Received: 1 May 2023 | Revised: 11 March 2024 | Accepted: 12 April 2024

DOI: 10.1002/ecy.4326

STATISTICAL REPORT

ECOLOGY
ECOLOGICAL SOCIETY OF AMERICA

A multispecies hierarchical model to integrate count and distance-sampling data

Neil A. Gilbert[1,2] | **Caroline M. Blommel**[2,3] | **Matthew T. Farr**[1,2,4] | **David S. Green**[5] | **Kay E. Holekamp**[1,2] | **Elise F. Zipkin**[1,2]

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

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

[3]Department of Fish, Wildlife, and Conservation Biology, Colorado State University, Fort Collins, Colorado, USA

[4]Washington Cooperative Fish and Wildlife Research Unit, School of Aquatic and Fishery Sciences, University of Washington, Seattle, Washington, USA

[5]Institute for Natural Resources, Portland State University, Portland, Oregon, USA

Correspondence
Neil A. Gilbert
Email: neil.allen.gilbert@gmail.com

Funding information
Directorate for Biological Sciences, Grant/Award Numbers: 2208894, DBI-1954406, DEB-1755089, IOS-1755089, OIS-1853934

Handling Editor: Brian D. Inouye

Abstract

Integrated community models—an emerging framework in which multiple data sources for multiple species are analyzed simultaneously—offer opportunities to expand inferences beyond the single-species and single-data-source approaches common in ecology. We developed a novel integrated community model that combines distance sampling and single-visit count data; within the model, information is shared among data sources (via a joint likelihood) and species (via a random-effects structure) to estimate abundance patterns across a community. Parameters relating to abundance are shared between data sources, and the model can specify either shared or separate observation processes for each data source. Simulations demonstrated that the model provided unbiased estimates of abundance and detection parameters even when detection probabilities varied between the data types. The integrated community model also provided more accurate and more precise parameter estimates than alternative single-species and single-data-source models in many instances. We applied the model to a community of 11 herbivore species in the Masai Mara National Reserve, Kenya, and found considerable interspecific variation in response to local wildlife management practices: Five species showed higher abundances in a region with passive conservation enforcement (median across species: 4.5× higher), three species showed higher abundances in a region with active conservation enforcement (median: 3.9× higher), and the remaining three species showed no abundance differences between the two regions. Furthermore, the community average of abundance was slightly higher in the region with active conservation enforcement but not definitively so (posterior mean: higher by 0.20 animals; 95% credible interval: 1.43 fewer animals, 1.86 more animals). Our integrated community modeling framework has the potential to expand the scope of inference over space, time, and levels of biological organization, but practitioners should carefully evaluate whether model assumptions are met in their systems and whether data integration is valuable for their applications.

KEYWORDS
abundance, Bayesian, data integration, distance sampling, hierarchical modeling

INTRODUCTION

Improved and expanded inferences on biodiversity patterns and trends are required to address concerns over widespread species losses (Cardinale et al., 2012; Dirzo et al., 2014). Although there is much value in single-species analyses, syntheses across taxa may grant managers and policymakers a broader perspective when enacting conservation decisions. Similarly, syntheses across data sources may enhance confidence—or avoid overconfidence—in conservation decisions, especially as data are limited for many species worldwide (Borgelt et al., 2022; IUCN, 2022). Despite the benefits of cross-species and cross-data synthesis, most ecological analyses focus on a single species or a single data source, likely because of challenges inherent in implementing complex analyses. However, two innovations in statistical ecology offer a path forward for the unified analysis of communities using multiple data sources: hierarchical community models and model-based data integration.

Hierarchical community models estimate species-specific parameters under the assumption that they come from shared community-level parameters, thereby permitting synthesis across species (Figure 1; Devarajan et al., 2020; Dorazio et al., 2006). Community models are structured such that species-specific parameters (e.g., intercepts, covariate coefficients) are treated as random effects drawn from community-level distributions (Kéry & Royle, 2015; Zipkin et al., 2009). Biologically, this model structure implies that species within communities show similar abundance patterns. For example, land-cover change from grassland to forest may drive declines for species that use grasslands—though not necessarily all such species. Community models can accommodate this type of heterogeneity while also providing inferences on the mean community-level response to a covariate such as land-cover change. Hierarchical community models hold two primary advantages over single-species approaches. First, community models expand the scope of inference by providing both species- and community-level inferences that summarize the community average as well as among-species variation in ecological quantities. Thus, community-level patterns (e.g., community-level responses to a covariate) can be succinctly reported with quantitative metrics and associated uncertainty, in contrast to post hoc summaries of multiple single-species models (Kéry & Royle, 2015). Second, community models can provide more accurate and precise inferences than single-species models, particularly for data-poor species, as parameters for these species are informed by data across the community (Guillera-Arroita, 2017; Kéry & Royle, 2015; Zipkin et al., 2009). Importantly, hierarchical community

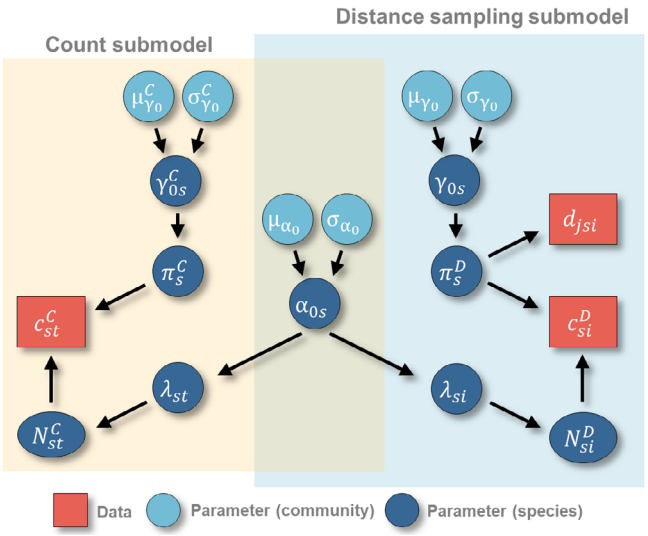

FIGURE 1 Graphical overview of integrated community model that combines distance sampling and count data. The parameters within the overlapping rectangles are shared between data sources. d_{jsi} is the distance measurements to animals from the distance sampling surveys. c^C_{st} and c^D_{si} are the number of animals observed from the single-visit counts and distance sampling data, respectively, while N^C_{st} and N^D_{si} are the corresponding latent abundances and λ_{st} and λ_{si} the corresponding expected abundances. α_{0s} is the species-level intercept for expected abundance and is modeled from a community distribution with mean μ_{α_0} and SD σ_{α_0}. π^C_s and π^D_s are transect-level detection probabilities for the count and distance-sampling data, respectively, and are modeled via detection functions that include scale parameter intercepts γ^C_{0s} (counts) and γ_{0s} (distance sampling). These species-level intercepts are drawn from community-level distributions with means $\mu^C_{\gamma_0}$ (counts) and μ_{γ_0} (distance sampling) and SDs $\sigma^C_{\gamma_0}$ (counts) and σ_{γ_0} (distance sampling). An updated graphical overview of the model for the case study is provided in Appendix S1: Figure S5.

modeling is a framework rather than a specific model; many types of ecological models can be analyzed using a community modeling approach, including occupancy models (Dorazio & Royle, 2005), N-mixture models (Yamaura et al., 2012), and distance-sampling models (Farr et al., 2019; Sollmann et al., 2016).

Integrated models combine multiple data sources within a single analytical framework and provide opportunities to expand the scope of inference by estimating parameters that are unidentifiable in analyses of individual data sources (Zipkin et al., 2019). Perhaps the most widely recognized type of integrated model is the integrated population model in which demographic data (e.g., mark–recapture) are combined with count data to make inferences about population dynamics (Schaub & Kéry, 2021; Zipkin & Saunders, 2018). However, data integration can be performed with any number of data sources (e.g., counts, presence-only data, detection–nondetection surveys); the

common theme is the combination of multiple data sources to expand or improve inferences (Schaub & Kéry, 2021). Integrated models combine distinct data sources via a joint likelihood, analyzing each data source with a distinct submodel (rather than indiscriminately pooling discrepant data) to account for differences between data sources (Fletcher et al., 2019; Isaac et al., 2020). Describing each data set with a submodel means that integrated models can combine different data sources (e.g., count and presence–absence data) or data with mismatched spatiotemporal resolution (Gilbert et al., 2021; Isaac et al., 2020; Pacifici et al., 2019). Beyond the conceptual appeal of applying all available sources of information to ecological problems, integrated models can provide more precise inferences than nonintegrated models due to the greater amount of available data (Fletcher et al., 2019; Gilbert et al., 2021). Limitations of data integration include increased model complexity (which may preclude use by many practitioners) and compromised estimation and prediction of key ecological parameters (e.g., abundance, occurrence) if large volumes of biased, low-quality data swamp smaller amounts of high-information-content data (Simmonds et al., 2020). All told, data integration remains an active area of research (Isaac et al., 2020) as ecologists identify new combinations of data sources to integrate and move beyond single-species perspectives (Doser et al., 2022).

Hierarchical community models and data integration have both advanced biodiversity analyses but have developed in separate spheres. Integrated community models—that is, hierarchical community models that incorporate multiple data sources—are an appealing prospect that combine the benefits of integrated and community modeling (Barraquand & Gimenez, 2019; Doser et al., 2022; Péron & Koons, 2012; Quéroué et al., 2021; Zipkin et al., 2023). In this study, we developed an integrated community model that combines distance sampling (a high-information data source that permits estimation of detection probability) and single-visit counts (a low-information data source containing no information on the detection process). Despite recent developments in community distance-sampling models using a single data source (Farr et al., 2019; Goyert et al., 2016; Sollmann et al., 2016) and single-species integrated models that combine distance-sampling and other data sources (Farr et al., 2021; Kéry et al., 2022; Schmidt et al., 2022; Schmidt & Robison, 2020; Zhao et al., 2021), community-level integrated models coupling distance sampling with other data sources have yet to be described or applied. Potential benefits of this modeling approach include estimating ecological quantities (e.g., covariate effects) with greater precision, particularly for data-poor species, and estimating latent abundances for multiple species in locations where only single-visit count data are available. We developed a simulation study to evaluate our model's performance when detection probabilities vary between count and distance-sampling data. We also present a case study in which we applied the model to count and distance-sampling data sets on a community of 11 herbivore species from the Masai Mara National Reserve in Kenya. A rise in anthropogenic activity has led to the decline of both resident and migrant herbivores in parts of the reserve (Green et al., 2019), yet species-specific responses to disturbance along with population abundances have yet to be reported for this community. Our integrated community model provides a method to estimate herbivore abundances and the effects of disturbance in the reserve by combining distance sampling and transect count data.

METHODS

Model description

Overview

The key aspect of the integrated community model is that information is shared between data sources (via a joint likelihood) and among species (via a random effects structure; Figure 1). We specify a species-specific model of abundance and integrate distance sampling and single-visit count data such that both data sources inform the parameters underpinning abundance (Farr et al., 2021; Kéry et al., 2022). The observation process can be modeled such that detection probability is shared between data types (requiring a strong assumption about identical observation processes between data types) or separate for the two data types via a latent detection function for the single-visit count data (Kéry et al., 2022). Relevant covariates can be incorporated into the models for both abundance (e.g., habitat) and detection probability (e.g., species-specific traits, time of day). Information is also shared among species by assuming that each species-level parameter in both the abundance and detection model components come from parameter-specific, community-level distributions (Kéry & Royle, 2015). By specifying an integrated community model, inferences can be made for rare or infrequently detected species for which single-species and/or single-data-source models provide highly uncertain inference or fail to converge. We describe the model below and how to apply it using a Bayesian framework. As a Bayesian model, prior distributions are required for model parameters; we provide suggestions for choosing appropriate priors for some of these to guide future practitioners. Throughout this model description, we use D superscripts to distinguish parameters

related to distance sampling from parameters related to the count submodel, indicated with *C* superscripts.

Distance sampling submodel

Distance sampling: Data

Distance sampling is an established method of estimating wildlife abundance via the assumption that the detectability of a focal individual (or group of animals) decays with distance (Buckland et al., 2001). The required data are (1) the count of animals detected during a survey and (2) the distances measured from the observers to each animal or group of animals. We describe the model using distance bands to accommodate scenarios in which observers collect categorical distance observations (e.g., animals classified as being within, for instance, 0–100 or 101–200 m of observers); for cases in which observers collect continuous distance measurements, distance measurements can be categorized into distance bands post hoc. Selecting the number of bins is an arbitrary decision; fewer coarse bands leads to less precise estimates of abundance, but increasing the number of bands comes with the trade-off of increased computational burden (Kéry & Royle, 2015). Distance sampling does not require temporal replication since detection probability is inferred from the observed distances (rather than from temporally replicated surveys, as in other abundance estimation methods; Royle, 2004). Therefore, we present a context in which multiple species (indexed with *s*; Appendix S1: Table S1) are surveyed using distance sampling across multiple transects (indexed with *i*). The transects are assumed to be independent. While we describe the model for use with data collected along line transects (e.g., along trails or roads), our approach can be adapted for use with data collected at points (e.g., avian point counts) with a modification of the detection function; we refer readers to chapter 8.5 of Kéry and Royle (2015) for a demonstration of how this is performed.

Distance sampling: Observation model

In the distance sampling submodel, the detection probability for an animal or group of animals follows a half-normal function (see Buckland et al., 2001 for alternative detection functions) of the group's distance to the observer and a scale parameter that governs how quickly detectability decays over distance. Distances to animals are classified into $k = 1,...,K$ distance categories. Detection probability π_{ks} within distance category k for species s is a half-normal function of the distance category's midpoint x_k and scale parameter ω_s, multiplied by each bin's proportion of the total width of the surveyed area P_k:

$$\pi_{ks} = \exp\left(-\frac{x_k^2}{2\omega_s^2}\right) \times P_k. \quad (1)$$

In the simplest case, the scale parameter ω_s is modeled using only a species-specific intercept, γ_{0s}:

$$\log(\omega_s) = \gamma_{0s}. \quad (2)$$

However, covariates that describe variation in detection over space (e.g., vegetation height), time (e.g., weather), or among species (e.g., body mass) could be added. The species-specific intercepts are treated as random variables drawn from a community-level distribution:

$$\gamma_{0s} \sim \text{Normal}\left(\mu_{\gamma_0}, \sigma_{\gamma_0}\right), \quad (3)$$

where μ_{γ_0} is the average decay rate (on a log scale) across the community of species, and σ_{γ_0} is the SD among species-specific intercepts. Priors are required for these community-level hyperparameters. For μ_{γ_0}, we suggest using weakly informative priors based on prior predictive checks using the distances over which data are collected, since what constitutes an appropriate prior depends on the scale of the distances used in Equation (1) (Kéry & Royle, 2015). For example, a Normal(6, 2) prior would be appropriate for moderately detectable taxa surveyed out to 1000 m (Appendix S1: Figure S3). For σ_{γ_0}, an Exponential (1) prior would be appropriate for a community in which among-species variation in detectability is moderate or poorly known (McElreath, 2020). Detection probability for the whole transect, π_s, is the sum of the detection probabilities for each distance category:

$$\pi_s = \sum_{k=1}^{K} \pi_{ks}. \quad (4)$$

The distance class d_{jsi} of the observed group or individual j of species s at transect i is modeled using a categorical distribution:

$$d_{jsi} \sim \text{Categorical}(\boldsymbol{\nu}_\mathbf{s}), \quad (5)$$

where $\boldsymbol{\nu}_\mathbf{s}$ is a vector of length K that holds the conditional cell probabilities. Cell probabilities in a categorical distribution must sum to one; thus, the probability of an observation being in distance category k, $\boldsymbol{\nu}_\mathbf{s}$, is constrained as

the detection probability within each distance class divided by the sum of detection probabilities of all distance classes:

$$\boldsymbol{\nu}_{\mathbf{s}} = \frac{\pi_{ks}}{\pi_s}. \quad (6)$$

Finally, the count c_{si}^D is modeled using a binomial distribution, with the number of trials being the latent abundance N_{si}^D and with detection probability π_s:

$$c_{si}^D \sim \text{Binomial}\left(N_{si}^D, \pi_s\right). \quad (7)$$

Distance sampling: Latent abundance model

Latent abundance N_{si}^D is modeled as a Poisson random variable (Equation 8). For cases where abundance shows evidence of overdispersion (i.e., greater variation than expected with a Poisson distribution in which the mean and variance are equal), alternative forms such as a Poisson–gamma mixture (also known as the negative binomial distribution) or a Poisson–lognormal distribution can be used (Kéry & Royle, 2015); see the case study for an example.

$$N_{si}^D \sim \text{Poisson}(\lambda_{si}), \quad (8)$$

where λ_{si} is the expected abundance for species s at transect i. Expected abundance can be modeled on a log scale as a function of covariates:

$$\log(\lambda_{si}) = \alpha_{0s} + \alpha_{1s} z_i, \quad (9)$$

where α_{0s} is the intercept and α_{1s} is the species-specific effect of transect-level covariate z_i. As with the detection model, the species-level parameters in the abundance model are assumed to be random variables drawn from community-level distributions:

$$\alpha_{0s} \sim \text{Normal}\left(\mu_{\alpha_0}, \sigma_{\alpha_0}\right), \quad (10)$$

$$\alpha_{1s} \sim \text{Normal}\left(\mu_{\alpha_1}, \sigma_{\alpha_1}\right). \quad (11)$$

Finally, priors are required for the community-level hyperparameters. While prior choice may vary depending on existing information about the system or study objectives, we suggest using common weakly informative priors such as Normal(0, 2) priors for the community means and Exponential(1) priors for the SDs (McElreath, 2020).

Count submodel

Counts: Data

Count data without additional information necessary to estimate detection probability (e.g., distance measurements, temporal replication) are commonly collected, for example, by many volunteer-based science programs. Data for the count submodel are thus total counts of species in a predefined sampling unit, which are assumed to be independent across space. The submodel for single-visit count data permits abundance estimation by integrating parameters related to abundance informed by the distance sampling submodel (Figure 1). Below, we use t to index transects to emphasize that the count data need not come from the same transects as the distance sampling data (indexed with i).

Counts: Observation model

Since the count data contain no information about the detection process, detection probability must either be borrowed from the distance sampling submodel (see case study) or estimated with a latent detection function (Kéry et al., 2022). Borrowing detection probability from the distance sampling submodel is only appropriate if detection probability can be assumed to be constant between data types and the maximum distance to which animals are surveyed when count data are known. If these assumptions are not met, the model-based approach we describe here offers a flexible approach to the common situation in which the observation process varies between data types. With this approach, count data are viewed as latent distance-sampling surveys in which distances are not collected, but critical assumptions—like detectability decaying over distance—still apply (Kéry et al., 2022). Below, we use C superscripts to differentiate parameters related to the count data from similar parameters in the distance-sampling submodel. Detection probability π_{ks}^C within distance category k for species s is a half-normal function of the distance category's midpoint x_k and scale parameter τ_s, multiplied by each bin's proportion of the total width of the surveyed area P_k^C:

$$\pi_{ks}^C = \exp\left(-\frac{x_k^2}{2\tau_s^2}\right) \times P_k^C. \quad (12)$$

As in the distance-sampling submodel (Equation 2), we describe the simplest case of the scale parameter τ_s being modeled with only a species-specific intercept, γ_{0s}^C:

$$\log(\tau_s) = \gamma_{0s}^C. \quad (13)$$

The species-level intercept γ_{0s}^C is modeled as a draw from a community-level distribution with hyperparameters $\mu_{\gamma_0}^C$ and $\sigma_{\gamma_0}^C$:

$$\gamma_{0s}^C \sim \text{Normal}\left(\mu_{\gamma_0}^C, \sigma_{\gamma_0}^C\right). \quad (14)$$

Detection probability for the whole transect, π_s^C, is the sum of the detection probabilities for each distance category:

$$\pi_s^C = \sum_{k=1}^{K} \pi_{ks}^C. \quad (15)$$

Finally, the count c_{st}^C of species s at transect t is modeled with a binomial distribution:

$$c_{st}^C \sim \text{Binomial}\left(N_{st}^C, \pi_s^C\right), \quad (16)$$

where N_{st}^C is the latent abundance of species s at transect t.

Counts: Latent abundance model

The latent abundance N_{st}^C of species s at transect t is modeled with a Poisson distribution:

$$N_{st}^C \sim \text{Poisson}(\lambda_{st}), \quad (17)$$

in which expected abundance λ_{st} is estimated via shared parameters with the latent abundance component of the distance-sampling submodel:

$$\log(\lambda_{st}) = \alpha_{0s} + \alpha_{1s} z_t. \quad (18)$$

Note that the intercept α_{0s} and covariate coefficient α_{1s} are shared with the distance-sampling submodel (Equation 9). Thus, abundance for each surveyed transect (via either distance sampling or counts) is estimated using the shared information across both data sources.

Model assumptions

Our integrated community model makes several assumptions that are important to understand in its application. First, the model makes the same assumptions as those found in traditional distance-sampling models, namely, that (1) animals are distributed uniformly in the surveyed space, (2) detection probability equals one on the transect (i.e., distance = 0), (3) individuals (or groups) are detected at their original location, and (4) distances are measured without error (Buckland et al., 2001; Kéry & Royle, 2015). In addition, data are assumed to be independent, both among spatial replicates within one data source and across data sources. Practically speaking, this assumption would be violated if neighboring transects "shared" animals—in the most extreme case, if all the animals from Transect A moved to Transect B between surveys. Next, the model assumes no errors in species identifications (i.e., no false positives). Finally, an important assumption is that shared parameters between data sources (Figure 1) reflect commonalities in the observational and ecological processes. Parameters underpinning abundance patterns—such as the intercepts and slopes in the equation for expected abundance—are shared and thus assumed to be the same between data sources. This is akin to a closure assumption: unless a model structure is added (e.g., covariates, overdispersion parameters, or parameters for abundance losses and gains), any changes in abundance between surveys violates the assumption. Similarly, if using the simplest form of the model in which detection probability is shared from the distance-sampling submodel (see case study), detection probability is assumed to be the same across the two types of data.

Simulation study

We developed a simulation study to evaluate the performance of the integrated community model. Our goal was to assess inferences under the realistic scenario of having greater amounts of count than distance-sampling data and of differing observation processes between data sources. We also compared the model's performance to alternative nonintegrated and single-species models.

We simulated communities of 15 species, choosing data-generating parameter values that paralleled the case study (see below). We designed the simulation such that the distance-sampling and count data came from different transects (indexed with i and t, respectively) within the same domain, meaning that abundance patterns (i.e., average abundances and abundance–covariate relationships) were identical between transects with the two data sources. However, we simulated twice as much count data (100 transects) as distance-sampling data (50 transects) under the assumption that it was more common to have more of the easier-to-collect count data than the labor-intensive distance-sampling data. We simulated the latent abundance of each species at a transect as a draw from a Poisson distribution. We simulated $\lambda_{si/t}$, the

expected abundance for species s at transect i or t (Equation 9), as a function of an intercept and one covariate. We simulated species-specific intercepts (i.e., α_{0s} in Equation 9) as Normal$\left(\mu_{\alpha_0} = 0.87, \sigma_{\alpha_0} = 1.95\right)$. Similarly, we drew each species' covariate effect (α_{1s} in Equation 10) from a Normal$\left(\mu_{\alpha_1} = 0.05, \sigma_{\alpha_1} = 0.25\right)$ distribution, meaning that species' abundances tended to show weak positive relationships with the simulated covariate but with considerable variation (Appendix S1: Figure S1).

To simulate distance-sampling data, we first divided the simulated true abundances of each species into randomly sized groups to mimic the case study in which we measured distances to groups of animals (rather than individuals). We then randomly assigned distances up to 1000 m to each simulated group at distance sampling transects $i = 1,...,50$. We simulated species' scale parameter ω_s (which governs how quickly detection decays over distance; Equation 1) as a function of intercept γ_{0s} (Equation 2), which we drew from a Normal$\left(\mu_{\gamma_0} = 5.5, \sigma_{\gamma_0} = 0.25\right)$ distribution. For reference, with intercept $\gamma_{0s} = 5.5$, detection probability would be 0.59 at 250 m and 0.12 at 500 m (Appendix S1: Figure S2). We simulated the detection of each group as a Bernoulli trial with probability dependent on the distance to the group and the species' detection function (Appendix S1: Figure S2). Finally, we computed the observed count for a transect as the sum of the sizes of observed groups.

We simulated count data on transects $t = 1, ..., 100$ in the same way (thus treating single-visit counts as latent distance sampling surveys) (Kéry et al., 2022), except we drew the scale parameter's intercept γ_{0s} from a Normal$\left(\mu_{\gamma_0} = 5.0, \sigma_{\gamma_0} = 0.25\right)$ distribution. For reference, with intercept $\gamma_{0s} = 5.0$, the detection probability would be 0.24 at 250 m and 0.003 at 500 m (Appendix S1: Figure S2). Thus, the detection probability for the count data tended to be lower than the distance-sampling data, which may be common in unstructured or semistructured monitoring in which generating count data is not the primary focus of data collection (Johnston et al., 2022). Finally, we derived the observed count as the sum of the sizes of observed groups and subsequently discarded the distance observations. We treated the two data sources as coming from transects of equal spatial extent (i.e., assuming animals were counted within 1000 m of transects for both distance sampling and count data).

Additionally, we evaluated the following five alternative models that used a single data source and/or included only one species: community distance sampling, community count only, single-species integrated, single-species distance sampling, and single-species count only (see Appendix S2 for details on each of these). We used the same parameter settings to simulate data for each of these alternative models. For the single-species models, we only ran models for the rarest (species observed at least once with the lowest sum of observed counts) and most common species (species with the highest sum of observed counts) in the simulated community, since rare species may be sensitive to biases induced through shrinkage of model estimates toward community averages (Harrison et al., 2018).

We fit the models with Nimble (de Valpine et al., 2017) through R (R Core Team, 2023), using three Markov chain Monte Carlo (MCMC) chains run for 200,000 iterations, discarding the initial 100,000 as burn-in and thinning by 100. Models would have converged with fewer iterations (likely half the number used), but we wished to be conservative, and runtimes were short (~30 min) with the settings used. We used Normal(0, 2) priors for abundance intercepts and covariate coefficients, Exponential(1) priors for SD parameters, and an uninformative Uniform(0,10) prior for the scale parameter intercept in the detection function (Appendix S1: Figure S3; McElreath, 2020). We ran 1000 replicate simulations for each model and considered chains with convergence diagnostic $\widehat{R}$ <1.1 to have converged (Brooks & Gelman, 1998). We assessed accuracy by visualizing the distribution of absolute bias ($\overline{x} - x$, where $\overline{x}$ is the posterior mean for a parameter estimate and x is the true value) and interpret relative bias across replicates as $100\frac{\sum \overline{x} - x}{\sum |x|}$. We evaluated model precision by calculating the coefficient of variation as $\frac{s}{|\overline{x}|}$, where s is the posterior SD of parameter estimate. Scripts to reproduce the simulations are archived on Zenodo in Gilbert (2024).

Case study: Herbivores in Masai Mara National Reserve

We applied the model to a community of 11 resident herbivore species (Table 1) from the Masai Mara National Reserve in southwestern Kenya (Appendix S1: Figure S4; henceforth "reserve"). Herbivore declines in the region are associated with rising anthropogenic pressures, such as livestock grazing and tourism, which vary in intensity within the reserve (Green et al., 2019). The reserve is unfenced, 1510 km^2 in size, dominated by grassland, and divided into two zones managed by different entities (Appendix S1: Figure S4). The Mara Triangle is managed by a not-for-profit that actively enforces rules restricting grazing and poaching (Walpole & Leader-Williams, 2001). In contrast, the Talek region is managed by local government; similar laws restricting grazing and poaching exist but are only passively enforced (Green et al., 2018, 2019). The Talek region is also adjacent to Talek town

TABLE 1 The 11 species of herbivores from the case study in the Masai Mara National Reserve, Kenya.

Common name	Scientific name	Distance sampling	Counts
African buffalo	*Syncerus caffer*	46.1 ± 118.0	2.5 ± 17.4
Eland	*Taurotragus oryx*	4.8 ± 16.9	0.2 ± 1.7
Elephant	*Loxodonta africana*	16.7 ± 30.4	0.5 ± 2.2
Giraffe	*Giraffa camelopardalis*	6.6 ± 8.8	0.1 ± 1.0
Grant's gazelle	*Nanger granti*	6.2 ± 17.1	0.7 ± 2.2
Hartebeest	*Alcelaphus buselaphus*	3.4 ± 7.6	0.3 ± 1.5
Impala	*Aepyceros melampus*	142.0 ± 206.0	7.0 ± 17.4
Thomson's gazelle	*Eudorcas thomsonii*	261.0 ± 526.0	29.7 ± 60.7
Topi	*Damaliscus lunatus*	158 ± 223.0	13.7 ± 31.9
Warthog	*Phacochoerus africanus*	19.9 ± 27.1	1.6 ± 3.1
Waterbuck	*Kobus ellipsiprymnus*	9.4 ± 26.4	1.5 ± 6.6

Note: The distance sampling column reports the mean and SD of counts from distance sampling surveys, while the counts column shows the mean and SD of the counts from single-visit counts. The difference in magnitude between data types is because transects for the single-visit counts were ~3.4% of the area of the transects for distance sampling.

(Appendix S1: Figure S4), an urban area that has grown in recent decades, with concomitant increases in livestock grazing and tourism (Green et al., 2019).

Distance-sampling data were collected from July 2012 to March 2014. We divided the area surveyed into 18 transects, each ~10 km (10.3 ± 1.0 km) in length; 13 were in the Mara Triangle and five in the Talek region (Appendix S1: Figure S4). Observers drove the transects on three consecutive days every 4–6 weeks; transects in the Mara Triangle were surveyed 16 times, while the Talek transects were surveyed 13 times (with one transect only surveyed on three occasions). Observers began driving transects at sunrise, counted groups of herbivores within 1000 m of the transect, and measured their distance from a vehicle with a laser rangefinder (Nikon Laser 1200). Conspecifics within 200 m of each other were considered a single group. Observers recorded 19,240 groups during 288 surveys (18 transects × 16 repeated visits), ranging from a maximum of 5548 groups (Thomson's gazelle, *Eudorcas thomsonii*) to a minimum of 147 groups (eland, *Taurotragus oryx*) per species. The mean number of animals observed (Table 1) ranged from a minimum of 3.4 (hartebeest, *Alcelaphus buselaphus*) to a maximum of 261 (Thomson's gazelle). In summary, the distance sampling data provide (1) counts of each species and (2) distance measurements to each group of animals (Figure 1, Appendix S1: Figure S5). We binned the distance measurements into 40 distance bands, each 25 m in width.

Count data came from a long-term (since 1988) but less intensive (no distance measurements) monitoring program in the reserve (Appendix S1: Figure S4; Green et al., 2019). We used count data that overlapped with the 2-year timeframe of distance sampling to ensure that species' abundances and detectability were comparable between data sources. Observers surveyed nine transects ~3 km (3.4 ± 1.2 km) in length: six were in the Mara Triangle and three in the Talek region (Appendix S1: Figure S4). Observers drove the transects biweekly for a total of 294 surveys (the number of surveys ranged from 27 to 36 per transect) between sunrise and 10:00 h and counted the total number of herbivores within 100 m of the transect (Boydston et al., 2003; Green et al., 2019). Mean counts across surveys ranged from 0.1 (giraffe, *Giraffa camelopardalis*) to 29.7 (Thomson's gazelle; Table 1). In summary, the count data provide species-specific herbivore counts per transect with no accompanying distance measurements. While the same transects were surveyed multiple times, populations were not closed between repeated surveys (since animals could move during the 2-week periods between surveys), thereby making application of N-mixture models inadvisable (Link et al., 2018; Royle, 2004).

We made several adjustments to the model for the case study (Appendix S1: Figure S5). Since the distance-sampling and count data were collected within the same area by the same observers using a similar methodology and the maximum distance to which animals were surveyed for the counts was known, we used the simplest form of the model in which detection probability is shared from the distance-sampling submodel to the count submodel, which assumes that detection probabilities are equal between the two data types (Appendix S1: Figure S5). Next, anticipating that larger-bodied species would be more detectable over greater distances, we included average species-specific

body mass $MASS_s$ from the PanTHERIA database (Jones et al., 2009) as a covariate in the scale parameter regression within the detection function (Equation 19). In addition, expecting differences in detection probability between regions due to the taller vegetation in the Mara Triangle, we allowed intercepts to vary between regions, indexed with r (1 = Mara Triangle, 2 = Talek):

$$\log(\omega_{sr}) = \gamma_{0sr} + \gamma_1 MASS_s. \quad (19)$$

We drew species-level intercepts from community-level distributions for each of the two regions:

$$\gamma_{0sr} \sim \text{Normal}\left(\mu_{\gamma_0 r}, \sigma_{\gamma_0 r}\right). \quad (20)$$

Since the count data showed evidence of overdispersion (Table 1), we modeled latent abundance with a Poisson–gamma mixture rather than a Poisson distribution. Furthermore, since the same transects were surveyed multiple times, we included an index for visit h in addition to species s and transect i (distance sampling) or t (counts) for the latent abundance. We allowed the overdispersion parameter ρ to take on unique values for each transect–visit combination, thereby accommodating variation in latent abundance between repeated visits (e.g., due to the movement of animals in and out of transects). Thus, for example, we modeled latent abundance at distance sampling transects as

$$N^D_{sih} \sim \text{Poisson}(\rho_{sih}\lambda_{si}), \quad (21)$$

where ρ_{sih} is an overdispersion parameter drawn from species-specific gamma distributions (modeled independently by species, i.e., not drawn from a community-level distribution) with shape and scale hyperparameters ζ_s:

$$\rho_{sih} \sim \text{Gamma}(\zeta_s, \zeta_s). \quad (22)$$

Hyperparameter ζ_s requires a prior distribution; we chose a weakly informative Gamma(4, 2) prior. Because animals were counted within 1000 m of transects for distance sampling but within only 100 m for counts, we calculated separate bin-level detection probabilities for the two data sources because the 25-m bins represented 0.025 and 0.25 of the total width surveyed for distance sampling and count data, respectively (Equation 1). We subsequently calculated a separate transect-level detection probability π^C_s for the count data by summing the bin-level detection probabilities for only the first four bins, rather than all 40 (Equation 4), since animals were counted to only 100 m (i.e., four 25-m-wide bins). Next, we updated the log-linear regression for expected abundance such that the intercept was indexed by region r (1 = Mara Triangle, 2 = Talek). Species' intercepts for the individual regions were drawn from a community-level distribution specific to each region. We also included area offsets A_i since transects were of differing lengths:

$$\log(\lambda_{si}) = \alpha_{0sr[i]} + \log(A_i), \quad (23)$$

$$\alpha_{0sr} \sim \text{Normal}\left(\mu_{\alpha_0 r}, \sigma_{\alpha_0 r}\right). \quad (24)$$

To understand the differences in abundance between the two regions, we computed the differences between intercepts for the two regions as derived variables, for example, $\exp(\alpha_{s2}) - \exp(\alpha_{s1})$. We fit the model with Nimble (de Valpine et al., 2017) through R (R Core Team, 2023) using three MCMC chains run for 900,000 iterations, discarding the initial 300,000 iterations as burn-in and thinning by 200. As in the simulation study, we used standard weakly informative priors for model parameters (McElreath, 2020). To ensure that the model had converged, we inspected parameter traceplots and checked that the convergence diagnostic $\widehat{R}$ was <1.1 (Brooks & Gelman, 1998). We report posterior means and 95% credible intervals (95% CI) for parameters in the case study results below. Data and code are archived on Zenodo in Gilbert (2024).

RESULTS

Simulation

The integrated community model provided unbiased estimates of species-level parameters even when detection probabilities varied between data types (Figure 2). Across species, relative bias was less than 5% for the abundance intercept and slope parameters, the scale parameter intercepts for the distance-sampling and count detection functions, and latent abundance at distance-sampling and count sites (Figure 2, Appendix S1: Table S2). Model estimates of community-level parameters were somewhat less accurate (probably because the simulated communities were relatively small, i.e., 15 species); for example, relative bias for σ_{α_1}, or the SD among species-level abundance slopes, was 15.35%, while absolute bias was 0.04 (Appendix S1: Tables S2 and S3).

The integrated community model tended to provide more accurate and more precise estimates of model parameters compared to alternative single-species and single-data-source models (Figure 3, Appendix S1: Tables S2–S4, Figures S6 and S7). These improvements

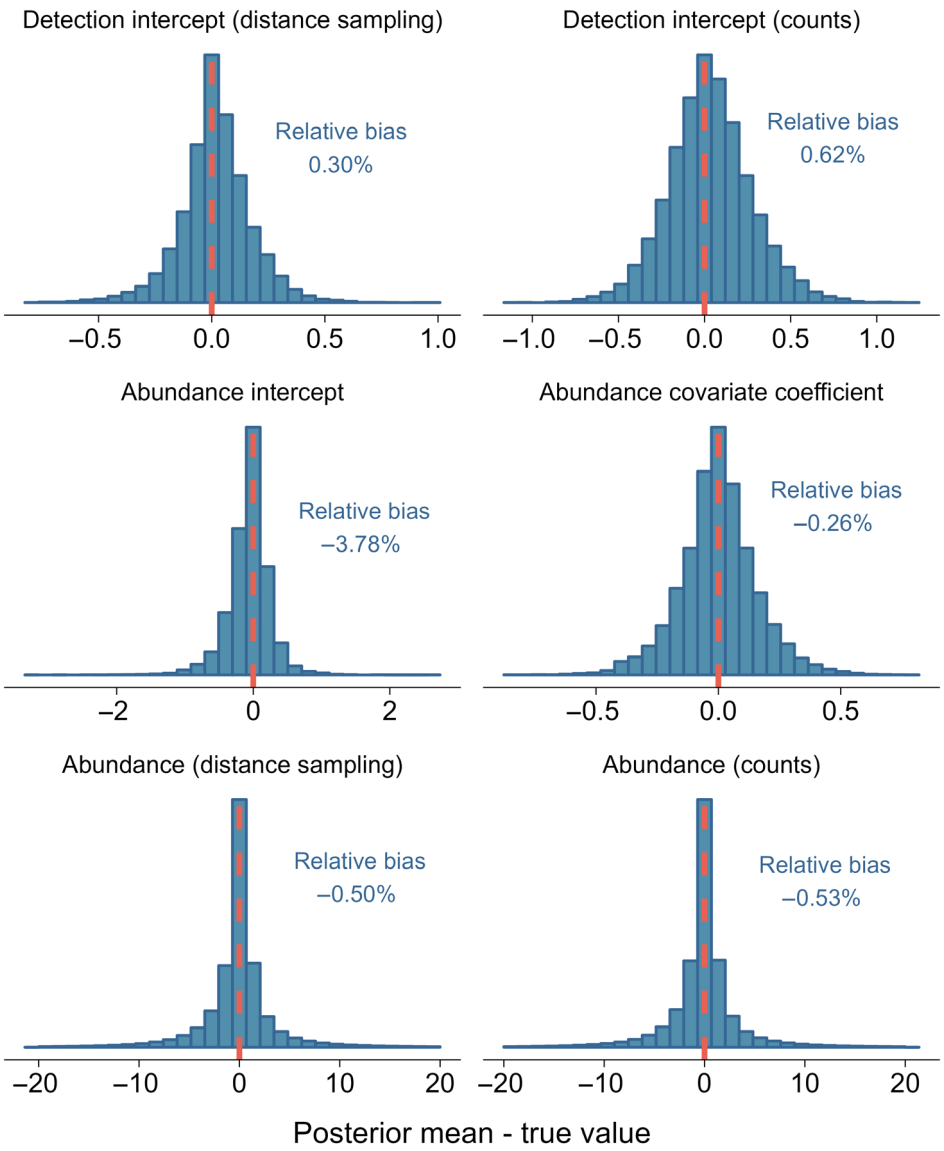

FIGURE 2 Distribution of absolute bias for selected parameters across 1000 replicate simulations for the integrated community model fit to distance sampling data from 50 sites and count data from 100 sites and with lower detection probability for count data on average (Appendix S1: Figure S2). The model provides unbiased estimates of parameters even when detection probability varies between data types; the annotations show relative bias aggregated across replicates and species.

were particularly apparent for rare species (Figure 3, Appendix S1: Figure S7). For example, relative bias for the abundance intercept of rare species was −6.65% for the integrated community model but −36.66% for the integrated single-species model (Appendix S1: Table S2, Figure S6). Similarly, the mean coefficient of variation for the abundance intercept of rare species was 0.43 for the integrated community model but 0.48 for the community distance-sampling model (Appendix S1: Table S4, Figure S7). Models containing only count data performed poorly (Figure 3), in many cases not converging; for example, 100% of the replicate simulations had at least one parameter that did not converge for the community count-only model, and 51.3% of parameters across replicates did not converge (Appendix S1: Table S5). Thus, results for count-only models should be interpreted with caution; for example, while the community count-only model gave the most precise estimates of the abundance intercept for rare species (Appendix S1: Figure S7), these estimates showed extreme negative bias (Figure 3). For reference, 13.9% of the integrated community model replicate simulations had at least one unconverged parameter, while only $1.5e^{-04}$ (i.e., 163 out of 2.318 million) parameters across all simulations did not meet the convergence threshold (Appendix S1: Table S5). All parameters that had convergence problems were latent abundance variables at either distance-sampling (N_{si}^{D}) or count (N_{si}^{C}) transects

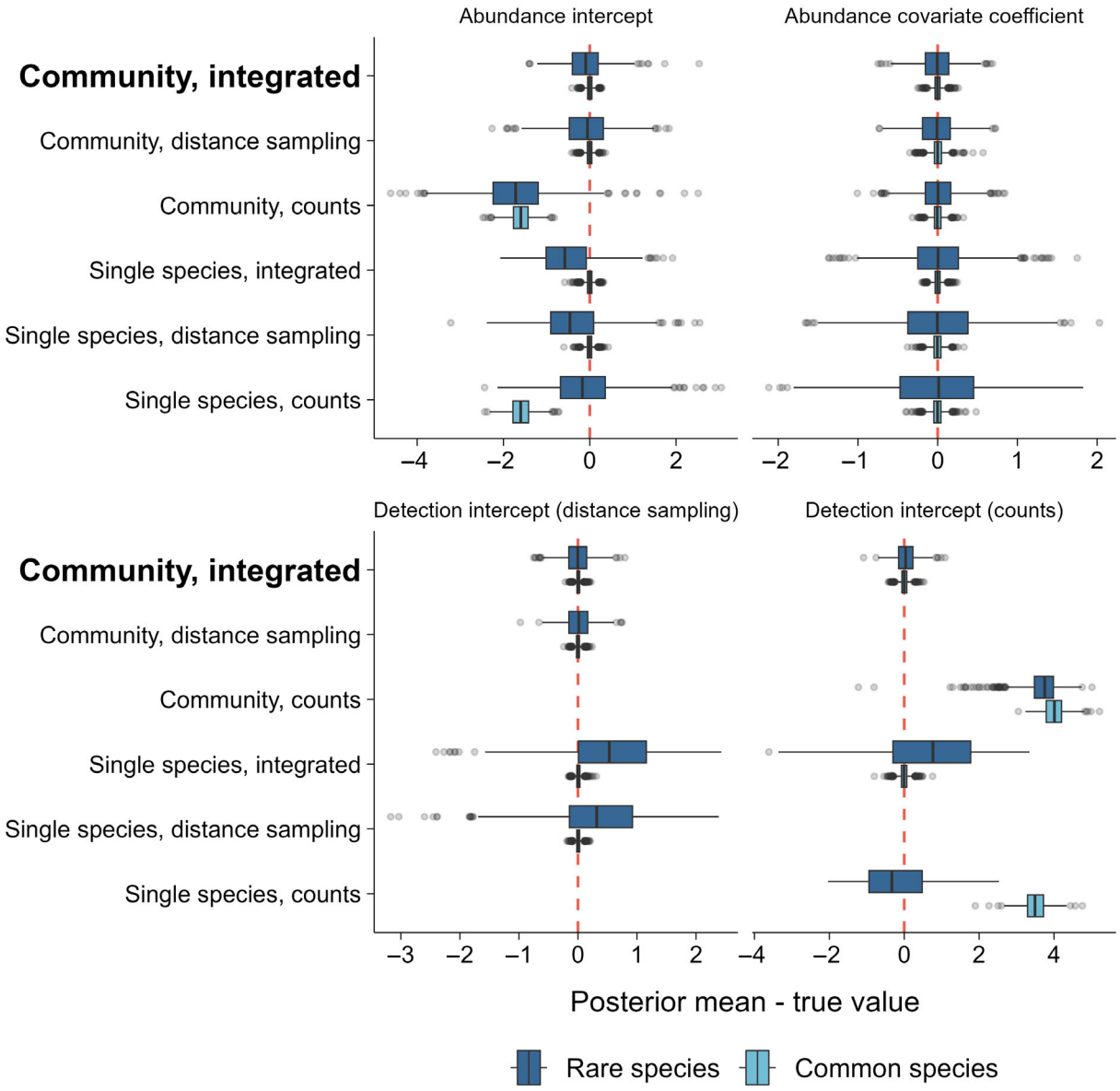

FIGURE 3 Absolute bias of integrated community model relative to five alternative single-species and single-data-source models for rare (dark blue) and common (light blue; most are too narrow to see immediately) species. Boxplots show distributions (over 1000 replicate simulations) of the difference between the posterior mean and the true value (dashed red line indicates a difference of zero, i.e., a perfect estimate). Note that certain parameters are not estimated in some of the alternative models; for example, the models containing only distance-sampling data do not estimate a detection intercept for the count submodel.

and would have likely converged if run for more iterations (maximum $\widehat{R} = 1.22$).

Case study: Herbivores in Masai Mara National Reserve

Most species (eight out of 11) showed abundance differences between the two regions with differing management regimes (Figure 4). Thomson's gazelle, topi (*Damaliscus lunatus*), and Grant's gazelle (*Nanger granti*) showed higher abundances on average in the Talek region, which experiences higher levels of human influence and only passive enforcement of grazing and poaching restrictions (Figure 4). In contrast, impala (*Aepyceros melampus*), buffalo (*Syncerus caffer*), waterbuck (*Kobus ellipsiprymnus*), warthog (*Phacochoerus africanus*), and elephant (*Loxodonta africana*) were more abundant in the Mara Triangle, where there is active conservation enforcement (Figure 4). Average community-level abundance was slightly higher (but not definitively so) in the Mara Triangle (0.20 more animals; 95% CI: 1.43 fewer

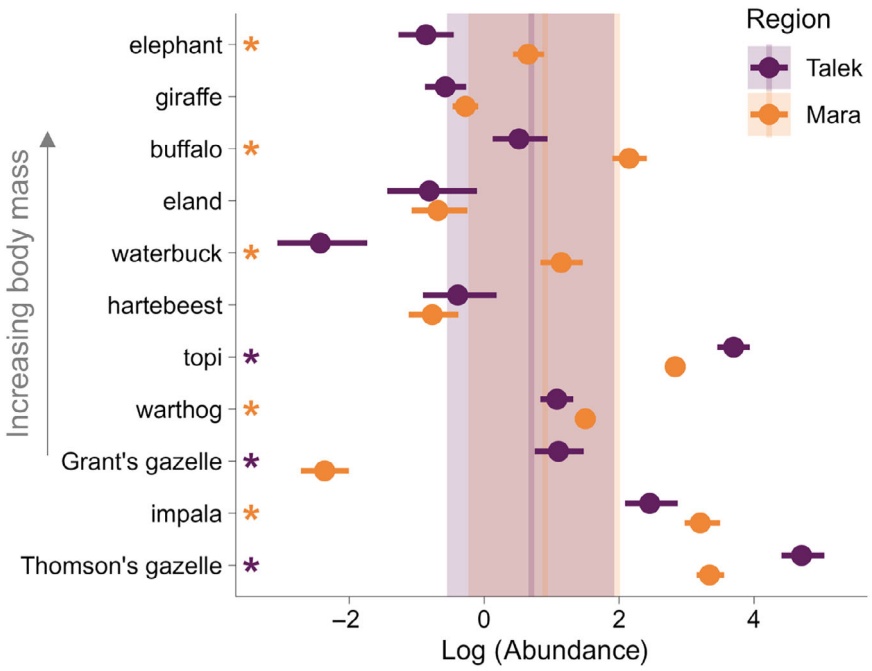

FIGURE 4 Average abundance (log scale) of herbivores within each region. Points and bars show means and 95% credible intervals for species' abundance estimates. The vertical lines and shaded rectangles show the means and 95% credible intervals for the community average of abundance. Species accompanied by an asterisk show between-region differences in abundance with 95% probability; the color of the asterisk denotes the region in which abundance is greater.

animals, 1.86 more animals). Additionally, the community average for abundance was more uncertain for Talek, likely because of the greater variation in abundance among species compared to the Mara Triangle (Figure 4). Beyond these between-region generalizations, most species showed considerable variation in abundance among transects within regions and between surveys on individual transects (Appendix S1: Figures S8 and S9).

Species had higher detectability in the Talek region than in the Mara Triangle; for example, the community average of detection probability at 500 m was 0.19 [95% CI: 0.12, 0.28] in Talek but only 0.07 [95% CI: 0.03, 0.12] in the Mara Triangle. The detectability of larger species decayed more gradually over distance compared to smaller species; the effect of body mass on the detection function scale parameter was estimated to be 0.16 [95% CI: 0.07, 0.24]. For context, this relationship implies that, at 500 m, the smallest species would have detection probabilities of 0.02 [95% CI: 0.00, 0.05] and 0.08 [95% CI: 0.03, 0.15] in Mara and Talek, respectively, while the largest species would have detection probabilities of 0.23 [95% CI: 0.10, 0.37] in Mara and 0.41 [95% CI: 0.26, 0.57] in Talek. Average transect-level detection probability π_s for the distance-sampling data ranged from 0.160 [95% CI: 0.156, 0.165] for Thomson's gazelle in the Mara Triangle to 0.45 [95% CI: 0.36, 0.59] for waterbuck in the Talek region. Transect-level detection probability for the count data was considerably higher (since animals were surveyed to only 100 m); for example, count detection probability was 0.91 [95% CI: 0.90, 0.91] for Thomson's gazelle in the Mara Triangle and 0.99 [95% CI: 0.98, 0.99] for waterbuck in the Talek region.

DISCUSSION

Our integrated community model combines distance-sampling and single-visit count data for multiple species, bringing together the respective benefits of hierarchical community modeling and data integration. Simulations showed that the integrated community model gives unbiased estimates of abundance and detection parameters under the realistic scenario of having larger amounts of count data than distance-sampling data and when the two data types differ in the way they are observed. Simulations also demonstrated that data integration generally increases precision of modeled quantities (by effectively increasing the sample size) and that integrated community models offer greater precision gains than integrated single-species models (Figure 3, Appendix S1: Table S4, Figure S7). Finally, simulations also suggested that the integrated community model can improve the accuracy of parameter estimates for rare species, especially intercepts for abundance and the latent detection function within

the count submodel (Figure 3). In real-world applications, biases may arise because the model shares abundance parameters between submodels for each data source (Figure 1), meaning that the model assumes that abundance patterns are comparable within the areas sampled by the two data sources (or that major differences can be described with covariates).

Under what circumstances is data integration for communities appropriate? Given the assumption of shared abundance parameters between data sources, we recommend applying the model to data sets that sample shared abundance patterns, such as data collected from overlapping or similar regions and time spans (Fink et al., 2010; Pease et al., 2022; Rollinson et al., 2021). While integrating data for overlapping locations and time periods might raise concerns about nonindependence between data sources, previous studies (focusing on integrated population models) have found few ramifications of dependence between data sources (Abadi et al., 2010; Weegman et al., 2021). In the case study, the distance sampling and count transects overlapped (Appendix S1: Figure S4), which provides confidence that abundance patterns sampled by the two data sources were similar. In contrast, it would be inadvisable to integrate distance sampling data (e.g., avian point transects) with counts (e.g., eBird data) from completely disparate regions, when underlying abundance patterns (i.e., parameters within Equations 9 and 18, and their corresponding hyperparameters) have major differences that are not modeled. Similarly, in the case study, we restricted analysis to the 2 years during which distance sampling data were collected. While the count data stretch back to the late 1980s, estimating abundance over many years of sampling might be challenging since average abundances of some species have likely shifted over 30 years with increasing levels of human disturbance in the Talek region (Green et al., 2019). We also used the simplest form of the model, which assumes shared detection processes between data sources; since both data sources were collected by the same field biologists using a similar methodology (driving transects) and surveying to a known distance (out to 100 m), this assumption seems reasonable. In contrast, the separate-detection form of the model is preferable for most data integration scenarios because count and distance-sampling data are likely to be collected differently (e.g., by different observers, or maximum distances to which animals are surveyed for counts might be unknown). A final consideration is how to define the focal community (Pacifici et al., 2014). Defining the community should follow study objectives (Pacifici et al., 2014), particularly with regard to community-level parameters (Figure 1). Grouping species with different life histories may impair interpretation of community-level parameters. For example, in the case study, we omitted hippo (*Hippopotamus amphibius*) due to its preference for aquatic habitats, which are unlike the habitats preferred by the other herbivores considered.

The case study highlights how the integrated community model can be used to estimate abundance underlying both distance-sampling and single-visit count data, the latter of which contain no information about detection probability (Appendix S1: Figures S8 and S9). Variation in herbivore abundance between the two regions with differing management regimes seemed to be associated with the species' sizes (Figure 4). Smaller-bodied taxa like Grant's and Thomson's gazelles were more abundant in the Talek region (higher anthropogenic disturbance), while larger-bodied species like elephant and buffalo were more abundant in the Mara Triangle (less disturbance due to active conservation enforcement). Three (non-mutually-exclusive) mechanisms may have contributed to this pattern. First, grazing produces short, high-productivity forage preferred by small herbivores (especially gazelles) while reducing resources for larger species such as elephants that require large amounts of forage (Estes, 1991). Second, the greater presence of humans in the Talek region may create a "human shield" for smaller herbivores from disturbance-sensitive predators, which prefer to prey upon smaller herbivores rather than large ones (Berger, 2007; Farr et al., 2019; Steyaert et al., 2016). Third, larger herbivores may experience higher levels of persecution by humans due to the destruction of crops or perceived competition for forage (Hoare, 1999; Teixeira et al., 2021). Regardless of the mechanism, our integrated community model permits inference for both species-specific and community-level responses to management regime (Figure 4).

Integrated community modeling is a promising innovation in the evolution of quantitative methods for biodiversity data (Zipkin et al., 2023). This framework combines the respective advantages of integrated and community modeling, allowing researchers to leverage multiple sources of information for multiple species within a streamlined model. The framework we present here integrates distance-sampling and single-visit count data for communities. Distance sampling is a wildlife monitoring staple (e.g., Buckland et al.'s [2001] distance-sampling book has >6000 citations as of 2023) and is applied in many systems, including avian point counts (Sillett et al., 2012), aerial surveys of large mammals (Schmidt et al., 2012), and transect counts of insects (Moranz et al., 2012). Our model can be used for any of these distance-sampling applications, although minor adaptations are required for use with point—rather than line—transects (Kéry & Royle, 2015). Single-visit counts are an even more common type of data and often come from volunteer-science programs like

eBird (Johnston et al., 2021). Counts have the advantage of being relatively easy to collect, potentially over broad spatiotemporal scales or for many species. However, such counts have a disadvantage: They cannot provide inference on abundance since they contain no information on the observation process. Researchers with single-visit count data could collect distance-sampling data with which to "inoculate" the count data to estimate detection probability via our model. If meeting assumptions about shared abundance patterns between data sets is not feasible, users might build model structure (e.g., random effects or covariates) to account for differences in abundance patterns between data sources. In summary, our approach using these two common data sources in biodiversity monitoring will likely soon be joined by additional integrated community models that include other data sources. We believe such models will see application in many biological systems and play an important role in quantifying biodiversity patterns and trends in light of ecology's data revolution.

ACKNOWLEDGMENTS

We thank the many former graduate students and research assistants on the Mara Hyena Project who contributed to data collection. We thank the Kenyan National Commission for Science, Technology and Innovation, the Narok County Government, the Mara Conservancy, and the Kenya Wildlife Service for permission to conduct the fieldwork. NG was supported by a National Science Foundation (NSF) Postdoctoral Research Fellowship in Biology (2208894). DSG was supported by the NSF Graduate Research Fellowship Program and by grants from the College of Natural Sciences and the program in Ecology, Evolution, and Behavior at Michigan State University. The work in this paper was supported by NSF Grants DBI-1954406, DEB-1755089, OIS-1853934, and IOS-1755089. We thank Mason Fidino and Joshua Twining for constructive feedback that improved the manuscript.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data and code (Gilbert, 2024) are available on Zenodo at https://doi.org/10.5281/zenodo.10968910.

ORCID

Neil A. Gilbert https://orcid.org/0000-0003-0949-5612
Matthew T. Farr https://orcid.org/0000-0003-1011-6851
David S. Green https://orcid.org/0000-0001-6031-0076
Elise F. Zipkin https://orcid.org/0000-0003-4155-6139

REFERENCES

Abadi, F., O. Gimenez, R. Arlettaz, and M. Schaub. 2010. "An Assessment of Integrated Population Models: Bias, Accuracy, and Violation of the Assumption of Independence." *Ecology* 91(1): 7–14. https://doi.org/10.1890/08-2235.1.

Barraquand, F., and O. Gimenez. 2019. "Integrating Multiple Data Sources to Fit Matrix Population Models for Interacting Species." *Ecological Modelling* 411: 108713. https://doi.org/10.1016/j.ecolmodel.2019.06.001.

Berger, J. 2007. "Fear, Human Shields and the Redistribution of Prey and Predators in Protected Areas." *Biology Letters* 3(6): 620–23. https://doi.org/10.1098/rsbl.2007.0415.

Borgelt, J., M. Dorber, M. A. Høiberg, and F. Verones. 2022. "More than Half of Data Deficient Species Predicted to be Threatened by Extinction." *Communications Biology* 5(1): 679. https://doi.org/10.1038/s42003-022-03638-9.

Boydston, E. E., K. M. Kapheim, H. E. Watts, M. Szykman, and K. E. Holekamp. 2003. "Altered Behaviour in Spotted Hyenas Associated with Increased Human Activity." *Animal Conservation* 6(3): 207–219. https://doi.org/10.1017/S1367943003003263.

Brooks, S. P., and A. Gelman. 1998. "General Methods for Monitoring Convergence of Iterative Simulations." *Journal of Computational and Graphical Statistics* 7(4): 434–455. https://doi.org/10.1080/10618600.1998.10474787.

Buckland, S. T., D. R. Anderson, K. P. Burnham, J. L. Laake, D. L. Borchers, and L. Thomas. 2001. *Introduction to Distance Sampling: Estimating Abundance of Biological Populations*. Oxford, UK: Oxford University Press.

Cardinale, B. J., J. E. Duffy, A. Gonzalez, D. U. Hooper, C. Perrings, P. Venail, A. Narwani, et al. 2012. "Biodiversity Loss and its Impact on Humanity." *Nature* 486(7401): 59–67. https://doi.org/10.1038/nature11148.

de Valpine, P., D. Turek, C. J. Paciorek, C. Anderson-Bergman, D. T. Lang, and R. Bodik. 2017. "Programming with Models: Writing Statistical Algorithms for General Model Structures with NIMBLE." *Journal of Computational and Graphical Statistics* 26(2): 403–413. https://doi.org/10.1080/10618600.2016.1172487.

Devarajan, K., T. L. Morelli, and S. Tenan. 2020. "Multi-Species Occupancy Models: Review, Roadmap, and Recommendations." *Ecography* 43(11): 1612–24. https://doi.org/10.1111/ecog.04957.

Dirzo, R., H. S. Young, M. Galetti, G. Ceballos, N. J. B. Isaac, and B. Collen. 2014. "Defaunation in the Anthropocene." *Science* 345(6195): 401–6. https://doi.org/10.1126/science.1251817.

Dorazio, R. M., and J. A. Royle. 2005. "Estimating Size and Composition of Biological Communities by Modeling the Occurrence of Species." *Journal of the American Statistical Association* 100(470): 389–398. https://doi.org/10.1198/016214505000000015.

Dorazio, R. M., J. A. Royle, B. Söderström, and A. Glimskär. 2006. "Estimating Species Richness and Accumulation by Modeling Species Occurrence and Detectability." *Ecology* 87(4): 842–854. https://doi.org/10.1890/0012-9658(2006)87[842:ESRAAB]2.0.CO;2.

Doser, J. W., W. Leuenberger, T. S. Sillett, M. T. Hallworth, and E. F. Zipkin. 2022. "Integrated Community Occupancy Models: A Framework to Assess Occurrence and Biodiversity

Dynamics Using Multiple Data Sources." *Methods in Ecology and Evolution* 13(4): 919–932. https://doi.org/10.1111/2041-210X.13811.

Estes, R. 1991. *The Behavior Guide to African Mammals, Including Hoofed Mammals, Carnivores*. Primates: The University of California Press.

Farr, M. T., D. S. Green, K. E. Holekamp, G. J. Roloff, and E. F. Zipkin. 2019. "Multispecies Hierarchical Modeling Reveals Variable Responses of African Carnivores to Management Alternatives." *Ecological Applications* 29(2): e01845. https://doi.org/10.1002/eap.1845.

Farr, M. T., D. S. Green, K. E. Holekamp, and E. F. Zipkin. 2021. "Integrating Distance Sampling and Presence-Only Data to Estimate Species Abundance." *Ecology* 102(1): e03204. https://doi.org/10.1002/ecy.3204.

Fink, D., W. M. Hochachka, B. Zuckerberg, D. W. Winkler, B. Shaby, M. A. Munson, G. Hooker, M. Riedewald, D. Sheldon, and S. Kelling. 2010. "Spatiotemporal Exploratory Models for Broad-Scale Survey Data." *Ecological Applications* 20(8): 2131–47. https://doi.org/10.1890/09-1340.1.

Fletcher, R. J., T. J. Hefley, E. P. Robertson, B. Zuckerberg, R. A. McCleery, and R. M. Dorazio. 2019. "A Practical Guide for Combining Data to Model Species Distributions." *Ecology* 100(6): e02710. https://doi.org/10.1002/ecy.2710.

Gilbert, N. 2024. "n-a-gilbert/multispecies_data_integration (v1.0.0)." Zenodo, Computer Software. https://doi.org/10.5281/zenodo.10968910.

Gilbert, N. A., B. S. Pease, C. M. Anhalt-Depies, J. D. J. Clare, J. L. Stenglein, P. A. Townsend, T. R. Van Deelen, and B. Zuckerberg. 2021. "Integrating Harvest and Camera Trap Data in Species Distribution Models." *Biological Conservation* 258: 109147. https://doi.org/10.1016/j.biocon.2021.109147.

Goyert, H. F., B. Gardner, R. Sollmann, R. R. Veit, A. T. Gilbert, E. E. Connelly, and K. A. Williams. 2016. "Predicting the Offshore Distribution and Abundance of Marine Birds with a Hierarchical Community Distance Sampling Model." *Ecological Applications* 26(6): 1797–1815. https://doi.org/10.1890/15-1955.1.

Green, D. S., L. Johnson-Ulrich, H. E. Couraud, and K. E. Holekamp. 2018. "Anthropogenic Disturbance Induces Opposing Population Trends in Spotted Hyenas and African Lions." *Biodiversity and Conservation* 27(4): 871–889. https://doi.org/10.1007/s10531-017-1469-7.

Green, D. S., E. F. Zipkin, D. C. Incorvaia, and K. E. Holekamp. 2019. "Long-Term Ecological Changes Influence Herbivore Diversity and Abundance inside a Protected Area in the Mara-Serengeti Ecosystem." *Global Ecology and Conservation* 20: e00697. https://doi.org/10.1016/j.gecco.2019.e00697.

Guillera-Arroita, G. 2017. "Modelling of Species Distributions, Range Dynamics and Communities under Imperfect Detection: Advances, Challenges and Opportunities." *Ecography* 40(2): 281–295. https://doi.org/10.1111/ecog.02445.

Harrison, X. A., L. Donaldson, M. E. Correa-Cano, J. Evans, D. N. Fisher, C. E. D. Goodwin, B. S. Robinson, D. J. Hodgson, and R. Inger. 2018. "A Brief Introduction to Mixed Effects Modelling and Multi-Model Inference in Ecology." *PeerJ* 6: e4794. https://doi.org/10.7717/peerj.4794.

Hoare, R. E. 1999. "Determinants of Human–Elephant Conflict in a Land-Use Mosaic." *Journal of Applied Ecology* 36(5): 689–700. https://doi.org/10.1046/j.1365-2664.1999.00437.x.

Isaac, N. J. B., M. A. Jarzyna, P. Keil, L. I. Dambly, P. H. Boersch-Supan, E. Browning, S. N. Freeman, et al. 2020. "Data Integration for Large-Scale Models of Species Distributions." *Trends in Ecology & Evolution* 35(1): 56–67. https://doi.org/10.1016/j.tree.2019.08.006.

IUCN. 2022. "The IUCN Red List of Threatened Species. Version 2022-2." https://www.iucnredlist.org.

Johnston, A., W. M. Hochachka, M. E. Strimas-Mackey, V. Ruiz Gutierrez, O. J. Robinson, E. T. Miller, T. Auer, S. T. Kelling, and D. Fink. 2021. "Analytical Guidelines to Increase the Value of Community Science Data: An Example Using eBird Data to Estimate Species Distributions." *Diversity and Distributions* 27(7): 1265–77. https://doi.org/10.1111/ddi.13271.

Johnston, A., E. Matechou, and E. B. Dennis. 2022. "Outstanding Challenges and Future Directions for Biodiversity Monitoring Using Citizen Science Data." *Methods in Ecology and Evolution* 14: 103–116. https://doi.org/10.1111/2041-210X.13834.

Jones, K. E., J. Bielby, M. Cardillo, S. A. Fritz, J. O'Dell, C. D. L. Orme, K. Safi, et al. 2009. "PanTHERIA: A Species-Level Database of Life History, Ecology, and Geography of Extant and Recently Extinct Mammals." *Ecology* 90(9): 2648. https://doi.org/10.1890/08-1494.1.

Kéry, M., and J. A. Royle. 2015. *Applied Hierarchical Modeling in Ecology*, 1st ed., Vol. 1. London, UK: Elsevier. https://doi.org/10.1016/C2013-0-19160-X.

Kéry, M., J. A. Royle, T. Hallman, W. D. Robinson, N. Strebel, and K. F. Kellner. 2022. "Integrated Distance Sampling Models for Simple Point Counts (arXiv:2211.17229)." arXiv. https://doi.org/10.48550/arXiv.2211.17229.

Link, W. A., M. R. Schofield, R. J. Barker, and J. R. Sauer. 2018. "On the Robustness of N-Mixture Models." *Ecology* 99(7): 1547–51. https://doi.org/10.1002/ecy.2362.

McElreath, R. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and STAN*, 2nd ed. Boca Raton, FL: CRC Press.

Moranz, R. A., D. M. Debinski, D. A. McGranahan, D. M. Engle, and J. R. Miller. 2012. "Untangling the Effects of Fire, Grazing, and Land-Use Legacies on Grassland Butterfly Communities." *Biodiversity and Conservation* 21(11): 2719–46. https://doi.org/10.1007/s10531-012-0330-2.

Pacifici, K., B. J. Reich, D. A. W. Miller, and B. S. Pease. 2019. "Resolving Misaligned Spatial Data with Integrated Species Distribution Models." *Ecology* 100(6): e02709. https://doi.org/10.1002/ecy.2709.

Pacifici, K., E. F. Zipkin, J. A. Collazo, J. I. Irizarry, and A. DeWan. 2014. "Guidelines for a Priori Grouping of Species in Hierarchical Community Models." *Ecology and Evolution* 4(7): 877–888. https://doi.org/10.1002/ece3.976.

Pease, B. S., K. Pacifici, and R. Kays. 2022. "Exploring Spatial Nonstationarity for Four Mammal Species Reveals Regional Variation in Environmental Relationships." *Ecosphere* 13(8): e4166. https://doi.org/10.1002/ecs2.4166.

Péron, G., and D. N. Koons. 2012. "Integrated Modeling of Communities: Parasitism, Competition, and Demographic Synchrony in Sympatric Ducks." *Ecology* 93(11): 2456–64. https://doi.org/10.1890/11-1881.1.

Quéroué, M., C. Barbraud, F. Barraquand, D. Turek, K. Delord, N. Pacoureau, and O. Gimenez. 2021. "Multispecies Integrated Population Model Reveals Bottom-up Dynamics in a Seabird

Predator–Prey System." *Ecological Monographs* 91(3): e01459. https://doi.org/10.1002/ecm.1459.

R Core Team. 2023. "R: A Language and Environment for Statistical Computing [Computer Software]." R Foundation for Statistical Computing. https://www.R-project.org/.

Rollinson, C. R., A. O. Finley, M. R. Alexander, S. Banerjee, K.-A. Dixon Hamil, L. E. Koenig, D. H. Locke, et al. 2021. "Working across Space and Time: Nonstationarity in Ecological Research and Application." *Frontiers in Ecology and the Environment* 19(1): 66–72. https://doi.org/10.1002/fee.2298.

Royle, J. A. 2004. "N-Mixture Models for Estimating Population Size from Spatially Replicated Counts." *Biometrics* 60(1): 108–115. https://doi.org/10.1111/j.0006-341X.2004.00142.x.

Schaub, M., and M. Kéry. 2021. *Integrated Population Models: Theory and Ecological Applications with R and JAGS*. London, UK: Academic Press.

Schmidt, J. H., K. L. Rattenbury, J. P. Lawler, and M. C. Maccluskie. 2012. "Using Distance Sampling and Hierarchical Models to Improve Estimates of Dall's Sheep Abundance." *The Journal of Wildlife Management* 76(2): 317–327. https://doi.org/10.1002/jwmg.216.

Schmidt, J. H., and H. L. Robison. 2020. "Using Distance Sampling-Based Integrated Population Models to Identify Key Demographic Parameters." *The Journal of Wildlife Management* 84(2): 372–381. https://doi.org/10.1002/jwmg.21805.

Schmidt, J. H., T. L. Wilson, W. L. Thompson, and B. A. Mangipane. 2022. "Integrating Distance Sampling Survey Data with Population Indices to Separate Trends in Abundance and Temporary Immigration." *The Journal of Wildlife Management* 86(3): e22185. https://doi.org/10.1002/jwmg.22185.

Sillett, T. S., R. B. Chandler, J. A. Royle, M. Kéry, and S. A. Morrison. 2012. "Hierarchical Distance-Sampling Models to Estimate Population Size and Habitat-Specific Abundance of an Island Endemic." *Ecological Applications* 22(7): 1997–2006. https://doi.org/10.1890/11-1400.1.

Simmonds, E. G., S. G. Jarvis, P. A. Henrys, N. J. B. Isaac, and R. B. O'Hara. 2020. "Is More Data Always Better? A Simulation Study of Benefits and Limitations of Integrated Distribution Models." *Ecography* 43(10): 1413–22. https://doi.org/10.1111/ecog.05146.

Sollmann, R., B. Gardner, K. A. Williams, A. T. Gilbert, and R. R. Veit. 2016. "A Hierarchical Distance Sampling Model to Estimate Abundance and Covariate Associations of Species and Communities." *Methods in Ecology and Evolution* 7(5): 529–537. https://doi.org/10.1111/2041-210X.12518.

Steyaert, S. M. J. G., M. Leclerc, F. Pelletier, J. Kindberg, S. Brunberg, J. E. Swenson, and A. Zedrosser. 2016. "Human Shields Mediate Sexual Conflict in a Top Predator." *Proceedings of the Royal Society B: Biological Sciences* 283(1833): 20160906. https://doi.org/10.1098/rspb.2016.0906.

Teixeira, L., K. C. Tisovec-Dufner, G. d. L. Marin, S. Marchini, I. Dorresteijn, and R. Pardini. 2021. "Linking Human and Ecological Components to Understand Human–Wildlife Conflicts across Landscapes and Species." *Conservation Biology* 35(1): 285–296. https://doi.org/10.1111/cobi.13537.

Walpole, M. J., and N. Leader-Williams. 2001. "Masai Mara Tourism Reveals Partnership Benefits." *Nature* 413(6858): 771. https://doi.org/10.1038/35101762.

Weegman, M. D., T. W. Arnold, R. G. Clark, and M. Schaub. 2021. "Partial and Complete Dependency among Data Sets Has Minimal Consequence on Estimates from Integrated Population Models." *Ecological Applications* 31(3): e2258. https://doi.org/10.1002/eap.2258.

Yamaura, Y., J. A. Royle, N. Shimada, S. Asanuma, T. Sato, H. Taki, and S. Makino. 2012. "Biodiversity of Man-Made Open Habitats in an Underused Country: A Class of Multispecies Abundance Models for Count Data." *Biodiversity and Conservation* 21(6): 1365–80. https://doi.org/10.1007/s10531-012-0244-z.

Zhao, Q., K. Heath-Acre, D. Collins, W. Conway, and M. D. Weegman. 2021. "Integrated Population Modelling Reveals Potential Drivers of Demography from Partially Aligned Data: A Case Study of Snowy Plover Declines under Human Stressors." *PeerJ* 9: e12475. https://doi.org/10.7717/peerj.12475.

Zipkin, E. F., A. DeWan, and J. Andrew Royle. 2009. "Impacts of Forest Fragmentation on Species Richness: A Hierarchical Approach to Community Modelling." *Journal of Applied Ecology* 46(4): 815–822. https://doi.org/10.1111/j.1365-2664.2009.01664.x.

Zipkin, E. F., J. W. Doser, C. L. Davis, W. Leuenberger, S. Ayebare, and K. L. Davis. 2023. "Integrated Community Models: A Framework Combining Multispecies Data Sources to Estimate the Status, Trends and Dynamics of Biodiversity." *Journal of Animal Ecology* 92: 2248–62. https://doi.org/10.1111/1365-2656.14012.

Zipkin, E. F., B. D. Inouye, and S. R. Beissinger. 2019. "Innovations in Data Integration for Modeling Populations." *Ecology* 100(6): e02713. https://doi.org/10.1002/ecy.2713.

Zipkin, E. F., and S. P. Saunders. 2018. "Synthesizing Multiple Data Types for Biological Conservation Using Integrated Population Models." *Biological Conservation* 217: 240–250. https://doi.org/10.1016/j.biocon.2017.10.017.

SUPPORTING INFORMATION

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

How to cite this article: Gilbert, Neil A., Caroline M. Blommel, Matthew T. Farr, David S. Green, Kay E. Holekamp, and Elise F. Zipkin. 2024. "A Multispecies Hierarchical Model to Integrate Count and Distance-Sampling Data." *Ecology* 105(7): e4326. https://doi.org/10.1002/ecy.4326