLE 1 Comparison of Poisson, negative binomial (mean parameterization) and geometric integrated distribution models (IDMs).

	Poisson (λ)	NB (λ, ν)	Geometric (λ)
$\mathbb{E}(N \mid \lambda, \nu)$	λ	λ	λ
$\mathrm{Var}(N \mid \lambda, \nu)$	λ	$\lambda + \frac{\lambda^2}{\nu}$	$\lambda(1+\lambda)$
$\Pr(N>0 \mid \lambda, \nu)$	$1-e^{-\lambda}$	$1-\left(\frac{\nu}{\nu+\lambda}\right)^{\nu}$	$\frac{\lambda}{1+\lambda}$
Dispersion	Regular	Variable	Overdispersed
Pólya-Gamma Augmentation	−	−	+

Note: For each distribution model, we assume that $\log(\lambda_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + \eta_i$. IDMs that are amenable to Pólya-Gamma augmentation are denoted by '+'. The conditional probability of species presence is given by $\Pr(N>0|\lambda,\nu)$, and note that $\mathrm{logit}^{-1}(\log(\lambda)) = \frac{\lambda}{1+\lambda}$.

3 | SIMULATION STUDY

We performed a simulation study to validate our geometric IDM and compared it with a commonly applied Poisson and negative binomial IDMs. We simulated datasets from the following Poisson IDM with perfectly observed abundance and presence–absence data.

$$N_i \sim \mathrm{Poisson}(\lambda_i), \tag{23}$$

$$z_i \sim \mathrm{Bernoulli}(1-e^{-\lambda_i}), \tag{24}$$

$$\log(\lambda_i) = \boldsymbol{x}_i'\boldsymbol{\beta} + \eta_i, \tag{25}$$

$$\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{R}), \tag{26}$$

where $i = 1, \ldots, n$ indexes spatial locations, $\boldsymbol{x}_i$ is a vector of 10 covariates and an intercept, and η_i is the spatial random effect for location i. We simulated spatial locations within a $(0,1)\times(0,1)$ square, and let $\boldsymbol{R} = \exp(-\boldsymbol{D}/\phi)$ be an exponential covariance matrix with spatial range parameter ϕ. Across simulations, we varied the number of sites or total sample size using $n = n_z + n_N = 100, 500, 1000$. We randomly assigned 80% of observations to be binary, following the observation in ecology that easy to collect data are generally more plentiful (Fletcher et al., 2019; Zhao et al., 2024). For each simulated dataset, we drew unique values of the fixed effect $\boldsymbol{\beta}$, spatial variance σ^2 and spatial range ϕ from $\mathcal{N}(0,1)$, $\mathrm{Gamma}(1,1)$ and $\mathrm{Uniform}(0,1)$ distributions, respectively. We simulated 10 continuous covariates from standard normal distributions.

We simulated 100 datasets for each sample size for a total of 300 datasets. For each dataset, we also simulated an additional 20 site-level abundances to quantify out-of-sample predictive performance. Because our Pólya-Gamma augmentation framework only pertains to sampling the fixed and spatial random effect, we treated ϕ as fixed and known so that differences in performance would better reflect the sampling efficiency of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$. In general, fixing the

spatial range is not required, and various sampling approaches for ϕ are compatible within our Pólya-Gamma augmentation framework. While the spatial random effect, $\boldsymbol{\eta}$, is not typically of inferential interest, we report its performance metrics in place of those for site-level densities, $\lambda_{i,t}$, to avoid duplicating performance differences attributable to estimation of the fixed effect, $\boldsymbol{\beta}$. Nevertheless, performance for $\lambda_{i,t}$ can be inferred from the combined contributions of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$, as inadequate performance in either component would likely propagate to $\lambda_{i,t}$.

We implemented the Poisson and negative binomial IDMs with HMC, using the No-U-Turn Sampler (NUTS) proposed by Hoffman and Gelman (2014), which automatically tunes the step size and step number hyperparameters. The negative binomial IDM followed the mean parameterization:

$$N_i \sim [N_i|\,\lambda_i, \nu] = \frac{\Gamma(N_i+\nu)}{\Gamma(\nu)N_i!}\left(\frac{\nu}{\nu+\lambda_i}\right)^{\nu}\left(\frac{\lambda_i}{\nu+\lambda_i}\right)^{N_i}, \tag{27}$$

$$z_i \sim \text{Bernoulli}\left(1-\left(\frac{\nu}{\nu+\lambda_i}\right)^{\nu}\right). \tag{28}$$

We also implemented the Poisson IDM using variational inference (VI) based on a Gaussian approximation with diagonal covariance matrix, also known as a fully factorized Gaussian approximation (Kucukelbir et al., 2015). We also considered a full-rank Gaussian VI approximation but do not report results because they were uniformly worse than the factorized approximation. The aforementioned algorithms were implemented within probabilistic programming language `Stan` (Carpenter et al., 2017). All three IDMs have the same conditional mean for abundance, but they differ in their conditional variance and probability of species presence (Table 1). Implementing the negative binomial IDM requires choosing a prior distribution for ν. We set $\nu \sim \text{half}-\text{Cauchy}(0, 1)$, which is still weakly informative because of its heavy-tail (Gelman, 2006; Polson & Scott, 2012).

We also attempted to implement the Poisson IDM with `JAGS`, but we were unable to obtain posterior inference because of prohibitive convergence issues. The convergence issues associated with `JAGS` are not unique to the model described below and have been documented more broadly for models with high-dimensional structured random effects (Monnahan et al., 2017). The simple Poisson (Equation 23) and negative binomial (Equation 27) models could also be fit with the software `INLA` and `TMB` (Kristensen et al., 2016). Given our broader interests in imperfectly observed data, we restricted our analysis to inferential approaches applicable to latent count models, such as those considered in Sections 4 and 5.

We assessed performance using four metrics. We used the calibrated Bayesian coverage (Coverage) proportion (Little, 2006) to determine whether the inferred posterior distributions were overly narrow. We measured parameter accuracy (Bias) and precision (Variability) using the bias of posterior means and posterior standard deviations, respectively. To quantify differences in computational efficiency ('log(Time)'), we calculated the time to acquire 1000 uncorrelated posterior samples. For the posterior sampling based approaches (i.e. MCMC and HMC), this equated to calculating 1000 / (ESS / time), where ESS refers to effective sample size, a measure of the number of independent samples obtained from a Markov chain (Gelman et al., 1995). This quantity relates to run time for the VI implementation because all Monte Carlo samples drawn from the selected variation distribution are independent.

The Poisson IDM coincides with the model used to generate the data. The negative binomial IDM can approximate a Poisson model as $\nu \to \infty$. Our geometric IDM differed from the data generating model because of the assumed geometric distribution for abundance. Posterior inference for $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ was similar between the geometric IDM and HMC implementations (Figure 2a). Median coverage was near 1 for $n = 100$, but the HMC implementation of the Poisson and negative binomial IDMs achieved the targeted 90% nominal level for $n = 500$. The HMC implementation resulted in the highest accuracy and precision for $\boldsymbol{\beta}$ followed closely by the negative binomial IDM. The negative binomial IDM resulted in slightly worse performance than the Poisson IDM implemented with HMC for $n = 100$ but differences were negligible for $n = 500$. The Poisson IDM implemented with VI was anti-conservative (i.e. posterior variances were too small and resulted in low coverage) and had the highest bias.

All models resulted in slightly elevated coverage for $\boldsymbol{\eta}$ (Figure 2b), although our geometric IDM resulted in coverage closer to the nominal 90% level. Bias was elevated for the VI implementation but similar for the other three methods. Our geometric IDM provided more precise estimates of $\boldsymbol{\eta}$. All four methods resulted in similar posterior predictive distributions for out-of-sample, site-level abundance (Figure 2c). On average, obtaining 1000 independent posterior samples of $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ with our geometric IDM approach was nearly 10 and 100 times faster than the Poisson IDM implemented with HMC for sample sizes 100 and 500, respectively. Because of the prohibitive computation time, we did not implement the Poisson IDM with HMC for the largest sample size, $n = 1000$. The coverage, bias and variability of our geometric IDM for $\boldsymbol{\beta}$ was on par with the performance of the Poisson IDM for $n = 500$. Disconcertingly, moving from $n = 500$ to $n = 1000$, bias for $\boldsymbol{\beta}$ actually increased for the VI implementation.

The VI implementation of the Poisson IDM resulted in similar performance to our geometric IDM for the smallest sample size and greater sampling efficiency. For $n = 500$, the VI implementation still resulted in greater sampling efficiency but had greater bias than our geometric IDM. Our geometric IDM approach had similar sampling efficiency to VI for the largest sample size and had less bias and variability for $\boldsymbol{\eta}$. As anticipated, the VI implementation had a tendency to underestimate uncertainty (Duan et al., 2018; Grimmer, 2011), which resulted in suboptimal coverage for $\boldsymbol{\beta}$. Bias for $\boldsymbol{\beta}$, in the VI implementation, did not decrease with sample size.

The previous simulations demonstrated that our geometric IDM provides a good approximation to the Poisson IDM and that the negative binomial IDM results in similar performance to the Poisson IDM for larger sample sizes. We also assessed the performance of each approach for varying levels of overdispersion. Specifically, we retained the same expression for λ_i as in (25) but simulated abundance N_i and presence–absence z_i using

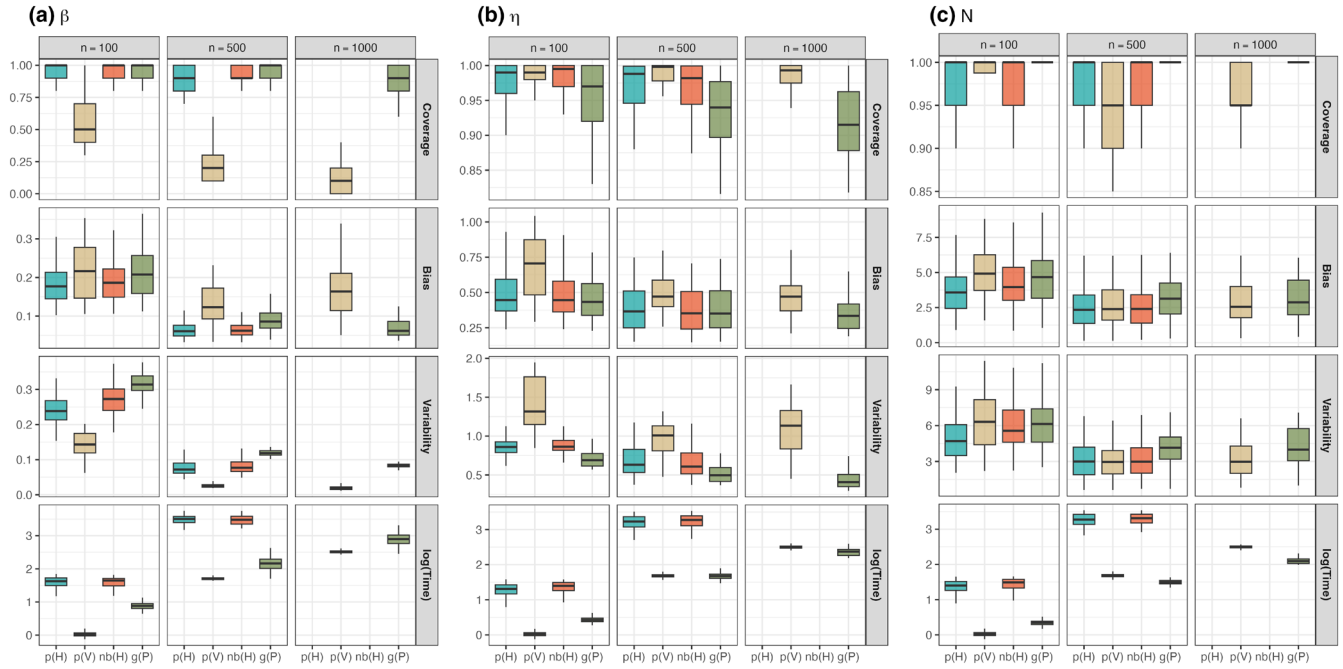

FIGURE 2 Metrics for posterior inference on fixed effect $\boldsymbol{\beta}$ (a) and spatial random effect $\boldsymbol{\eta}$ (b). Metrics related to posterior predictive distributions for site-level abundance are also shown (c). Methods compared include our geometric integrated distribution model (IDM) implemented with Pólya-Gamma augmentation 'g(P)', a Poisson IDM implemented with Hamiltonian Monte Carlo (HMC), 'p(H)' and variational inference 'p(V)', and a negative binomial IDM implemented with HMC, 'nb(H)'. Coverage is the proportion of credible intervals that included the true simulated value. Bias is the absolute difference between the true simulated effect and posterior mean. Variability is the posterior standard deviation. 'log(Time)' is the total run time in log, base ten, seconds required to obtain 1000 independent posterior samples. Coverage, Bias, Variability and 'log(Time)' are averaged over the full parameter vectors $\boldsymbol{\beta}$, $\boldsymbol{\eta}$ and $\mathbf{N}$. Note that the log mean Bias and Variability are plotted in panel (c) to facilitate visualization. Column strips labels denote the number of sites or total number of abundance and presence–absence observations. Row strip labels denote the metric assessed.

Equation (27) for $\nu = 1, 10, 100$, which is inversely related to the level of overdispersion. These specifications for ν resulted in a geometric, moderately overdispersed and nearly Poisson distribution for N, respectively, and are plotted in Figure 1. Hence, for $\nu = 1$, our geometric IDM was the correct model, but for $\nu = 100$, the Poisson IDM, as implemented with HMC and VI, was approximately correct. Only the negative binomial IDM followed the data-generating model for $\nu = 10$. We simulated 100 datasets, all with $n = 100$ spatial locations, for each value of ν. To improve estimation of ν in the negative binomial IDM, we let $n_n = n_z = 50$ to increase the proportion of abundance data.

Across all levels of overdispersion, the Poisson IDM implemented with HMC resulted in median coverage for $\boldsymbol{\beta}$ near the targeted 90% nominal level (Figure 3a). The HMC implementation of the Poisson IDM also resulted in the least variability (excluding the VI implementation which was anti-conservative) across all levels of overdispersion, followed closely by the negative binomial IDM. The geometric and negative binomial IDMs were slightly conservative for larger values of ν. Bias for $\boldsymbol{\beta}$ was similar across approaches. The VI implementation had the lowest posterior uncertainty, which explains its lower coverage.

Coverage for $\boldsymbol{\eta}$ was slightly above the targeted 90% nominal level for each approach (Figure 3b). Our geometric IDM resulted in the least bias and posterior uncertainty for $\boldsymbol{\eta}$ followed closely by the negative binomial IDM. Differences in performance between approaches were greatest for $\nu = 1$. The VI implementation of the Poisson IDM had higher coverage, greater bias and greater posterior uncertainty compared with the HMC implementation. Sampling efficiency for both $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ did not vary with the level of overdispersion. All IDM approaches provided similar posterior predictive distributions for site-level abundance (Figure 3c).

To summarize our findings, Bayesian computation was burdensome for the Poisson and negative binomial distribution models, but considerably faster for the geometric model because of Pólya-Gamma data augmentation. Effective sampling rate is a function of both time per iteration and mixing. Hamilitonian Monte Carlo as implemented in `Stan` resulted in slightly less correlated posterior samplers (i.e. nearly 4 × greater effective sample size) than Pólya-Gamma data augmentation but was nearly 100 times slower per iteration for $n = 500$. Hence, overall Pólya-Gamma data augmentation was much more efficient. The Poisson IDM implemented with variational inference was the least computationally demanding but resulted in high bias and suboptimal coverage for $\boldsymbol{\beta}$.

The Poisson IDM exhibited low posterior mean bias and variance for $\boldsymbol{\beta}$ across sample sizes and levels of overdispersion, but was mildly biased and conservative for $\boldsymbol{\eta}$. In contrast, the geometric IDM showed low posterior mean bias and variance for $\boldsymbol{\eta}$ across scenarios, but was mildly biased and conservative for $\boldsymbol{\beta}$ when counts were

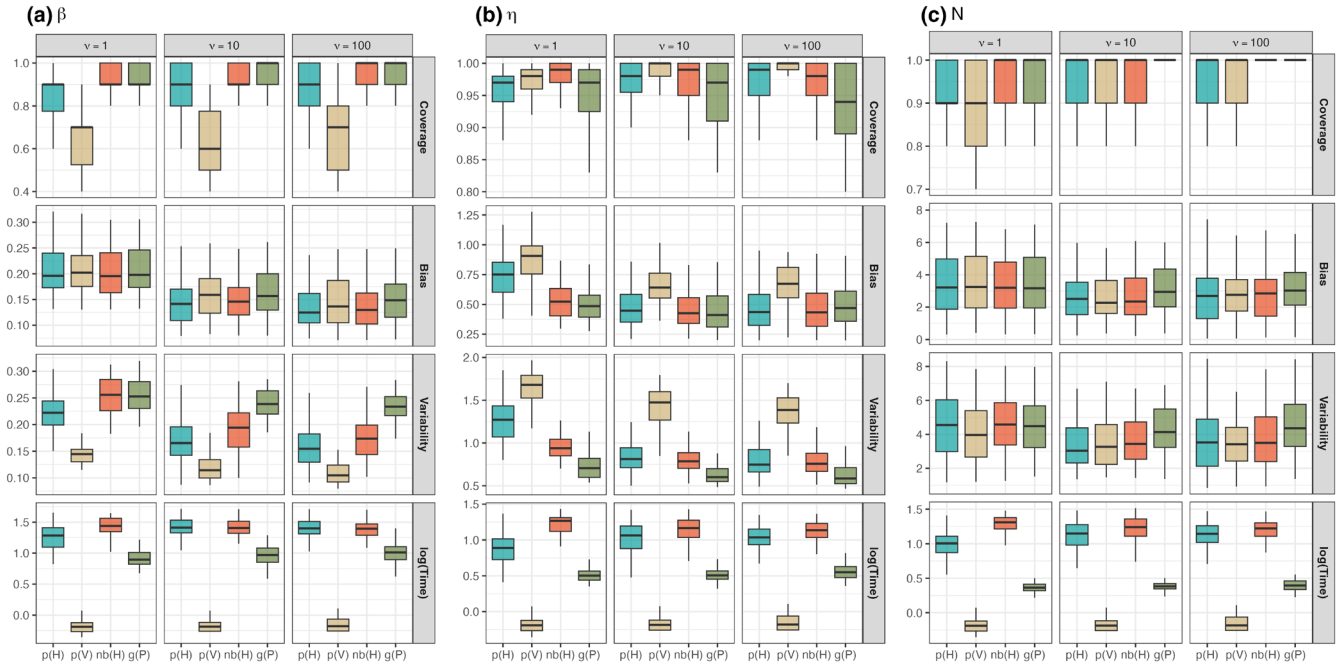

FIGURE 3 Metrics for posterior inference on fixed effect $\boldsymbol{\beta}$ (a) and spatial random effect $\boldsymbol{\eta}$ (b). Metrics related to posterior predictive distributions for site-level abundance are also shown (c). Methods compared include our geometric integrated distribution model (IDM) implemented with Pólya-Gamma augmentation 'g(P)', a Poisson IDM implemented with Hamiltonian Monte Carlo (HMC), 'p(H)' and variational inference 'p(V)', and a negative binomial IDM implemented with HMC, 'nb(H)'. Coverage is the proportion of credible intervals that included the true simulated value. Bias is the absolute difference between the true simulated effect and posterior mean. Variability is the posterior standard deviation. 'log(Time)' is the total run time in log, base ten, seconds required to obtain 1000 independent posterior samples. Coverage, Bias, Variability and 'log(Time)' are averaged over the full parameter vectors $\boldsymbol{\beta}$, $\boldsymbol{\eta}$ and $\boldsymbol{N}$. Note that the log mean Bias and Variability are plotted in panel (c) to facilitate visualization. Column strips labels denote the level of overdispersion. Row strip labels denote the metric assessed.

generated from a near-Poisson model. While the negative binomial IDM has the potential to adjust to the level of overdispersion dynamically, we found that it generally resulted in an intermediate performance relative to Poisson and geometric IDMs for small sample sizes. We did not consider levels of overdispersion beyond that implied by the geometric distribution (i.e. $\nu < 1$), for which the negative binomial IDM might demonstrate the best performance. All distribution models resulted in similar posterior predictive distributions for abundance.

4 | INTEGRATED DISTANCE SAMPLING OF *TURDUS MIGRATORIUS*

Turdus migratorius (henceforth, American robin) is a widely distributed migratory species that can be found year-round across the contiguous United States. We used publicly available data for distance sampling and point counts of American robins in the William L. Finley National Wildlife Refuge of western Oregon (Kéry et al., 2024b). The distance sampling data comprise 2020 detections across 2912 sites between 2011 and 2013. Point counts were derived from opportunistic citizen science surveys (Sullivan et al., 2014) in the same region from 2011 and 2017. After filtering (Kéry et al., 2024a), the point counts had 819 detections of robins from 1059 sites.

Kéry et al. (2024a) proposed an integrated distance sampling model for joint analysis of the two count data types, and other complex data integration approaches have been developed for distance sampling (Schliep et al., 2024; Van Ee et al., 2023, 2024). We developed a simpler model to highlight differences in performance related to our central modelling choice (i.e. treating site-level abundances as geometric rather than Poisson random variables). In particular, we censored the point counts to $\{0, 1\}$ and treated the resulting binary data as the true presence–absence of the species. Furthermore, we did not model the availability (i.e. probability that an individual occupying the study region is available for detection) process. Neither of these assumptions is required for the implementation of our geometric IDM, and both types of imperfect detection could be accounted for within an occupancy modelling framework (MacKenzie et al., 2002). In what follows, we describe a commonly applied distance sampling model for these data.

We let $y_{l,i}$ denote the detection of the lth individual at site i. The L_i observed distances for individuals detected at site i are denoted by $d_{l,i}$. We took a parameter-expanded data augmentation (PX-DA) approach to facilitate posterior inference from the distance sampling data (Hooten & Hefley, 2019; Royle et al., 2009; Royle & Dorazio, 2012). Specifically, we augmented each site with $M - L_i$ undetected individuals, and let $c_{l,i}$ be the indicator of whether the lth

TABLE 2 Posterior means (standard deviations) of covariate associations from various integrated distribution modelling (IDM) approaches including: proposed geometric IDM, 'g(P)', a negative binomial IDM, 'nb(H)', implemented with Hamiltonian Monte Carlo (HMC), a Poisson IDM implemented with HMC, 'p(H)' and variational inference, 'p(V)', and the integrated distance sampling model, 'Kery', of Kéry et al. (2024a).

Approach	Intercept	Canopy	Canopy^2	Elevation	Elevation^2	Time
g(P)	1.98 (0.39)	−0.94 (0.34)	−6.57 (0.68)	−0.90 (0.09)	0.11 (0.02)	355
nb(H)	1.86 (0.39)	−0.63 (0.31)	−5.99 (0.60)	−0.94 (0.08)	0.12 (0.02)	480
p(H)	1.74 (0.41)	−0.67 (0.28)	−5.91 (0.51)	−0.92 (0.07)	0.13 (0.02)	325
p(V)	1.85 (0.02)	0.29 (0.07)	−3.87 (0.11)	−1.06 (0.03)	0.12 (0.02)	13
Kery	3.63 (0.24)	−0.98 (0.28)	−7.46 (0.52)	−0.67 (0.07)	0.08 (0.02)	26,215

Note: Canopy is the percentage of canopy cover within a 315 m radius of the observer. Canopy was scaled such that 0%, 50% and 100% corresponded to values −0.5, 0 and 0.5, respectively. Elevation was scaled to have mean 0 and variance 1 across all sites. Computational efficiency was assessed as the 'Time' (seconds) required to obtain 1000 independent posterior samples. The integrated distance sampling model of Kéry et al. (2024a) was fit using JAGS.

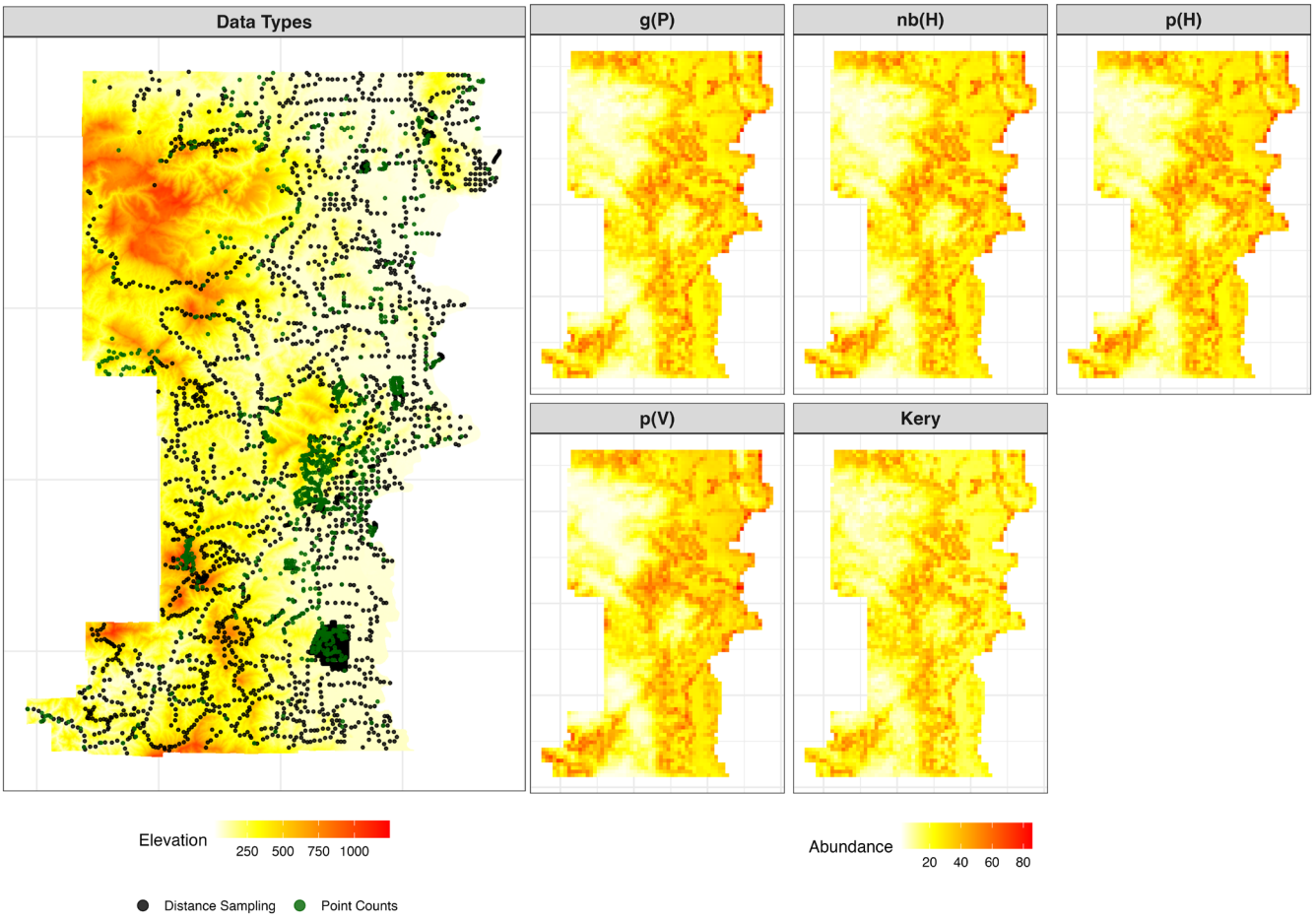

FIGURE 4 Distance sampling surveys and points counts in William L. Finley National Wildlife Refuge (left panel). Right panel shows predicted median abundance at 3874 1 km^2 grid cells using proposed geometric integrated distribution model (IDM), 'g(P)', a negative binomial IDM, 'nb(H)', implemented with Hamiltonian Monte Carlo (HMC), a Poisson IDM implemented with HMC, 'p(H)' and variational inference, 'p(V)' and the integrated distance sampling model, 'Kery', of Kéry et al. (2024a).

individual at site i was available for detection. This model can be formulated hierarchically as a binomial mixture-model:

$$y_{l,i} \sim \begin{cases} \text{Bernoulli}(p_{l,i}) & c_{l,i} = 1, \\ 0 & c_{l,i} = 0, \end{cases}, \tag{29}$$

$$c_{l,i} \sim \text{Bernoulli}(\chi_i). \tag{30}$$

We modelled the detection probability of each individual with a half-normal detection function $p_{l,i} = \exp\left(-d_{l,i}^2 / \sigma_i^2\right)$ and accounted for heterogeneity in detection across sites as a log-linear function of covariates (i.e. $\log(\sigma_i^2) = \boldsymbol{h}_i'\boldsymbol{\gamma}$).

We fit the PX-DA distance sampling model for 10,000 iterations using a Metropolis-within-Gibbs sampling framework. The total run time was 2.1 min (128 Gb, 4.05 GHz machine). For each iteration, we calculated the abundance of individuals at site i using the model derived quantity $N_i = \sum_{l=1}^{M} c_{i,l}$. We fit our geometric IDM, negative binomial IDM and a Poisson IDM implemented with HMC and VI to the inferred abundances and censored point count data. To account for heterogeneity in abundance across sites, we utilized the same linear predictor as Kéry et al. (2024a) by specifying:

$$E(N_i \mid \boldsymbol{\beta}, \boldsymbol{\xi}) = a_i \exp(\boldsymbol{x}_i' \boldsymbol{\beta} + \xi_{t(i)}), \tag{31}$$

where a_i is the area of site i, $\boldsymbol{x}_i$ is a vector of covariates, including an intercept, and $\xi_{t(i)}$ is a random effect for year.

Technical details regarding the implementation of each model are provided in the Supporting Information. Kéry et al. (2024a) reported evidence of slight underdispersion in the distance sampling counts. Hence, while the true model for these data is unknown, a Poisson distribution for the latent abundance is likely more appropriate than our geometric IDM. Although not directly comparable for the reasons mentioned above, we also report posterior inference from the integrated distance sampling model of Kéry et al. (2024a).

We provide posterior means and standard deviations for covariate associations in Table 2. The VI implementation of the Poisson IDM had the greatest sampling efficiency but provided different inference. In particular, the VI implementation suggested a positive association between abundance and the linear effect of canopy, whereas negative associations were inferred with the other four approaches. Based on its elevated bias in the simulation study (Section 3), we suspect this is an estimation error of the VI implementation. Posterior standard deviations were also generally the lowest for the VI implementation. The HMC implementation of the Poisson and negative binomial IDMs and our geometric IDM resulted in similar inference and sampling efficiency. Posterior standard deviations were lowest for the Poission IDM followed by the negative binomial and lastly our geometric IDM. The posterior mean and 90% credible interval for ν estimated from the negative binomial IDM were 1.40 and $(1.25, 1.55)$, respectively. Hence, while other sources of evidence indicated slight underdispersion in the distance sampling counts, the negative binomial IDM suggested overdispersed site-level abundances. Inference from the integrated distance sampling model of Kéry et al. (2024a) was similar to the other approaches with the exception that Kéry et al. (2024a) estimated a larger intercept. Intuitively, incorporating the availability process inflates the estimated abundance at all locations.

Figure 4 shows the expected median abundance of American robins for 3874 1 km^2 grid-cells in William L. Finley National Wildlife Refuge using each approach. We also show the locations of the 2912 distance sampling and 1059 point count surveys. For ease of visualization with the map inferred by Kéry et al. (2024a), we corrected for the negative bias in the IDMs we implemented post hoc by setting their posterior means to 3.63 (Table 2). The VI implementation resulted in the smoothest map and generally inferred greater abundances than the other approaches in the lowest elevation zone (right side of map). The HMC implementation of the Poisson and negative binomial IDMs and our geometric IDM resulted in nearly indistinguishable maps of species abundance. Inferred species distributions were similar to the map reported by Kéry et al. (2024a) but had slightly more spatial variation in expected abundance.

5 | DEPLETION SURVEYS OF *SALVELINUS FONTINALIS*

Salvelinus fontinalis (brook trout) is a salmonid native to eastern North America. Brook trout are economically valuable and culturally significant as the state fish of Vermont, New Hampshire, Michigan, New York, Pennsylvania, New Jersey, West Virginia, Virginia and North Carolina. Ecologically, brook trout are recognized as an indicator species for high-quality coldwater streams (Kanno et al., 2010). Mapping the distribution of brook trout in the Appalachian mountains is challenging because of the dense network of streams and spatial heterogeneity. Because populations can persist in short stream segments (<1 km), the distribution of self-sustaining populations is highly fragmented (Kanno et al., 2011; Letcher et al., 2007), and new populations are discovered every year. A prerequisite for effective conservation is an accurate inventory of species abundance (Primack, 2006). SDM provides a statistically robust framework for habitat prioritization (Guisan et al., 2013) and identifying the streams most suitable for restoration and species reintroduction. Existing distribution models for brook trout have been informed exclusively by presence–absence (Hudy et al., 2008) or count data (Lu, Kanno, Valentine, Kulp, & Hooten, 2024; Lu, Kanno, Valentine, Rash, & Hooten, 2024). We proposed an IDM to facilitate a joint analysis of both data types.

We compiled a dataset of $n = 2064$ sites surveyed across $T = 13$ years (2012–2024). Sites were distributed across 14 subbasins (i.e. 8-digit hydrologic units) as described by Jones et al. (2022) and 15 stream networks, which we define as a collection of hydrologically connected stream segments. All sites were surveyed between June and October using electro-fishing. The majority of sites ($n_z = 1926$) were surveyed to infer brook trout presence at previously unsampled locations. Surveyors sampled the site until a brook trout was caught or until, based on expert opinion, it was determined that no trout were present at the site. At a reduced number of sites ($n_N = 138$), count data were acquired using depletion surveys. Block nets or natural barriers were used to ensure population closure during sampling, and between one and five successive electro-fishing passes were conducted with one backpack electro-fisher and one netter for every three metres of stream width. Most sites were sampled for several consecutive years to facilitate inference on temporal trends in abundance. Figure 5 provides a map of our study area and sampling locations. Permission for field studies was granted by the North Carolina Wildlife Resources Commission.

Following Lu, Kanno, Valentine, Kulp, and Hooten (2024) and Lu, Kanno, Valentine, Rash, and Hooten (2024), we accommodated the

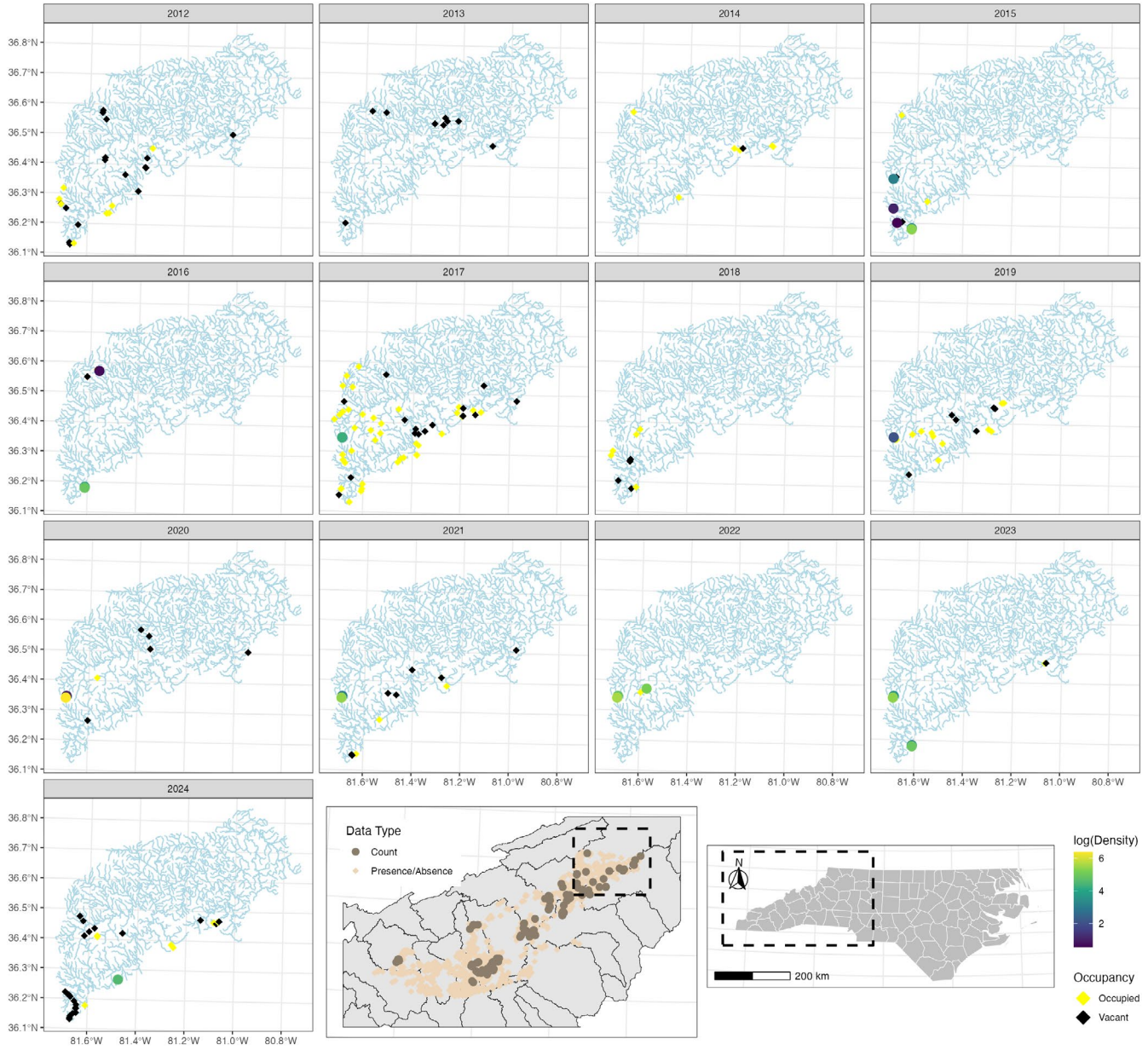

FIGURE 5 Map of study area and survey design. Bottom-right inset plot shows the study region in relation to the state of North Carolina. Bottom-middle inset depicts the distribution of depletion and presence–absence surveys over the United States Geological Survey subbasins. Main plot shows sites for a single stream network over the entire study period (2012–2024).

observational uncertainty in the depletion counts using a removal-based N-mixture model (Dorazio et al., 2005; Wyatt, 2002) such that:

$$y_{i,t,j} \sim \begin{cases} \text{Binomial}\left(N_{i,t}, p_{i,t}\right), & j = 1 \\ \text{Binomial}\left(N_{i,t} - \sum_{l=1}^{j-1} y_{i,t,l}, p_{i,t}\right), & j > 1 \end{cases}, \tag{32}$$

where $y_{i,t,j}$ denotes the count of brook trout caught at site i during year t on pass j. We modelled site-level capture probabilities as $\text{logit}\left(p_{i,t}\right) = \boldsymbol{h}_{i,t}'\boldsymbol{\gamma}$ where the vector $\boldsymbol{h}_{i,t}$ included an intercept and the slope of the stream segment (scaled to have mean 0 and variance 1 across all sites).

We accounted for dependence in brook trout abundance across space and time by modelling the intensity surface as a log-linear combination of multiple effects:

$$\log\left(\lambda_{i,t}\right) = \log\left(a_{i,t}\right) + \boldsymbol{x}_{i,t}'\boldsymbol{\beta} + \xi_t + \zeta_{w(i)} + \eta_i, \tag{33}$$

where $a_{i,t}$ is the stream habitat area sampled at site i in year t in square metres (divided by 1000 to improve numerical stability; Lu, Kanno, Valentine, Kulp, & Hooten, 2024; Lu, Kanno, Valentine, Rash, & Hooten, 2024), $\boldsymbol{x}_{i,t}$ is vector of covariates (intercept included), ξ_t is a temporal random effect, $\zeta_{w(i)}$ describes broad-scale spatial patterns related to the subbasin identity, w, of site i, and η_i accounts for within-network spatial dependence based on hydrologic distance.

The covariates included in $\boldsymbol{x}_{i,t}$ were informed by recent analyses (Lu, Kanno, Valentine, Rash, & Hooten, 2024; Valentine et al., 2024) and are described in the Supporting Information. We modelled temporal dependence in mean brook trout abundance using a first-order autoregressive process.

$$\xi_t \sim \mathcal{N}\left(\rho_\xi \xi_{t-1}, \sigma_\xi^2\right). \tag{34}$$

Because the HUC8 watersheds are contiguous polygons, a natural choice for modelling spatial dependence was the following conditional autoregressive (CAR) process:

$$\boldsymbol{\zeta} \sim \mathcal{N}\left(\mathbf{0}, \sigma_\zeta^2 \left(\text{diag}(\boldsymbol{A}\mathbf{1}) - \rho_\zeta \boldsymbol{A}\right)^{-1}\right), \tag{35}$$

where $\boldsymbol{A}$ is the adjacency matrix defined by:

$$A_{w(i),w(l)} = \begin{cases} 1, & w(i) \neq w(l), \text{HUCs } w(i) \text{ and } w(l) \text{ are adjacent} \\ 0, & \text{otherwise} \end{cases}. \tag{36}$$

To facilitate model parsimony and induce moderate spatial correlation, we let $\rho_\zeta \to 1$, which corresponds to an intrinsic conditional autoregressive (ICAR) process (Ver Hoef et al., 2018). For within stream network dependence, we used a tail-down model (Ver Hoef & Peterson, 2010), which has been suggested as a suitable spatial covariance function for species that can travel upstream. Specifically, we specified the following exponential covariance model:

$$\boldsymbol{\eta}_q \sim \mathcal{N}\left(\mathbf{0}, \sigma_{\eta q}^2 \exp\left(-\boldsymbol{D}_q/\phi_q\right)\right), \tag{37}$$

where $\boldsymbol{D}_q$ is the hydrologic distance matrix (i.e. the distance between all sites when movement is restricted to the stream network) and $\boldsymbol{\eta}_q$, a subvector of $\boldsymbol{\eta}$, is the spatial effects for stream network $q = 1, \ldots, 14$. We specified moderately informative inverse-gamma $\left(2, 1.5^2\right)$ priors for all variance parameters, gamma $(0.1, 0.001)$ for the spatial range ϕ_q, independent $\mathcal{N}\left(0, 1.5^2\right)$ priors for the effects, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and a flat Uniform$(0, 1)$ prior for the temporal autocorrelation parameter ρ_ξ.

As in the simulation study and *Turdus migratorius* case study, we fit our geometric IDM, a negative binomial IDM and a Poisson IDM implemented with HMC and VI to the depletion counts and presence–absence data. We assessed model fit by calculating the ranked probability score (RPS) (Epstein, 1969). For binary data, RPS is equivalent to the Brier Score (Brier, 1950). Because posterior predictive distributions for site-level abundance are heavy-tailed, we report the scaled RPS, which down-weights influence of site-level predictions with high uncertainty (Bolin & Wallin, 2023). We specifically report the scaled RPS multiped by − 1 such that lower values relate to better performance as is the case for the traditional RPS.

The true site-level abundance was unknown for sites sampled using depletion surveys. We used the posterior median from our geometric IDM to approximate the unknown abundance because `Stan` does not provide posterior distributions for discrete parameters.

TABLE 3 Model fit and computational efficiency for proposed geometric integrated distribution model (IDM), 'g(P)', a negative binomial IDM, 'nb(H)', implemented with Hamiltonian Monte Carlo (HMC) and a Poisson IDM implemented with HMC, 'p(H)' and variational inference, 'p(V)'.

Approach	RPS presence–absence	RPS abundance	Time
g(P)	0.08 (0.00, 0.34)	2.93 (1.94, 3.57)	14
nb(H)	0.08 (0.00, 0.25)	2.52 (1.54, 3.18)	598
p(H)	0.37 (0.00, 0.99)	2.45 (1.51, 3.72)	10
p(V)	0.27 (0.00, 0.94)	2.46 (1.63, 3.67)	1673

Note: Average ranked probability score, 5th and 95th quantiles across sites in parentheses, is reported for presence–absence and abundance. Computational efficiency was assessed as the 'Time' (minutes) required to obtain 1000 independent posterior samples of $\boldsymbol{\beta}$. Note that true site-level abundances were unknown and inferred from posterior medians of our geometric IDM.

This was a reasonable approach for comparison because posterior distributions for site-level abundance are primarily inferred from the depletion survey component of the hierarchical model, which was identical across the four IDMs compared. Treating the median abundance as a point estimate of the true abundance was also reasonable as posterior distributions had little variation. The average coefficient of variation, $\frac{\text{Var}(N_i|\boldsymbol{y},\boldsymbol{z})^{1/2}}{\mathbb{E}(N_i|\boldsymbol{y},\boldsymbol{z})}$, across sites was 0.13.

Our motivation for calculating the RPS within-sample was to quantify the relative weight each IDM assigns to the two data types. Heuristically, a relatively high RPS for presence–absence but low RPS for abundance, for example, would suggest that the resulting posterior distribution was more heavily influenced by the count data. We note that this procedure was not meant to assess predictive performance or guide model selection, for which out-of-sample data would be more relevant. Both implementations of the Poisson IDM had a substantially poorer fit to the presence–absence data but slightly better fit for abundance relative to our geometric IDM (Table 3). The negative binomial IDM resulted in a relatively low RPS for both presence–absence and abundance. The posterior mean and 90% credible interval for ν estimated from the negative binomial IDM were 3.90 and (3.17, 4.74), suggesting a moderate level of overdispersion that is neither perfectly captured by the geometric nor Poisson IDM. The sampling efficiency for the VI implementation of the Poisson IDM and our geometric IDM were similar, and roughly 100 times slower for the HMC implementations. The negative bimomial IDM, despite being a more complex model, had a greater computational efficiency than the Poisson IDM. We qualify, however, there is variation in run time for complex models implemented in `Stan`, and in the simulation study, computational burden was generally greater for the negative binomial IDM.

We found limited temporal variation in brook trout abundance but considerable variation within stream networks. We calculated the posterior mean density of fish (individuals per 1000 square metres) based on our geometric IDM for each stream centroid in the network shown in Figure 6. Predicted mean density is only shown for 2024 because inferred patterns in abundance were similar for other years. Densities

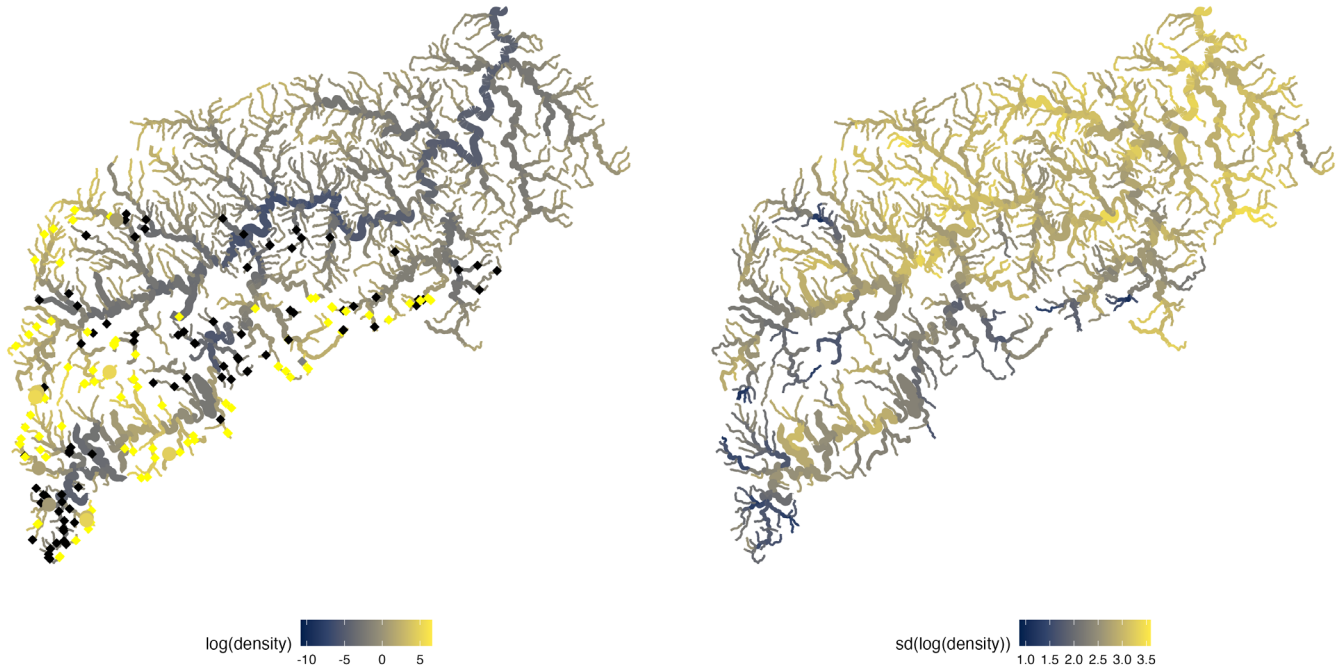

FIGURE 6 Predicted mean density (fish/1000 m^2), left panel, and posterior standard deviation of fitted value, right panel, for selected stream network in North Carolina. Means and variances are for 2024 but all sites visited between 2012 and 2024 are shown in top panel. Predicted means were calculated for each stream segment in the network.

were generally greater for higher elevation and low order stream segments (i.e. tributaries). Lower predicted densities were inferred near and in segments where brook trout were absent. Uncertainty in density was also higher for high-elevation tributaries with the exception of tributaries near sites with count or presence–absence data.

6 | DISCUSSION

We proposed a computationally efficient and flexible approach for facilitating joint inference from potentially noisy count and presence–absence data. Across a simulation study and the analysis of two real datasets, we found that our approach provides inference similar to commonly applied Poisson-based IDMs but can improve sampling efficiency by up to two orders of magnitude. Computing time for our proposed method does not scale adversely with the dimension of the predictors and, thereby, enables posterior inference for high-dimensional parameters like a spatial random effect. For IDMs with a large number of effects, our geometric IDM has a similar computational efficiency to variational approaches but also provides more reliable inference. These computational gains are likely understated because we implemented our approach in `R` (R Core Team, 2024), while the competing approaches have been optimized using `C++` (Gabry et al., 2025).

The linear predictor in the brook trout model included 8 covariates, a temporal effect for 13 years, a broad-scale spatial effect for 14 watersheds, and a stream-level spatial effect for all 2064 locations. Using Pólya-Gamma augmentation, we could sample all 2000+ parameters with a block update, and mixing was such that only 20,000 iterations were needed to obtain a representative sample. Sampling these parameters with Metropolis-Hastings would be challenging (Andrieu et al., 2010; Duan et al., 2018; Turek et al., 2017) and obtaining a posterior sample using HMC via `Stan` was less efficient by two orders of magnitude. One could consider estimating the high-dimensional parameters with one data type, but this approach has drawbacks in the context of IDMs. presence–absence data are minimally informative for inferring spatial patterns in abundance on their own, and count data are often only available for a reduced number of sites. Hence, methods that provide computationally feasible joint inference for random effects are vital to inferring patterns in abundance at broader scales (Fletcher et al., 2019; Zhao et al., 2024).

We compared four IDM approaches that relate the linear predictor to the log-expected abundance across a simulation study and two case studies. Corroborating more general results, we found that variational inference approaches tend to underestimate posterior uncertainty in IDMs and can result in substantial bias (Duan et al., 2018; Grimmer, 2011). For the types of datasets, we considered in Section 3, we found that commonly applied Poisson IDMs provide reliable inference for covariate associations, $\boldsymbol{\beta}$, even when site-level abundances are overdispersed. Inference for the spatial random effect, $\boldsymbol{\eta}$, was also valid but became increasingly conservative as the level of overdispersion increased. Our proposed geometric IDM generally improved inference for the spatial random effect but resulted in slightly inflated estimates of posterior uncertainty for the fixed effect, especially for regularly varying species abundance. More general negative binomial IDMs have the potential to dynamically adjust to the level of overdispersion, but this flexibility

also comes with the challenge of estimating the overdispersion parameter v. All approaches resulted in similar posterior predictive distributions for abundance regardless of the level of overdispersion. Presumably, the Poisson IDM compensates for overdispersed abundance by pushing more variation into site-level differences in the spatial random effect.

The case studies generally reinforced the findings of our simulation study. All IDMs, excluding the Poisson IDM implemented with VI, provided similar posterior means for covariate associations for point counts of American robins, and trends in posterior variances mirrored those observed in simulation. The three IDMs also resulted in nearly identical maps of predicted median abundance. Likewise, in our analysis of brook trout, we found that all IDMs provided similar within-sample fit for abundance. In contrast, we found that the Poisson IDM resulted in a substantially poorer fit for species presence–absence. While this does not necessarily provide evidence that our geometric or the negative binomial IDM are the preferred model for understanding spatial patterns in brook trout abundance, it does highlight differences in the contributions of each data type to posterior inference under the different modelling frameworks. In the Poisson IDM, inference is dominated by the count data because of the assumption of regular variation in abundance. Our geometric IDM, as well as more general negative binomial IDMs, attenuates the weight given to the count data by specifying more dispersed distributions for abundance. A modelling approach that allocates more weight to the observed presence–absence data may be more appropriate when there are uncertainties or biases associated with the count data that are difficult to quantify. We qualify the generality of our results by noting that we primarily have considered scenarios in which the amount of presence–absence data exceeds the amount of abundance data. Because presence–absence data are often used to supplement the spatial, temporal, or taxonomic coverage of harder to collect data (Strebel et al., 2022), this tendency is common in practice.

We conclude that there are three main advantages of our proposed geometric IDM. First, our approach substantially reduces computational burden compared with other state-of-the-art Bayesian implementations of IDMs. Second, modelling abundance as a geometric random variable unifies interpretations of the fixed effect from Poisson and logistic regression. In particular, a unit change in the predictor corresponds to one unit change in the expected log-abundance and log-odds of species presence. Third, our geometric IDM provides a parsimonious alternative to general negative binomial models for modelling overdispersed abundance and may improve inference in small samples, where the dispersion parameter can be difficult to estimate reliably.

Our method assumes a shared process for the presence and abundance of a species. While this is a standard approach for integrated distribution modelling, there is some evidence to suggest that presence and abundance are governed by different processes (Cingolani et al., 2007; Dibner et al., 2017). In plant communities, Cingolani et al. (2007) posited that certain environmental filters preclude species from occupying a site and an additional set of filters may regulate if a species can flourish. Likewise, Maxwell et al. (2019) noted widespread decreases in the abundance of many species but relatively few extirpations. These previous studies suggest partitioning the presence and abundance processes within a hurdle model framework. An avenue for future research would involve modifying our proposed approach to accommodate shared predictors across the two processes.

We have proposed an efficient implementation of IDMs. Our approach is flexible and can be applied to count, presence–absence and presence–only data with complex observer models. The critical modelling assumption is that abundance follows a conditionally geometric distribution. While this assumption may not be met in practice, the simulation study and case study of American robins showed that the trade-off for applying our approach to regularly varying or possibly even underdispersed abundance data is only a slight reduction in the accuracy and precision of posterior distributions for covariate associations. Extending our geometric IDM to larger spatial scales and additional applied settings is a promising direction for future research. Further investigation is also needed to generalize the conditionally conjugate framework to allow estimation of v.

AUTHOR CONTRIBUTIONS

Justin J. Van Ee, Xinyi Lu and Mevin B. Hooten developed the statistical methods. Jacob M. Rash and Yoichiro Kanno contributed to data collection and analysis. Justin J. Van Ee and Mevin B. Hooten drafted the manuscript with substantial input from all authors.

ACKNOWLEDGEMENTS

This project was funded by a Multistate Conservation Grant F25AP00123 from the US Fish and Wildlife Service and jointly administered with the Association of Fish and Wildlife Agencies.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to disclose.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210x.70329.

DATA AVAILABILITY STATEMENT

The data and corresponding R and Stan scripts used for the analyses are available on GitHub and archived at Zenodo (Van Ee, 2026): https://doi.org/10.5281/zenodo.20128893.

ORCID

Justin J. Van Ee https://orcid.org/0000-0002-1305-6922
Yoichiro Kanno https://orcid.org/0000-0001-8452-5100

REFERENCES

Aarts, G., Fieberg, J., & Matthiopoulos, J. (2012). Comparative interpretation of count, presence–absence and point methods for species distribution models. *Methods in Ecology and Evolution*, *3*(1), 177–187.

Andrieu, C., Doucet, A., & Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *72*(3), 269–342.

Baddeley, A., Berman, M., Fisher, N., Hardegen, A., Milne, R., Schuhmacher, D., Shah, R., & Turner, R. (2010). Spatial logistic regression and change-of-support in Poisson point processes. *Electronic Journal of Statistics*, *4*, 1151–1201.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877.

Bolin, D., & Wallin, J. (2023). Local scale invariance and robustness of proper scoring rules. *Statistical Science*, *38*(1), 140–159.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Brooks, S., King, R., & Morgan, B. (2004). A Bayesian approach to combining animal abundance and demographic data. *Animal Biodiversity and Conservation*, *27*(1), 515–529.

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*, 1–32.

Cingolani, A. M., Cabido, M., Gurvich, D. E., Renison, D., & Díaz, S. (2007). Filtering processes in the assembly of plant communities: Are species presence and abundance driven by the same traits? *Journal of Vegetation Science*, *18*(6), 911–920.

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

Clark, A. E., & Altwegg, R. (2023). A Gibbs sampler for multi-species occupancy models. *Environmental and Ecological Statistics*, *30*(2), 189–204.

Clark, A. E., Altwegg, R., & Ormerod, J. T. (2016). A variational Bayes approach to the analysis of occupancy models. *PLoS One*, *11*(2), e0148966.

Clark, J. S. (2005). Why environmental scientists are becoming Bayesians. *Ecology Letters*, *8*(1), 2–14.

Conway, R. W., & Maxwell, W. L. (1961). A queueing model with state dependent service rate. *Journal of Industrial Engineering*, *12*, 132.

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., & Bodik, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, *26*(2), 403–413.

Dennis, E. B., Diana, A., Matechou, E., & Morgan, B. J. (2024). Efficient statistical inference methods for assessing changes in species' populations using citizen science data. *Journal of the Royal Statistical Society. Series A, Statistics in Society*, *188*(3), 641–657.

Dhaka, A. K., Catalina, A., Welandawe, M., Andersen, M. R., Huggins, J., & Vehtari, A. (2021). Challenges and opportunities in high dimensional variational inference. *Advances in Neural Information Processing Systems*, *34*, 7787–7798.

Di Febbraro, M., Bosso, L., Fasola, M., Santicchia, F., Aloise, G., Lioy, S., Tricarico, E., Ruggieri, L., Bovero, S., Mori, E., & Bertolino, S. (2023). Different facets of the same niche: Integrating citizen science and scientific survey data to predict biological invasion risk under multiple global change drivers. *Global Change Biology*, *29*(19), 5509–5523.

Diana, A., Dennis, E. B., Matechou, E., & Morgan, B. J. T. (2023). Fast Bayesian inference for large occupancy datasets. *Biometrics*, *79*(3), 2503–2515.

Dibner, R. R., Doak, D. F., & Murphy, M. (2017). Discrepancies in occupancy and abundance approaches to identifying and protecting habitat for an at-risk species. *Ecology and Evolution*, *7*(15), 5692–5702.

Dorazio, R. M. (2014). Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, *23*(12), 1472–1484.

Dorazio, R. M. (2016). Bayesian data analysis in population ecology: Motivations, methods, and benefits. *Population Ecology*, *58*(1), 31–44.

Dorazio, R. M., Jelks, H. L., & Jordan, F. (2005). Improving removal-based estimates of abundance by sampling a population of spatially distinct subpopulations. *Biometrics*, *61*(4), 1093–1101.

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

Doser, J. W., Finley, A. O., Kéry, M., & Zipkin, E. F. (2022). `spOccupancy`: An `R` package for single-species, multi-species, and integrated spatial occupancy models. *Methods in Ecology and Evolution*, *13*(8), 1670–1678.

Doser, J. W., Finley, A. O., Saunders, S. P., Kéry, M., Weed, A. S., & Zipkin, E. F. (2025). Modeling complex species-environment relationships through spatially-varying coefficient occupancy models. *Journal of Agricultural, Biological and Environmental Statistics*, *30*(1), 146–171.

Dovers, E., Popovic, G. C., & Warton, D. I. (2024). A fast method for fitting integrated species distribution models. *Methods in Ecology and Evolution*, *15*(1), 191–203.

Duan, L. L., Johndrow, J. E., & Dunson, D. B. (2018). Scaling up data augmentation MCMC via calibration. *Journal of Machine Learning Research*, *19*(64), 1–34.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, *8*(6), 985–987.

Farr, M. T., Green, D. S., Holekamp, K. E., & Zipkin, E. F. (2021). Integrating distance sampling and presence-only data to estimate species abundance. *Ecology*, *102*(1), e03204.

Fithian, W., Elith, J., Hastie, T., & Keith, D. A. (2015). Bias correction in species distribution models: Pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, *6*(4), 424–438.

Fletcher, R. J., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A., & Dorazio, R. M. (2019). A practical guide for combining data to model species distributions. *Ecology*, *100*(6), e02710.

Fletcher, R. J., McCleery, R. A., Greene, D. U., & Tye, C. A. (2016). Integrated models that unite local and regional data reveal larger-scale environmental relationships and improve predictions of species distributions. *Landscape Ecology*, *31*, 1369–1382.

Frost, F., McCrea, R., King, R., Gimenez, O., & Zipkin, E. (2023). Integrated population models: Achieving their potential. *Journal of Statistical Theory and Practice*, *17*(1), 6.

Fukaya, K., Kusumoto, B., Shiono, T., Fujinuma, J., & Kubota, Y. (2020). Integrating multiple sources of ecological data to unveil macroscale species abundance. *Nature Communications*, *11*(1), 1695.

Gabry, J., Češnovar, R., Johnson, A., & Bronder, S. (2025). `cmdstanr`: `R` Interface to '`CmdStan`'. R package version 0.9.0. https://discourse.mc-stan.org

Gelfand, A. E., & Shirota, S. (2019). Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. *Ecological Monographs*, *89*(3), e01372.

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

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Griffin, J. E., Matechou, E., Buxton, A. S., Bormpoudakis, D., & Griffiths, R. A. (2020). Modelling environmental DNA data; Bayesian variable selection accounting for false positive and false negative errors. *Journal of the Royal Statistical Society: Series C: Applied Statistics*, *69*(2), 377–392.

Grimmer, J. (2011). An introduction to Bayesian inference via variational approximations. *Political Analysis*, *19*(1), 32–47.

Grüss, A., Charsley, A., Thorson, J., Anderson, O., O'Driscoll, R., Wood, B., Breivik, O., & O'Leary, C. (2023). Integrating survey and observer data improves the predictions of New Zealand spatio-temporal models. *ICES Journal of Marine Science*, *80*(7), 1991–2007.

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I. T., Regan, T. J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T. G., Rhodes, J. R., Maggini, R., Setterfield, S. A., Elith, J., Schwartz, M. W., Wintle, B. A., Broennimann, O., Austin, M., … Buckley, Y. M. (2013). Predicting species distributions for conservation decisions. *Ecology Letters*, *16*(12), 1424–1435.

Han, P., & Si, Y. (2025). Frontiers in data integration. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *188*(1), 24–26.

Hanks, E. M., Hooten, M. B., & Baker, F. A. (2011). Reconciling multiple data sources to improve accuracy of large-scale prediction of forest disease incidence. *Ecological Applications*, *21*(4), 1173–1188.

Hector, E. C., Tang, L., Zhou, L., & Song, P. X. (2024). Data integration and model fusion in the Bayesian and frequentist frameworks. In *Handbook of Bayesian, fiducial, and frequentist inference* (pp. 238–263). Chapman and Hall/CRC.

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

Hirzel, A. H., Hausser, J., Chessel, D., & Perrin, N. (2002). Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? *Ecology*, *83*(7), 2027–2036.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Hooten, M. B., & Hefley, T. J. (2019). *Bringing Bayesian models to life*. CRC Press.

Hooten, M. B., Johnson, D. S., & Brost, B. M. (2021). Making recursive Bayesian inference accessible. *The American Statistician*, *75*(2), 185–194.

Hudy, M., Thieling, T. M., Gillespie, N., & Smith, E. P. (2008). Distribution, status, and land use characteristics of subwatersheds within the native range of brook trout in the eastern United States. *North American Journal of Fisheries Management*, *28*(4), 1069–1085.

Hui, F. K., Warton, D. I., Ormerod, J. T., Haapaniemi, V., & Taskinen, S. (2017). Variational approximations for generalized linear latent variable models. *Journal of Computational and Graphical Statistics*, *26*(1), 35–43.

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., Jarvis, S., Lahoz-Monfort, J., Pagel, J., Pescott, O. L., Schmucki, R., Simmonds, E. G., & O'Hara, R. B. (2020). Data integration for large-scale models of species distributions. *Trends in Ecology & Evolution*, *35*(1), 56–67.

Johnson, D. S., Brost, B. M., & Hooten, M. B. (2022). Greater than the sum of its parts: Computationally flexible Bayesian hierarchical modeling. *Journal of Agricultural, Biological and Environmental Statistics*, *27*(2), 382–400.

Johnston, A., Matechou, E., & Dennis, E. B. (2023). Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods in Ecology and Evolution*, *14*(1), 103–116.

Jones, K. A., Niknami, L. S., Buto, S. G., & Decker, D. (2022). *Federal standards and procedures for the national watershed boundary dataset (WBD). U.S.* Geological Survey Techniques and Methods.

Kanno, Y., Vokoun, J., & Beauchene, M. (2010). Development of dual fish multi-metric indices of biological condition for streams with characteristic thermal gradients and low species richness. *Ecological Indicators*, *10*(3), 565–571.

Kanno, Y., Vokoun, J. C., & Letcher, B. H. (2011). Fine-scale population structure and riverscape genetics of brook trout (*Salvelinus fontinalis*) distributed continuously along headwater channel networks. *Molecular Ecology*, *20*(18), 3711–3729.

Kéry, M., Royle, J. A., Hallman, T., Robinson, W. D., Strebel, N., & Kellner, K. F. (2024a). Integrated distance sampling models for simple point counts. *Ecology*, *105*(5), e4292.

Kéry, M., Royle, J. A., Hallman, T., Robinson, W. D., Strebel, N., & Kellner, K. F. (2024b). Integrated distance sampling models for simple point counts. (v1.0). *Zenodo*. https://doi.org/10.5281/zenodo.10666980

Koshkina, V., Wang, Y., Gordon, A., Dorazio, R. M., White, M., & Stone, L. (2017). Integrated species distribution models: Combining presence-background data and site-occupancy data with imperfect detection. *Methods in Ecology and Evolution*, *8*(4), 420–430.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., & Bell, B. M. (2016). `TMB`: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, *70*, 1–21.

Kucukelbir, A., Ranganath, R., Gelman, A., & Blei, D. (2015). Automatic variational inference in Stan. In *Proceedings of the 29th International Conference on Neural Information Processing Systems – Volume 1*, *NIPS'15*, Montreal, Canada (pp. 568–576). MIT Press, Cambridge, MA, USA.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of Machine Learning Research*, *18*(14), 1–45.

Letcher, B. H., Nislow, K. H., Coombs, J. A., O'Donnell, M. J., & Dubreuil, T. L. (2007). Population response to habitat fragmentation in a stream-dwelling brook trout population. *PLoS One*, *2*(11), e1139.

Lindén, A., & Mäntyniemi, S. (2011). Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology*, *92*(7), 1414–1421.

Lindgren, F., & Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, *63*, 1–25.

Little, R. J. (2006). Calibrated Bayes: A Bayes/frequentist roadmap. *The American Statistician*, *60*(3), 213–223.

Lu, X., Kanno, Y., Valentine, G. P., Kulp, M. A., & Hooten, M. B. (2024). Regularized latent trajectory models for spatio-temporal population dynamics. *Journal of Agricultural, Biological, and Environmental Statistics*, *30*(3), 1–17.

Lu, X., Kanno, Y., Valentine, G. P., Rash, J. M., & Hooten, M. B. (2024). Using multi-scale spatial models of dendritic ecosystems to infer abundance of a stream salmonid. *Journal of Applied Ecology*, *61*(7), 1703–1715.

Lunn, D., Barrett, J., Sweeting, M., & Thompson, S. (2013). Fully Bayesian hierarchical modelling in two stages, with application to meta-analysis. *Journal of the Royal Statistical Society: Series C: Applied Statistics*, *62*(4), 551–572.

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2013). *The BUGS book: A practical introduction to Bayesian analysis*. Chapman Hall.

Lynch, H. J., Thorson, J. T., & Shelton, A. O. (2014). Dealing with under- and over-dispersed count data in life history, spatial, and community ecology. *Ecology*, *95*(11), 3173–3180.

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

Maxwell, S. L., Butt, N., Maron, M., McAlpine, C. A., Chapman, S., Ullmann, A., Segan, D. B., & Watson, J. E. (2019). Conservation implications of ecological responses to extreme weather and climate events. *Diversity and Distributions*, *25*(4), 613–625.

Meehan, T. D., Michel, N. L., & Rue, H. (2020). Estimating animal abundance with N-mixture models using the `R-INLA` package for `R`. *Journal of Statistical Software*, *95*, 1–26.

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.

Monnahan, C. C., Thorson, J. T., & Branch, T. A. (2017). Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution*, *8*(3), 339–348.

Morera-Pujol, V., Mostert, P. S., Murphy, K. J., Burkitt, T., Coad, B., McMahon, B. J., Nieuwenhuis, M., Morelle, K., Ward, A. I., & Ciuti, S. (2023). Bayesian species distribution models integrate presence-only and presence–absence data to predict deer distribution and relative abundance. *Ecography*, *2023*(2), e06451.

Mostert, P. S., & O'Hara, R. B. (2023). `PointedSDMs`: An `R` package to help facilitate the construction of integrated species distribution models. *Methods in Ecology and Evolution*, *14*(5), 1200–1207.

Niku, J., Hui, F. K., Taskinen, S., & Warton, D. I. (2019). `gllvm`: Fast analysis of multivariate abundance data with generalized linear latent variable models in `R`. *Methods in Ecology and Evolution*, *10*(12), 2173–2182.

Osgood-Zimmerman, A., & Wakefield, J. (2023). A statistical review of template model builder: A flexible tool for spatial modelling. *International Statistical Review*, *91*(2), 318–342.

Pacifici, K., Reich, B. J., Miller, D. A., Gardner, B., Stauffer, G., Singh, S., McKerrow, A., & Collazo, J. A. (2017). Integrating multiple data sources in species distribution modeling: A framework for data fusion. *Ecology*, *98*(3), 840–850.

Pacifici, K., Reich, B. J., Miller, D. A., & Pease, B. S. (2019). Resolving misaligned spatial data with integrated species distribution models. *Ecology*, *100*(6), e02709.

Pagel, J., Anderson, B. J., O'Hara, R. B., Cramer, W., Fox, R., Jeltsch, F., Roy, D. B., Thomas, C. D., & Schurr, F. M. (2014). Quantifying range-wide variation in population trends from local abundance surveys and widespread opportunistic occurrence records. *Methods in Ecology and Evolution*, *5*(8), 751–760.

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

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., & Pélissier, R. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, *11*(1), 4540.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*. March, 20–22. http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Drafts/Plummer.pdf

Polson, N. G., & Scott, J. G. (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, *7*(4), 887–902.

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.

Primack, R. B. (2006). *Essentials of conservation biology* (Vol. *23*). Sinauer Associates Sunderland.

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G., & Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, *6*(4), 366–379.

Renner, I. W., Louvrier, J., & Gimenez, O. (2019). Combining multiple data sources in species distribution models while accounting for spatial dependence and overfitting with combined penalized likelihood maximization. *Methods in Ecology and Evolution*, *10*(12), 2118–2128.

Richards, S. A. (2008). Dealing with overdispersed count data in applied ecology. *Journal of Applied Ecology*, *45*(1), 218–227.

Robinson, O. J., Ruiz-Gutierrez, V., Reynolds, M. D., Golet, G. H., Strimas-Mackey, M., & Fink, D. (2020). Integrating citizen science data with expert surveys increases accuracy and spatial extent of species distribution models. *Diversity and Distributions*, *26*(8), 976–986.

Royle, J. A., & Dorazio, R. M. (2012). Parameter-expanded data augmentation for Bayesian analysis of capture–recapture models. *Journal of Ornithology*, *152*(2), 521–537.

Royle, J. A., Karanth, K. U., Gopalaswamy, A. M., & Kumar, N. S. (2009). Bayesian inference in camera trapping studies for a class of spatial capture–recapture models. *Ecology*, *90*(11), 3233–3244.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *71*(2), 319–392.

Schaub, M., & Kéry, M. (2021). *Integrated population models: Theory and ecological applications with R and JAGS*. Academic Press.

Schindler, A. R., Cunningham, S. A., Schafer, T. L., Sinnott, E. A., Clements, S. J., DiDonato, F. M., Mosloff, A. R., Walters, C. M., Shipley, A. A., Weegman, M. D., & Qing, Z. (2022). Joint analysis of structured and semi-structured community science data improves precision of relative abundance but not trends in birds. *Scientific Reports*, *12*(1), 20289.

Schliep, E. M., Gelfand, A. E., Clark, C. W., Mayo, C. A., McKenna, B., Parks, S. E., Yack, T. M., & Schick, R. S. (2024). Assessing marine mammal abundance: A novel data fusion. *The Annals of Applied Statistics*, *18*(4), 3071–3090.

Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J., & O'Hara, R. B. (2020). Is more data always better? A simulation study of benefits and limitations of integrated distribution models. *Ecography*, *43*(10), 1413–1422.

Stillman, A. N., Howell, P. E., Zimmerman, G. S., Bjerre, E. R., Millsap, B. A., Robinson, O. J., Fink, D., Stuber, E. F., & Ruiz-Gutierrez, V. (2023). Leveraging the strengths of citizen science and structured surveys to achieve scalable inference on population size. *Journal of Applied Ecology*, *60*(11), 2389–2399.

Strebel, N., Kéry, M., Guélat, J., & Sattler, T. (2022). Spatiotemporal modelling of abundance from multiple data sources in an integrated spatial distribution model. *Journal of Biogeography*, *49*(3), 563–575.

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., Damoulas, T., Dhondt, A. A., Dietterich, T., Farnsworth, A., Fink, D., Fitzpatrick, J. W., Fredericks, T., Gerbracht, J., Gomes, C., Hochachka, W. M., Iliff, M. J., Lagoze, C., La Sorte, F. A., … Kelling, S. (2014). The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, *169*, 31–40.

Tirronen, M., & Kuparinen, A. (2024). Parameter estimation for allometric trophic network models: A variational Bayesian inverse problem approach. *Methods in Ecology and Evolution*, *15*(12), 2373–2384.

Turek, D., de Valpine, P., Paciorek, C., & Anderson-Bergman, C. (2017). Automated parameter blocking for efficient Markov chain Monte Carlo sampling. *Bayesian Analysis*, *12*(2), 465–490.

Twining, J. P., Fuller, A. K., Sun, C. C., Calderón-Acevedo, C. A., Schlesinger, M. D., Berger, M., Kramer, D., & Frair, J. L. (2024). Integrating presence–only and detection/non-detection data to estimate distributions and expected abundance of difficult-to-monitor species on a landscape-scale. *Journal of Applied Ecology*, *61*(6), 1441–1459.

Valentine, G. P., Lu, X., Childress, E. S., Dolloff, C. A., Hitt, N. P., Kulp, M. A., Letcher, B. H., Pregler, K. C., Rash, J. M., Hooten, M. B., & Kanno, Y. (2024). Spatial asynchrony and cross-scale climate interactions in populations of a coldwater stream fish. *Global Change Biology*, *30*(1), e17029.

Van Ee, J. (2026). Data and code for "Geometric overdispersion facilitates the integration of ecological data".

Van Ee, J., Hagen, C., Pavlacky, D., Fricke, K., Koslovsky, M., & Hooten, M. (2023). Melding wildlife surveys to improve conservation inference. *Biometrics*, *79*(4), 3941–3953.

Van Ee, J. J., Hagen, C. A., Pavlacky, D. C., Jr., Haukos, D. A., Lawrence, A. J., Tanner, A. M., Grisham, B. A., Fricke, K. A., Rossi, L. G., Beauprez, G. M., Kuklinski, K. E., Martin, R. L., Koslovsky, M. D., Rintz, T. B., & Hooten, M. B. (2024). Melded integrated population models. *Journal of Agricultural, Biological, and Environmental Statistics*, *30*, 769–799.

Ver Hoef, J. M., & Peterson, E. E. (2010). A moving average approach for spatial statistical models of stream networks. *Journal of the American Statistical Association*, *105*(489), 6–18.

Ver Hoef, J. M., Peterson, E. E., Hooten, M. B., Hanks, E. M., & Fortin, M.-J. (2018). Spatial autoregressive models for statistical inference from ecological data. *Ecological Monographs*, *88*(1), 36–59.

Vu, Q., Hui, F. K., Welsh, A., Muller, S., Cantoni, E., & Haak, C. R. (2025). A single index approach to integrated species distribution modeling for fisheries abundance data. *arXiv* 2509.15379.

Waldock, C., Stuart-Smith, R. D., Albouy, C., Cheung, W. W., Edgar, G. J., Mouillot, D., Tjiputra, J., & Pellissier, L. (2022). A quantitative review of abundance-based species distribution models. *Ecography*, *2022*(1), e05694.

Warton, D. I., & Shepherd, L. C. (2010). Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *The Annals of Applied Statistics*, *4*(3), 1383–1402.

Watson, J., Joy, R., Tollit, D., Thornton, S. J., & Auger-Méthé, M. (2021). Estimating animal utilization distributions from multiple data types: A joint spatiotemporal point process framework. *The Annals of Applied Statistics*, *15*(4), 1872–1896.

Winfree, R. W., Fox, J., Williams, N. M., Reilly, J. R., & Cariveau, D. P. (2015). Abundance of common species, not species richness, drives delivery of a real-world ecosystem service. *Ecology Letters*, *18*(7), 626–635.

Wyatt, R. J. (2002). Estimating riverine fish population size from single- and multiple-pass removal sampling using a hierarchical model. *Canadian Journal of Fisheries and Aquatic Sciences*, *59*(4), 695–706.

Zhang, L., Carpenter, B., Gelman, A., & Vehtari, A. (2022). Pathfinder: Parallel quasi-Newton variational inference. *Journal of Machine Learning Research*, *23*(306), 1–49.

Zhang, W., Bravington, M., & Fewster, R. (2019). Fast likelihood-based inference for latent count models using the saddlepoint approximation. *Biometrics*, *75*(3), 723–733.

Zhao, Q., Latif, Q. S., Nuse, B. L., Pavlacky, D. C., Jr., Kilner, C. L., Ryder, T. B., & Latimer, C. E. (2024). Integrating counts from rigorous surveys and participatory science to better understand spatiotemporal variation in population processes. *Methods in Ecology and Evolution*, *15*(8), 1380–1393.

Zipkin, E. F., Zylstra, E. R., Wright, A. D., Saunders, S. P., Finley, A. O., Dietze, M. C., Itter, M. S., & Tingley, M. W. (2021). Addressing data integration challenges to link ecological processes across scales. *Frontiers in Ecology and the Environment*, *19*(1), 30–38.

SUPPORTING INFORMATION

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

Figure S1. Directed acyclic graph of integrated distance sampling model.

Figure S2. Posterior distributions of brook trout covariate associations inferred using different data types.

How to cite this article: Van Ee, J. J., Kanno, Y., Lu, X., Rash, J. M., & Hooten, M. B. (2026). Geometric overdispersion facilitates the integration of ecological data. *Methods in Ecology and Evolution*, *17*, 2104–2122. https://doi.org/10.1111/2041-210x.70329