ABSTRACT

PEASE, BRENT STEVEN. Exploring Drivers of Spatial Non-Stationarity in Wildlife-Environment Relationships Through Species Distribution Modeling. (Under the direction of Drs. Krishna Pacifici and Roland Kays).

Understanding the spatial and temporal variation in the distribution and abundance of a species is a central interest in ecology and is vital information for wildlife management agencies to implement well-informed, successful management actions. Species distribution models (SDMs) are a widely used tool to understand environmental characteristics contributing to the presence or absence of a given species and predict their distribution. Despite the wide-reaching applications, most research focused on species distributions ignores fine-scale delineations (e.g., subspecies or populations) in their analyses, presuming that the ecological relationships discovered are uniform (i.e., stationary) over space. However, ecological relationships may vary spatially because of changes in the environment (i.e., land cover conversion) or the organism itself (i.e., ecotype or subspecies variation). Discovering this within-species variation requires dense datasets over large areas, which are increasingly available due to the recent proliferation of large-scale animal occurrence data, from citizen scientists and other efforts. Using nearly 4000 sampling locations from a state-wide camera trapping project, in Chapter 1 I present results from a spatially varying coefficients (SVC) model that allows for spatial non-stationarity in environmental relationships to explore space-varying effects of covariates on the distribution of four mammal species predicted by eight measures of the environment. The SVC model had superior fit and predictive performance in comparison with regional and stationary models for bobcat (*Lynx rufus*) and white-tailed deer (*Odocoileus virginianus*). Species that had inherent relationships with specific environmental characteristics (e.g., forest cover) tended to have the greatest support for spatial non-stationarity, whereas widespread, generalists tended to have coarser (e.g., physiographic

regions) variability in ecological relationships. Spatial non-stationarity was more common in relationships with landscape composition characteristics, such as housing density, than in landscape configuration metrics, such as patch richness density. Combining different datasets has the potential to increase the ability to document nonstationarity but remains statistically challenging. In Chapter 2, I developed a procedure for selecting among multiple data sources for use in integrated species distribution models using four publicly available data sources maintained by the North Carolina state wildlife agency. Using a structured decision making framework, we found that integrating multiple data sources reduced parameter uncertainty and improved predictive abilities in the SDMs. Single data source models underestimated ecological effects compared to integrated models, and the estimated distribution of the integrated models better reflected the known distribution of a suite of species across the state. Finally, in chapter 3, using a local spatial modeling approach, I modeled American black bear (*Ursus americanus*) distribution across North America, finding notable spatial non-stationarity in responses to land use, human pressure, and climatic conditions. The genetic affinity of subspecies boundaries explained the responses to land-use characteristics whereas ecoregion delineations explained climatic responses, suggesting local adaptation to climate. Together, this dissertation highlights the importance of accounting for varied ecological relationships over large geographic space and incorporating multiple data sources in SDMs to inform wildlife conservation and management.

Exploring Drivers of Spatial Non-Stationarity in Wildlife-Environment Relationships Through Species Distribution Modeling

by
Brent Steven Pease

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Fisheries, Wildlife, and Conservation Biology

Raleigh, North Carolina
2021

APPROVED BY:

Dr. Krishna Pacifici
Committee Co-Chair

Dr. Roland Kays
Committee Co-Chair

Dr. Brian Reich

Dr. Christopher Moorman

DEDICATION

To my grandparents.

BIOGRAPHY

Brent was born in Centralia, Illinois, a small town of 14,000 about an hour east of St. Louis, Missouri. Although much of his youth was focused on playing organized sports, a summer spent in Estes Park, Colorado following his high school graduation changed Brent's career trajectory almost overnight. Graduating in 2014 from Colorado State University with a BSc in Fish, Wildlife, and Conservation Biology, he began down a career-track focused on the conservation and management of wildlife. After nearly a year spent with Colorado Parks and Wildlife, Brent returned home to begin a MSc program in Forestry at Southern Illinois University where he graduated in 2017 with honors. The culmination of these experiences led him to pursue his PhD at North Carolina State University in Fisheries, Wildlife, and Conservation Biology to realize his professional goal of becoming a university professor and researcher. Starting Summer 2021, Brent will be living in Southern Illinois with his wife, Reanna, and their child, Arlo, as Brent transitions to his new position as an Assistant Professor in the Forestry Program at Southern Illinois University.

ACKNOWLEDGMENTS

I am grateful for the many individuals that have supported me in some way during the completion of this dissertation. I am especially thankful for my advisors, Krishna Pacifici and Roland Kays, who spent hours in support of my development as a scientist. Despite the many personal challenges he faced during my graduate training, Krishna always made time to discuss ideas, offer professional advice, and generally chat about anything (particularly if it involved Google Maps). Krishna truly instilled the process of science in his mentoring and worked to prepare me for my career every time we interacted. Thanks to Roland for always pushing ahead to the next "big thing"; his drive to continually learn and do more was inspiring. My committee members, Brian Reich and Christopher Moorman, were a great asset throughout my training. Brian could not have been more supportive or patient when working through an analysis problem. Thanks to Chris for taking the time to make sure I knew what "habitat" represented and for ensuring that none of my research was lost (or found); the many hours he spent editing my dissertation greatly improved the flow and readability and taught me many lessons along the way. Though not directly involved with my dissertation research, special thanks to other faculty who helped along the way including Jaime Collazo, Josh Gray, Kevin Gross, Erin Sills, Lara Pacifici, George Hess, and Nils Peterson.

While there were many highlights during my time at NC State, one of the greatest was the opportunity to interact with lab members and other graduate students. I consider time spent interacting with and learning from Arielle Parsons as second-to-none; Arielle is an incredible scientist and I am grateful to have had the chance to work with her. My graduate experience would not have been the same without Mike Cove – thanks to him for his friendship and mentorship. Losing intramural basketball games with Andy Maurer was an excellent use of my time – without Covid, by the end of the season we may have been walk-ons with the Wolfpack. Although the list is long, I have several in mind who helped along the way, including Nick Gould, Will Casola, Eric Teitsworth, Hailey Boone, Matt Snider, Riley Gallagher, and Dan Hannon.

I could not imagine completing this dissertation without the support of my wife, Reanna. Thank you for always compromising, listening, and supporting – so excited about the next stage of life with our family. Finally, I am continually grateful for the consistent interest and support from my parents and brother – it has been quite the winding path but I think we finally made it.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Understanding drivers of variation in animal distribution and abundance is fundamental to ecology (Odum and Barrett 1971, Krebs 1994). Animal distribution focuses on *where* a species occurs whereas abundance describes the *number* of individuals in some given area. A species' fundamental niche governs these, describing environment conditions in which given species can exist in the *absence* of biotic interactions (Hutchinson 1957). The realized niche, rather, is a subset of the fundamental niche that explicitly takes into consideration how biological interactions limit where a species occurs (Hutchinson 1957).

Species distribution models (SDMs) are a widely used tool to understand associations between the occurrence of a given species and the abiotic and biotic environment – a modeling framework for understanding ecological niches. Despite the wide-reaching applications, most broad-scale ecological research focused on species distributions ignores fine-scale delineations (e.g., subspecies or populations) in their analyses, presuming that the ecological relationships discovered are uniform (i.e., stationary) over space. However, ecological relationships may vary spatially because of changes in the environment (i.e., land cover conversion) or the organism itself (i.e., ecotype or subspecies variation). Discovering this within-species variation requires dense datasets over larger areas, which is now being provided by the recent proliferation of large-scale animal occurrence data, from citizen scientists and other efforts.

The focus of this dissertation is to explore how species-environment associations vary over geographic space and to attempt to elucidate the drivers of this spatial variation. Additionally, some of the work within works to identify ways to take advantage of the vast amount of biodiversity observations through data integration. In Chapter 2, we first test for spatial non-stationarity at the state-level using a state-wide volunteer camera trapping project.

We identify which species-environment relationships vary over space and attempt to uncover drivers of this variation. Chapter 3 stays within the boundaries of North Carolina to develop a data-checking procedure for integrated species distribution models. Finally, Chapter 4 returns to exploring spatial non-stationarity but across a species' entire range, allowing for the inclusion of genetic differences as a potential driver of varied species-environment relationships. Concluding remarks are discussed in Chapter 5.

CHAPTER 2: AT WHAT SCALE DO WILDLIFE-HABITAT RELATIONSHIPS VARY ACROSS SPACE?

Introduction

There is a growing interest and need to predict animal distribution and abundance across large geographic scales and time periods (Schimel 2011). The move towards estimating regional to range-wide trends is driven by the rapid pace of anthropogenic-induced global change (e.g., land conversion, climate change) and its impact on biodiversity (Lister and Stuart 2008, Moritz et al. 2008). Fortunately, the accumulation of long-term biodiversity monitoring programs, new digital sensors, and growing citizen science efforts is providing spatially dense animal occurrence data at the continental or global scale (Kays et al. 2020a). Coupled with a movement towards open-access data in ecology (Reichman et al. 2011), the wealth of animal occurrence records has prompted a notable increase in the use of species distribution models (SDMs) to relate animal observations to environmental conditions and predict how animals will respond to future changes. Although range wide SDMs have been generated using museum records for decades (Newbold 2010), the current quantity of observations of animal occurrences provides greater spatiotemporal detail to evaluate assumptions that were taken for granted with past traditional data sources.

Although SDMs can take many forms, a regression-based (e.g., Generalized Linear Models [GLMs]) statistical framework is commonly used to describe species-environment relationships in SDMs (Franklin 2010, Miller 2010). In this approach, accounting for variation among the observations is dealt primarily through the use of fixed-effects coefficients (i.e., covariates), where species-environment relationships are described by the slope of a given coefficient, and model fit may be improved with additional covariates. An important but often overlooked statistical assumption in this framework is that the estimated coefficients are assumed to be constant or "stationary" across the species' range. However, as the spatial extent of a modeling exercise increases, the assumption of fixed parameter estimates becomes increasingly unrealistic. That is, a constant set of regression coefficients may not represent the actual space-varying relationships between a response (e.g., animal abundance or occupancy) and environmental covariates (e.g., temperature) (Figure 1).

Non-stationarity in ecological processes may be a result of external (i.e., environmental) and/or internal (i.e., focal species) factors. External factors include functional responses to variation in resource availability, predator-prey dynamics, or interspecific competition, whereas internal factors include intraspecific variation and local adaptations. Non-stationarity as a result of varied resource availability could arise when a given environmental feature is completely absent in some parts of the study area but present in other areas, potentially resulting in a varied statistical response because the environmental feature can't be continually used by all populations in the study area. The relative abundance of an environmental feature could also create non-stationarity. For example, consider the common occurrence of some environmental feature being available only in isolated pockets in some regions but occurring in large, contiguous blocks in others; the statistical importance of this feature could be quite different between regions, especially where a species is more strongly drawn to the resource when its availability is restricted (Elton 1946, Chase and Leibold 2003). Other external factors include predator-prey interactions, where the importance of these interactions vary with space as community composition, relative abundances, and environmental conditions (e.g., connectivity, habitat-mediated predator-prey dynamics, etc.) differ across large space, ultimately resulting in varied behavior of the focal species that results in non-stationarity (Berger et al. 2001, Roemer et al. 2009, Gorini et al. 2012). Interspecific competition is induced by competing species and the presence/absence of those competitors across large space (and/or a difference suite of competitors across large space); this may result in varied species-environment relationships because of behavioral differences induced by competitors and/or predators (Schoener 1986, Jones and Barmuta 2000). Internal factors leading to non-stationarity may be a function of phylogenetic differences across a species' range and would indicate a lack of niche conservatism, reflecting recent local adaptation (Wiens and Graham 2005, Peterson 2011, Ourers et al. 2020).

Spatial non-stationarity is likely common in ecological systems (Bini et al. 2009) but has received little attention in wildlife research (Miller 2012a), in part because it requires more data than traditional approaches. Several methods to account for non-stationarity in SDM's have been developed including eigenvector spatial filtering (Griffith 1996, 2008), Generalized Additive Models (Wood 2017), Geographic Iighted Regression (GWR) (Fotheringham et al. 2003), and Bayesian Spatially-Varying Coefficients (SVC) Models (Gelfand et al. 2003, Banerjee et al. 2014). Bayesian SVC models are appealing because they are valid probabilistic models that permit full Bayesian inference (Gelfand et al. 2003, Finley 2011, Banerjee et al. 2014).

While previous research has warranted further exploration of space-varying effects (Shi et al. 2006, Fortin et al. 2008, Smith et al. 2019b), and despite extensive use of GWR (Fotheringham et al. 2003), a number of important ecological aspects are unresolved. Namely, it is unclear which species tend to respond differently across space, which type of ecological predictors the species are responding to differently, and whether the magnitude (i.e., strength) and resolution (i.e., fine vs large scale) of the space-varying effects relate to the ecology of the species so that some general trends arise. Our primary objective was to test for evidence of non-stationarity in effects of landscape composition and configuration (e.g., patch size) on medium-to-large sized mammals across the state of North Carolina, USA.

Based on previous research, we expected that non-stationarity would indeed exist, but it is unclear whether the non-constant relationships would be best delineated by large-scale characteristics such as physiographic regions (i.e., ecoregions) or better described by fine-scale responses (e.g., population level). We expected that the spatially-varying effects would differ among species and that they would reflect the ecology of each species. For highly adaptable, generalist species such as white-tailed deer (*Odocoileus virginianus*), northern raccoon (*Procyon lotor*) or coyote (*Canis latrans*), we predicted that non-stationarity would be detected and the spatially varying relationships would be fine-scale, reflecting localized environmental conditions and the species' ability to use and exploit a wide variety of conditions. In contrast, we predicted that species with a restricted niche breadth or those known to avoid specific environmental conditions (e.g., human development), such as bobcat (*Lynx rufus*), will have wide-reaching, consistent responses to some environmental conditions.

Materials and Methods

Study Area

The study area was the state of North Carolina, USA (Figure 1). North Carolina has an estimated human population of 10.3 million, covers an area of 139,390 km^2, lies between 33°50′ N to 36°35′ N and 75°28′ W to 84°19′ W, and ranges between 0 – 2037 m in elevation. Land cover across the state is approximately 40% forest (deciduous, coniferous, and mixed), 20% row crop agriculture, 10% developed, and 10% water, with the remaining 20% being comprised of wetlands, scrubland, or grasslands (Homer et al. 2015). The state consists of three primary physiographic regions: the Coastal Plain, Piedmont, and Appalachian Mountains (Halls 1984). The Coastal Plain comprises about 45% of the state and is characterized primarily by agricultural land cover, wetlands, and mixed coniferous forests. The Piedmont lies between the Coastal Plain and Mountains and topographically consists of gentle rolling hills, gaining in elevation towards the mountains. The Piedmont contains the primary urban centers of the North Carolina and is mosaic of developed, forested, and row crop agriculture land covers. Containing the state's most rugged topography, the Appalachian Mountain region is predominately forested but makes up just 10% of the state.

Data Source - Citizen Science Camera Trapping

We analyzed camera trap data from the North Carolina Candid Critters (NCCC) project (Lasky et al. In Review) that occurred in North Carolina during Oct 2015- Dec 2019 (Figure 1). The Candid Critters project relied on volunteer citizen scientists to deploy cameras that followed a standardized protocol: camera sites were unbaited, cameras were mounted 40-50 cm above ground, active 24 hours/day, and had a target deployment length of 21 days. All cameras had trigger speeds < 0.5 seconds and were equipped with infrared flash. Cameras were set to maximum trigger sensitivity, to have no rest period between triggers, and to record multiple photographs per trigger. Triggers within 60 seconds of each other were considered one detection and grouped into a sequence, and then we used these sequences as independent records for subsequent analysis of detection rate (sequences per day). Cameras continued to re-trigger if the animal was still in the detection zone of the camera. Camera locations were either at the volunteer's discretion or using pre-selected locations determined by stratified random sampling of land cover types in North Carolina. Previous analyses determined that the camera locations were representative of the state's land cover (Kays et al. In Review). Citizen science volunteers used the eMammal software to upload pictures and identify all photographed wildlife species. Uploaded pictures were then vetted by project staff using the eMammal Expert Review Tool, where species identification was confirmed or corrected (McShea et al. 2016).

We removed deployments that lasted < 7 days or > 35 days, as this indicated camera malfunction or deployments that did not adhere to protocol. We also spatially thinned deployments from the same month of a given year or that were separated by ≤50 m, resulting in 3,678 camera trap deployments. Cameras were deployed on both public (~40%) and private lands (~60%), and deployments were set in all 100 North Carolina counties.

We imposed a 5-km x 5-km regular hexagonal grid across North Carolina (resulting in 5,497 grid cells) for all analyses. We chose this resolution because it is representative of environmental conditions encountered by the focal wildlife species during daily movement and within seasonal home ranges. We assigned all camera deployments to their respective grid cell and summarized all environmental covariates within each grid cell. For simplicity, we summed the number of species-specific daily detections and recorded the cumulative number of days all cameras were deployed in a grid cell.

Covariates

We focused on covariates describing landscape composition and configuration known to influence mammal habitat use in North Carolina (Kays et al. 2017, Parsons et al. 2018) and beyond (Kie et al. 2002) (Table 1). Covariates describing land cover composition and configuration included forest cover, housing density, patch richness density, and contagion index (herein, focal covariates) and were derived from the 2011 30-m x 30-m National Land Cover

Data (NLCD) product (Homer et al. 2015)) and the North Carolina housing density estimates which were based on the 2010 TIGER census blocks (Bureau 2010). Contagion index and patch richness density were calculated using the R package *landscapemetrics* (Hesselbarth et al. 2019). We considered several other landscape metrics to represent configuration but determined that PRD and contagion captured many characteristics known to affect mammal distribution and abundance (Kie et al. 2002, Nielsen and Woolf 2002, Saïd and Servanty 2005, Constible et al. 2006, Walter et al. 2009, Cherry et al. 2017). Specifically, the two chosen metrics had correlation coefficients greater than or equal to 0.5 (positive or negative) with edge density, mean patch area, and Shannon's Diversity Index, suggesting that these two measures were capturing several important determinants of mammal distribution and abundance (Kie et al. 2002). The landscape composition and configuration covariates were summarized in 5-km x 5-km hexagonal grid cells across North Carolina; this resolution was based on estimated home ranges and scale of effects for each species (Kie et al. 2002, Woolf et al. 2002, Constible et al. 2006, Clare et al. 2015). Four additional site-level covariates were used to describe camera-level variation (Table 1); we used the Microsoft US Building Footprints for North Carolina to calculate camera distance to building structures, distances to roads used the North Carolina Department of Transportation's linear features database, and water source distance calculations used the Natural Earth waterways database. We also included detection distance – a field-based measurement of the distance from the mounted camera a human can be detected – determined by walk tests performed by the volunteer deploying the camera. Lastly, we used the length of the camera deployment to represent the amount of camera trapping effort. All covariates included in a model were checked to have correlation coefficients less than 0.6 and were mean-centered prior to analysis to aid computation and interpretation of parameter estimates.

Model Descriptions

We used the same animal and environmental data to build SDM's with three different modeling approaches to compare their performance, and thus quantify the importance of accounting for non-stationarity. Our starting point, representing a common approach in ecology, was a generalized linear regression specification (herein, "GLM model"), which included all covariates as fixed effects. The GLM model assumed that a given covariate effect is constant across space (i.e., assumes stationarity), reflecting the hypothesis that a species is responding consistently across the area of interest. We added a spatial random effect to the GLM model (and subsequent models) to account for autocorrelation and other remaining patterns of spatially structured residuals (see below for description of the spatial random effect). In the other two models, we allowed the effects of the landscape composition and configuration covariates to vary over space (Table 1). In the second model we used the three primary ecoregions ("ecoregion model") of North Carolina as boundaries where relationships were constant within the ecoregions but varied among the ecoregions, estimating effects of landscape composition and configuration metrics for each region (i.e., a total of 12 covariates – 4 covariates in 3 regions). The Ecoregion model reflect the hypothesis that non-stationarity is occurring but at coarse natural boundaries (i.e., coarse resolution). The third approach used a spatially varying coefficients model (Gelfand et al. 2003), which allowed a given parameter effect (e.g., forest cover) to vary continuously over the region of interest ("SVC model"); unlike the ecoregion model, this model does not impose boundaries where a given effect is thought to differ, but rather directly uses the response data to estimate where the variation in effect occurs. The SVC model reflects the hypothesis that non-stationarity exists and is occurring at a fine-scale resolution (e.g., population-level variation). The SVC model is a form of "local spatial modeling" that uses neighborhood response and predictor information to estimate ecological relationships at a fine scale (Lloyd 2010a). Ultimately, the three models evaluated whether a given species showed signs of nonstationary in ecological relationships, and if so, at which resolution (i.e., regionally, or finer), as evaluated by model fit and predictive performance.

We used single-season, single-species occupancy models as the underlying regression-based SDM; the occupancy model is a regression-based SDM that accounts for imperfect detection hierarchically by separately modeling the ecological and observation processes (MacKenzie et al. 2017a). We built models for four species – white-tailed deer, coyote, raccoon, and bobcat – that represent a range of ecological associations and abundances across the state. For example, bobcat is more of a specialist species relative to the others.

We implemented the single-species, single-season occupancy model described by (MacKenzie et al. 2017a) for each species. We start with a description of an occupancy model with covariates affecting the detection and state process ("GLM model"):

$$\begin{gathered} Z_i \sim Bernoulli(\psi_i) \\ logit(\psi_i) = \mathbf{X}_i^T \boldsymbol{\beta} + \theta_i \\ y_i \sim Binomial(n_i, \mu_i) \\ \mu_i = Z_i * p_i \\ logit(p_i) = \mathbf{X}_i^T \boldsymbol{\alpha} \end{gathered} \tag{1}$$

where Z_i is an unobserved parameter indicating whether grid cell i is truly occupied (1 = occupied, 0 = unoccupied), ψ_i is the probability that a grid cell is occupied, $\mathbf{X}_i$ is a vector of covariates ("Landscape Composition and Configuration" categories in Table 1), $\boldsymbol{\beta}$ is the corresponding vector of fixed regression coefficients, θ_i represents an improper conditional autoregressive (CAR; (Banerjee et al. 2014) random effect, y_i indicates the number of detections within grid cell i out of a total of n_i camera trap nights per grid cell, p_i is the probability of detecting a species conditional on its occurrence state, and $\boldsymbol{\alpha}$ is an additional vector of regression coefficients describing effects on the detection process ("Site Variation" categories in Table 1). The CAR random effect creates spatial dependence among neighboring grid cells. Specifically, for responses $y = (y_i, \dots, y_N)$ the improper CAR is

$$p(y_i | y_{-i}, \tau) \sim N\left(\frac{1}{w_{i+}} \sum_{j \in N_i} w_{ij} y_j, w_{i+}\tau\right) \quad (2)$$

where y_{-i} represents all elements of y except y_i, the neighborhood N_i of grid cell i is the set of all j for which region j is a neighbor of region i, $w_{i+} = \sum_{j \in N_i} w_{ij}$, and the Normal distribution is parameterized in terms of precision τ. Adding a CAR random effect to the linear predictor assumes that the response at grid cell i is a function of site covariates and the responses at neighboring grid cells.

To model ecoregion-specific focal covariate effects ("ecoregion model"), we indexed each regression coefficient of the ecological process in Eq. 1 with an indicator of which physiographic region site i is located within

$$logit(\psi_i) = \mathbf{X}_i^T \boldsymbol{\beta}_{(e)} + \theta_i \quad (3)$$

where e is the indicator of ecoregion taking on either coastal plains, mountains, or piedmont; this results in three independent estimated coefficients for each covariate in **X**.

Finally, we fit a full spatially varying coefficients model ("SVC model") that allows for the effect of a given covariate to vary over space. To implement this, we modified the linear predictor of the latent state Z_i in Eq.1

$$logit(\psi_i) = \mathbf{X}_i^T \boldsymbol{\omega}_i \quad (4)$$

where $\boldsymbol{\omega}_i$ is a vector of spatially-varying coefficients arising from a spatial Gaussian process (Gelfand et al. 2003). Each *j*-th element in the $\boldsymbol{\omega}_i$ vector at location i is distributed as an improper CAR random effect as described in Eq. 3.

For all spatially varying coefficients, we calculated a "stationarity index" (SI) to compare the global coefficient of each landscape composition and configuration metric estimated using the GLM model to the estimated SVCs. For each coefficient, we calculated the interquartile range in the spatially varying coefficient and divided it by two times the standard error of the global regression coefficient for the same variable from the GLM. Thus, higher values indicate more spatial variation in the ecological relationship and values >1 suggest support for nonstationarity (Osborne et al. 2007a).

We estimated parameters under Bayesian inference with Markov chain Monte Carlo (MCMC) sampled using the r-nimble package (de Valpine et al. 2017) within the R statistical computing environment (R Core Team 2020a). We generated 3 chains of 20,000 iterations after a burn-in of 10,000 iterations and thinned by 20 iterations to reduce file size. We used normally distributed, uninformative priors (mean = 0, SD = 5, precision = 0.04) for regression coefficients in both the observation and ecological models. We used WAIC calculated within r-nimble to evaluate and compare model fit. To evaluate predictive performance of models we used Brier Score, which is a proper score function described by the mean squared error of the probabilistic model predictions and the observed presence or absence in the data, and the well-known area under the receiving operator characteristic curve (AUC) (Faraggi and Reiser 2002, Long et al. 2011, Zipkin et al. 2012, Pacifici et al. 2016).

Additional file and data manipulation was done using packages *data.table* (Dowle and Srinivasan 2019), *here* (Müller 2017), *sf* (Pebesma 2018), and the *tidyverse* (Wickham et al.

2019). Data visualization was done using packages *ggplot2* (Wickham 2016), *tmap* (Tennekes 2018), and *RColorBrewer* (Neuwirth 2014).

Results

Camera trapping

Volunteers deployed camera traps in all 100 counties of North Carolina with a greater concentration of deployments around more populated areas (Figure 1). Forty-nine percent of the deployments were in the Piedmont region, followed by 27% in the Coastal Plains, and 24% in the Mountain region. We recorded 37,408 detections of deer, raccoons, coyotes, and bobcats over 73,967 trap nights of effort at 3,678 camera trap locations. Deer were consistently the most detected species, making up 42% of the detections (30,703 detections; naïve occupancy = 0.82), followed by raccoon (4,739 detections or 13% of detections; naïve occupancy = 0.34), coyote (1,653 detections or 4%; naïve occupancy = 0.21), and bobcat (313 detections or 0.8%; naïve occupancy = 0.05).

Model comparison

Evidence of non-stationarity varied among the species tested (Table 2; Table 3). White-tailed deer was the only species with top support for the SVC model, although the SVC model was also within 2 WAIC points of the top model for bobcat (Table 2). There was little support for the SVC model for coyote or raccoon, instead the Ecoregion model had the most support for both species (Table 2). Irrespective of model fit, the SVC models consistently had the highest AUC and lowest Brier Score (Table 2), indicating superior predictive performance.

The range and magnitude of estimated occupancy probabilities for each species were relatively consistent across all models (Figure 2); estimated occupancy probabilities were most uncertain under the SVC model for all species, though deer had comparable uncertainties to the GLM and Ecoregion model (Figure 2). Averaging across all models, mean occupancy was 0.88 (SD = 0.13), 0.79 (SD = 0.14), 0.67 (SD = 0.15), and 0.36 (SD = 0.18) for deer, raccoon, coyote, and bobcat, respectively. Similarly, we documented consistent cell-level detection probabilities across all models for each species (Figure S1); at the cell-level, we poorly detected most species (detection probabilities $p <= 0.1$), except deer (average $p = 0.23$) (Figure A1).

The estimated landscape coefficients from the GLM model had similar strength and direction as the Ecoregion model, but some relationships varied across ecoregions (Table 4; Table A1). Although the added flexibility of the ecoregion models improved fit for some species, the smaller, regional sample size also increased uncertainty in the posterior distributions of the landscape covariates (Table 4; Table A1). Site coefficient estimates were nearly identical across all models although we did not expect this vary because these components remained constant for all species and models (Figure A2).

Ecological relationships

All species tended to have stronger and more precise relationships with landscape composition (i.e., cover of forest and developed areas) than landscape configuration (i.e., contagion index and patch richness density), resulting in distribution estimates more reflective of the land cover of North Carolina (Figure 3; Figure A3; Table 4; Table A1). Further, some species tended to have spatially different relationships with landscape composition metrics in both the Ecoregion and SVC models (Table 3; Table 4). For deer, the sole species with the most support for the SVC model, the estimated coefficients for landscape composition effects had stronger and more varied relationships than configuration metrics (Figure 4; Figure A4; Table 4). Specifically, deer had the most spatially varying relationship with forest cover (SI = 2.13), ranging from strongly negative ($\hat{\beta} \leq -2$) to strongly positive ($\hat{\beta} \geq 3$) (Table 3; Figure 4). Interestingly, for deer the strongest relationships with landscape composition metrics tended to be at transition zones between the major ecoregions in the state (Figure 4). Although the SVC model was within 2 WAIC points of the top model for bobcat, all SIs were less than 1 (Table 3; Figure A5; Figure A6). However, across all species and predictors, the lowest overall SI was bobcat's negative relationship with human development, reflecting the most constant relationship across space (Table 3; Figure A5; Figure A6). Additionally, bobcat had the lowest contagion index SI suggesting their need for large, contiguous blocks of undeveloped land cover was ubiquitous across their distribution in North Carolina (Table 3).

We documented strong and precise relationships with most of the proximity-based, site-level covariates, although this varied across species. For bobcat, there was a strong, positive relationship with distance to buildings (i.e., greater distance, greater site use by bobcats), but little effect of proximity to roads or water sources (Figure A2). In contrast, raccoon, coyote, and deer all had strong, negative relationships with distance to building (Figure A2). Raccoon additionally had a strong, negative effect of distance to road, whereas deer and coyote responded positively (Figure A2). Coyote also had a precise, positive effect of distance to water source while deer and raccoon had negative relationships with these features (Figure A2).

Potential Drivers of Nonstationarity

We conducted additional post-hoc analyses to investigate what might be driving nonstationarity in the relationship between deer and forest cover. We developed five hypotheses for potential causes of the nonstationarity (Table 5) and selected additional covariates to add to the model. The four primary hypotheses included spatial variation in resource availability or predator pressure, intraspecific variation in deer populations across North Carolina, and missing environmental covariates. We believed that the effect of predator pressure may depend on resource availability and intraspecific deer variation, so we created models to describe this interaction. We used the amount of forest cover within a given cell based on the five classes of forest cover in 2018 MODIS data to reflect the resource availability hypothesis (Sulla-Menashe and Friedl 2018). We used the combined detection rate (ratio of detections to the number of camera deployment days) of American black bear (*Ursus americanus*), coyote, and bobcat to reflect wildlife predation risk to deer. We also included mean deer hunter effort (i.e., number of hours spent hunting) in each grid cell based on surveys from the North Carolina Wildlife Resources Commission (annual mail surveys 2016 – 2019) to reflect human predation risk. The hunter information is summarized at the county-level, so all grid cells within a given county were assigned the same hunter effort value. To test whether intraspecific variation was the driver of non-stationarity, we used five deer management units (DMUs) delineated by the North Carolina Wildlife Resources Commission as a categorical predictor. The North Carolina DMUs are based on biological differences across the state including peak conception date, average number of bucks harvested, latitude/longitude, and average weights of 1.5-and 2.5-year-old does (Howard et al. 2015, Shaw et al. 2016, Serenari et al. 2019). Given that we did not have genetic data to explicitly test intraspecific variation, we used the DMUs as a proxy because they reflect differences in deer populations driven by an interaction of internal (e.g., genetic differences) and external (e.g., soil types) differences across the state. To test each hypothesis, we fit linear models with the spatially varying coefficient of forest cover as the response, and using the hypothesis-specific covariates as predictors, and then compared with AIC to evaluate support.

The greatest support was for the hypothesis that the interactive effect of predator pressure and intraspecific deer variation explained the spatial variation in the effects of forest cover on deer occupancy (Table 5; Figure 5). For example, in the western DMU, deer exhibited a negative relationship with forest cover under low predator pressure scenarios whereas high predator pressure conditions resulted in a positive effect of forest cover (Figure 5).

Discussion

Although it is widely recognized that ecological relationships can vary across the range of a species, most SDMs presume spatial stationarity. Violations of this assumption could lead to inaccurate statistical inference (Gelman et al. 2001) and here we show how accounting for nonstationarity with spatial random effects and SVC models can substantially improve model fit and predictive performance. Further, our results showing spatial non-stationarity in ecological relationships illustrates how the typical assumption of one spatially fixed parameter describing ecological relationships is not valid for some species. Most importantly, given the range in resolution and strength at which non-stationarity occurred, it is important to explore the possible drivers of non-stationarity and how the spatially varying relationships reflect the ecology of the species.

Spatially varying relationships between wildlife and environmental characteristics reflect biological differences within species that could be driven by complex interactions of external and internal factors. Evidence of non-stationarity may be indicative of a given species' ability to exploit a range of environmental conditions by locally adapting, but the magnitude of variability is usually governed by the species' niche (Chase and Leibold 2003). That is, the degree to which species-environment relationships exhibit non-stationarity is likely a function of the species' niche breadth, or total niche width, which describes the variation in resources exploited by a species (Roughgarden 1972). Further, populations of species with greater individual foraging specialization (or other behavioral specializations) are more likely to create patterns of ecological non-stationarity in species-environment relationships, especially when analyzed over large areas (Sargeant 2007). This does not necessarily exclude the possibility of ecological non-stationarity in species with high niche conservatism because even if there is high tendency for species to retain ancestral ecological characteristics, if the species has a relatively high niche width, then it is reasonable to expect varied responses to environmental conditions across the range of the species (Wiens et al. 2010).

Local spatial modeling outperformed conventional SDMs and other data partitioning approaches (i.e., regional effects) for deer, which follows several studies that have formally accounted for non-stationarity either through Geographically weighted Regression (Fotheringham et al. 2003) or SVC models (Gelfand et al. 2003). Past research studying spatial nonstationarity has typically focused on ecological metrics such as species richness, diversity, or turnover, but our research shows that this also occurs at, and below, the species-level, reinforcing the idea that spatial non-stationarity is likely a common phenomenon in ecology (Foody 2004, Shi et al. 2006, Osborne et al. 2007a, Fortin et al. 2008, Windle et al. 2009, Jarzyna et al. 2014, Smith et al. 2019b).

Deer was the only species with the SVC model receiving the most support, driven primarily by the high variability in their relationship with forest cover across the state. We followed these findings with a post-hoc analysis to explore what might be the driver(s) behind this non-stationarity and found the most support for an interactive effect of external (predator pressure) and internal (population differences) factors. In the western-most part of the state where forest cover is most contiguous and relatively little human development exists, deer exhibited a negative relationship with forest cover under low predator pressure suggesting that local populations exploit different land cover types under low-risk scenarios when forest is abundant. However, in the central part of the state where there is the greatest human population densities and hunter effort, deer consistently had strong, positive relationships with forest cover – regardless of predator pressure levels – emphasizing the importance of this land cover feature for predator and human avoidance. Looking at individual hypotheses, deer population biology received far greater support than predation risk alone (R^2 = 0.57 and 0.30, respectively). Specifically, the central and northwestern management units were quite different than the other regions and had the strongest, positive relationship with forest cover, which corresponds with the hypothesis of forest as a predator avoidance tactic.

Although the top model for bobcat was the GLM model, the SVC model was within 2 WAIC points suggesting substantial support for spatially varying ecological relationships. Most importantly, by testing bobcat we may have observed the strongest signal of whether spatial non-stationarity is related to individual species ecology, an important yet unresolved research question. We estimated the lowest stationarity index - across all species and covariates – for the relationship between bobcat and housing density. This is a well-known relationship in North Carolina and our results corroborate previous work in the state (Parsons et al. 2019). Although we found strong evidence that bobcat tend to avoid highly populated areas, our study took place over a relatively small portion of their geographic range, and as studies increase in geographic scale, this relationship will not always be the case (Young et al. 2019). Indeed, Young et al. (2019) reported significant evidence of urban-dwelling bobcats in Texas, USA, and this variation highlights the importance of relaxing the stationarity assumption in large-scale SDMs as SVCs may highlight areas where ecological relationships deviate from the expectation (Tigas et al. 2002, Clare et al. 2015, Lombardi et al. 2017).

Coyote and raccoon both had top support for the Ecoregion model indicating that ecological relationships may also vary but those relationships are relatively uniform within a larger area. Coyote responded similarly to forest cover in the coastal and mountain region, but similarly to development in the mountains and piedmont regions. With forest cover, coyotes had a strong, positive response in the piedmont region; this may be due to fragmented green spaces providing the necessary resources to exist in developed areas (Parsons et al. 2018, 2019). Similarly, coyotes showed avoidance of developed spaces in the coastal region where relatively less forest cover exists, providing more support for the need of this resource to maintain high occurrence probabilities in human-dominated spaces. Raccoon, in contrast, tended to use developed areas but this relationship was strongest in the piedmont region. However, unlike the other regions, raccoon had a negative response to patch richness density in the piedmont, suggesting a positive relationship with contiguous land cover in highly developed areas.

The tradeoff of using local spatial models, like our ecoregion or SVC approaches, to account for nonstationarity is that the sample size for each model is smaller, and thus models should be able to reliably estimate fewer explanatory variables. That is, a conventional global regression analysis will generally be able to accommodate many more predictors (Green 1991) because they use all observations in one model, whereas local models only consider a subset of nearby observations and thus typically will be able to estimate fewer predictor effects. The result of this tradeoff is apparent when comparing the magnitude and uncertainty estimates of predictor variables across spatial and non-spatial models, where we observed an increase in uncertainty and moderate changes in magnitude as we incorporated spatial random effects and allowed the focal coefficients to vary over space. The increase in uncertainty, particularly for the spatial ecoregion and SVC models, in part reflects fewer observations available to inform estimates, and when combined with changes in magnitude, provides evidence that the independence assumption of the conventional GLM were not being met. This phenomenon is typically referred to as the *bias-variance tradeoff* and is a well-known property of increased model complexity (i.e., including spatial random effects; (Bolker et al. 2009). Previous work in ecological non-stationarity demonstrated this as well, where Meehan et al. (2019) had greater uncertainty in information-poor areas when using an SVC model compared to traditional analyses, and Osborne and Suárez-Seoane (2002) had larger variance estimates when partitioning data for large-scale SDMs. Additionally, Jarzyna et al. (2014) found differences in coefficient magnitude when comparing spatial and non-spatial models, and Byrne et al (2019) found changes in direction and strength of effect across ecoregions.

The interest in moving towards open-access, publicly-available animal occurrence repositories such as the Global Biodiversity Information Facility (GBIF) will likely become primary resources to inform ecological conservation and management (Kays et al. 2020a). One of the most appealing results of local spatial modeling is not only the improved predictions to future conditions across landscapes (e.g., regions, countries), but the high-resolution of responses to predictors at the management-unit level. For example, we documented strong support that deer management zones in North Carolina – zones based explicitly on population differences across the state – were also exhibiting different ecological relationships across the state. The powerful combination of predicting large scale patterns while also emphasizing areas small enough that management or protection highlights the usefulness of this approach in ecology. Thus, we suggest those interested in building SDMs should consider the use of SVCs when the spatial extent of the area of interest exceeds population-level inference.

Table 1: Ecological predictors used to model space-varying ecological effects across North Carolina, USA during October 2015 – December 2019.

Category	Covariate	Description
Landscape Composition	Forest Cover	Proportion of forest cover in 5 km x 5km grid cell
Landscape Composition	Housing Density	Mean housing density (houses/km) in 5km x 5km grid cell
Landscape Configuration	Contagion Index	The degree of clumping among land cove types in 5km x 5km grid cell, or the propensity for a raster pixel of a given land cover class to be neighboring a different land cover class
Landscape Configuration	Patch Richness Density	Number of land cover types per 100 ha in 5km x 5km grid cell
Site variation	Distance to Road	Mean minimum distance to road from camera traps within a grid cell
Site variation	Distance to Building Structure	Mean minimum distance to building structures from camera traps within a grid cell

Table 1 (continued).

Site variation	Distance to Water Source	Mean minimum distance to water source from camera traps within a grid cell
Site variation	Detection Distance	Mean detection distance for cameras within a grid cell
Survey Effort	Trap Nights	Length (days) of camera trap deployment used to represent effort of camera trapping.

Table 2: Model comparison of fit (WAIC) and predictive performance (Brier Score, AUC) for single-season occupancy analysis of bobcat, coyote, raccoon, and white-tailed deer in North Carolina, USA during Oct 2015 – Dec 2019. The three models used were: GLM – generalized linear model with a spatial random effect; Ecoregion – GLM with a spatial random effect and ecoregional random effects on landscape composition and configuration metrics; SVC – spatially varying coefficients model. We report on the change in (delta) the Widely Applicable Information Criterion (WAIC) from the best fitting model (0 indicates best fitting model), Area under the curve (AUC; higher values are better, perfect score = 1), and Brier Score (BRIER; lower values are better, perfect score = 0). Bold values indicate the best score for each category and models are sorted based on best lowest WAIC. Asterisk indicates model <= 2 WAIC points of top model.

Species	Model	Δ WAIC	AUC	Δ AUC	BRIER	Δ BRIER
	GLM	**0**	**0.77**	**0.02**	**0.16**	**0**
Bobcat	SVC*	1.57	0.79	0	0.17	0.01
	Ecoregion	17.37	0.74	0.05	0.19	0.03
	Ecoregion	**0**	**0.63**	**0.04**	**0.31**	**0.01**
Coyote	SVC	5.62	0.67	0	0.3	0
	GLM	6.09	0.62	0.05	0.3	0
Raccoon	***Ecoregion***	**0**	**0.64**	**0.01**	**0.25**	**0.01**

Table 2 (continued).

	SVC	9.63	0.65	0	0.24	0
	GLM	18.02	0.65	0	0.24	0
	SVC	**0**	**0.83**	**0**	**0.08**	**0**
Deer	Ecoregion	10.4	0.83	0	0.08	0
	GLM	16.82	0.8	0.03	0.09	0.01

Table 3: Stationarity Indices for bobcat, coyote, raccoon, and white-tailed deer in North Carolina, USA during October 2015 – December 2019. For each variable, we calculated the interquartile range in the spatially varying coefficient and divided it by two times the standard error of the global regression coefficient for the same variable from the GLM. Thus, higher values indicate more spatial variation in the ecological relationship and values >1 suggest support for nonstationarity.

Species	Landscape Configuration		Landscape Composition	
	Contagion	Patch Richness	Forest Cover	Development
Bobcat	0.34	0.49	0.44	0.27
Coyote	0.69	0.48	0.62	0.50
Raccoon	0.62	0.38	0.33	1.36
Deer	0.85	0.47	2.13	0.91

Table 4: Comparison of estimated coefficients showing which environmental covariates had significant positive (+) or negative (-) effects (determined by whether the 95% credible interval overlapped zero). Coefficients overlapping zero were designated with "NS". For all covariates we also show the ecoregional effects under sum-to-zero constraints (Red = Coastal; Blue = Mountains; Green = Piedmont). Using this constraint results in the magnitude of effects being relative to compared regions rather than absolute effects; absolute effects from Ecoregion models followed the directions listed in the GLM and SRE models. Absolute effects are reported in Table A1.

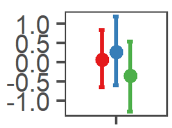
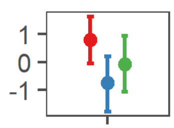
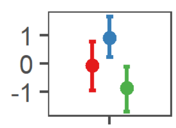
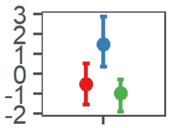
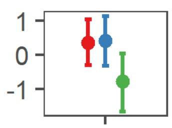
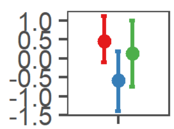
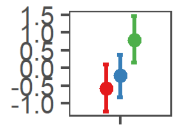
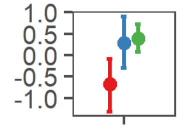
	Landscape Configuration				Landscape Composition			
	Contagion		Patch Richness		Forest Cover		Development	
Species	GLM	Ecoregion	GLM	Ecoregion	GLM	Ecoregion	GLM	Ecoregion
Bobcat	NS		NS		NS		-	
Coyote	+		NS		+		+	

Table 4 (continued).

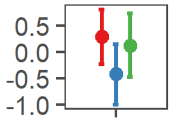
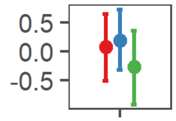
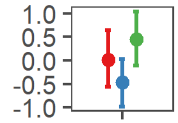
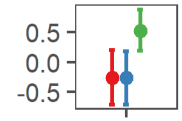
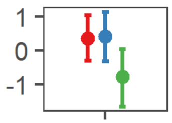
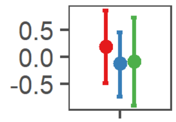
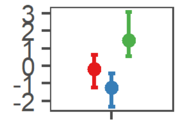
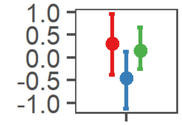
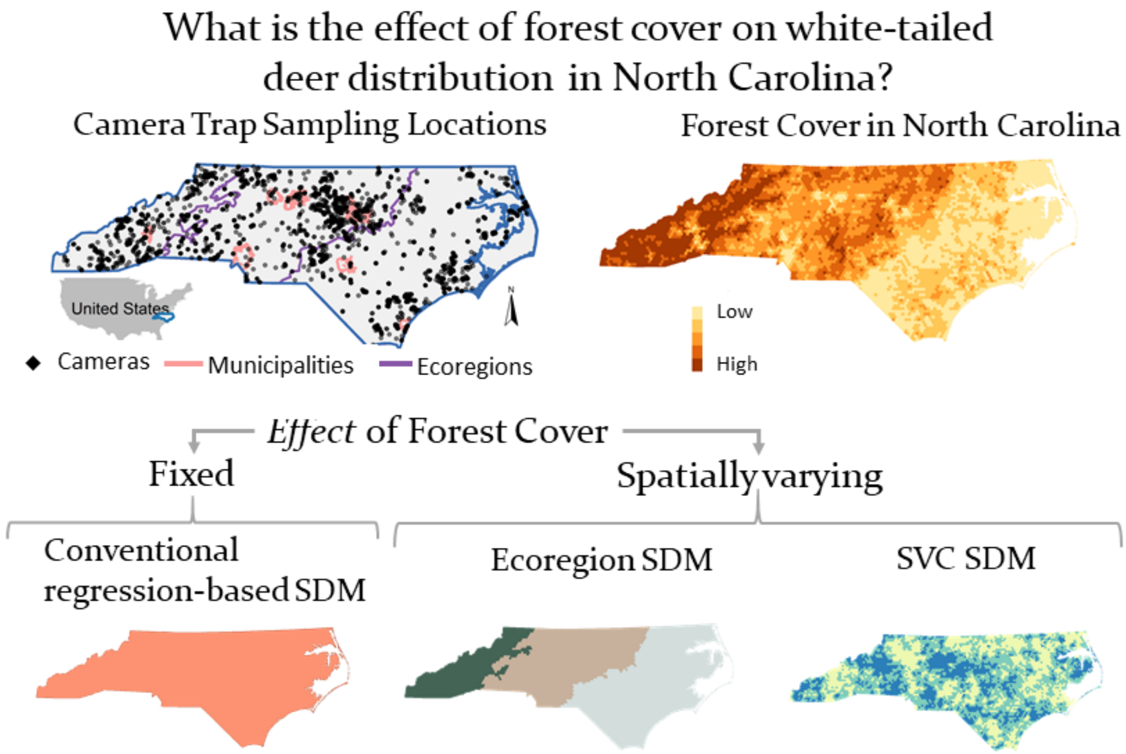
Raccoon	NS		NS		-		NS	
Deer	NS		NS		NS		NS	

Table 5: Hypothesis of potential drivers of non-stationarity in the relationship between white-tailed deer and forest cover.

Hypothesis	Covariates	Δ AIC	Degrees of Freedom	R-squared
Predator x Biology	Deer Management Units; wildlife predator detection rates; deer hunter harvest effort	0	16	0.65
Global	All	558	13	0.60
Deer Biology	Deer Management Units	1094	6	0.57
Predator x Resource	Wildlife predator detection rates; deer hunter harvest effort; forest availability	3786	7	0.30
Predator Pressure	Wildlife predator detection rates; deer hunter harvest effort	3796	4	0.30
Resource Availability	Forest availability	5589	7	0.02
Missing Covariate	Forest types (coniferous, deciduous, mixed)	5654	3	0.03

Figure 1: Conceptual diagram of non-stationary effects of forest cover on white-tailed deer (Odocoileus virginianus) distribution in North Carolina, USA. Conventional regression-based SDM refers to a Species Distribution Model where coefficient effects are assumed to be stationary across space. Ecoregion SDM depicts coarse-level non-stationarity in the effect of forest cover on deer distribution. A Bayesian Spatially Varying Coefficients Model to account for fine-scale variation in covariate effects is represented by SVC SDM. Under each scenario, we describe the spatial extent that the covariate is assumed to be constant (covariate relationship spatial extent), the number of observations from our empirical dataset that would be available to estimate this effect (No. of observations in spatial extent), the possible number of covariates the given number of observations could theoretically estimate (Possible No. of covariates; * = based on the standard guide of 30 observations per explanatory variable), and a qualitative evaluation of the risk of assumption violation for each approach. The inset of North Carolina in the top left represents actual camera trapping locations (n = 3678; black dots) during 2015 – 2019; the three primary physiographic regions (i.e., ecoregions) in North Carolina are delineated along with municipalities human populations greater than 90,000.



Model	GLM SDM	Ecoregion SDM	SVC SDM
Covariate relationship spatial extent:	139,391 km²	21,352 – 68,253 km²	Depends on estimated correlation range; ~20 km²
No. of observations in spatial extent:	3260	769-1600	1-386
Possible No. of covariates*:	108	25-58	1-12
Risk of conventional assumption violations:	High	Medium	Low

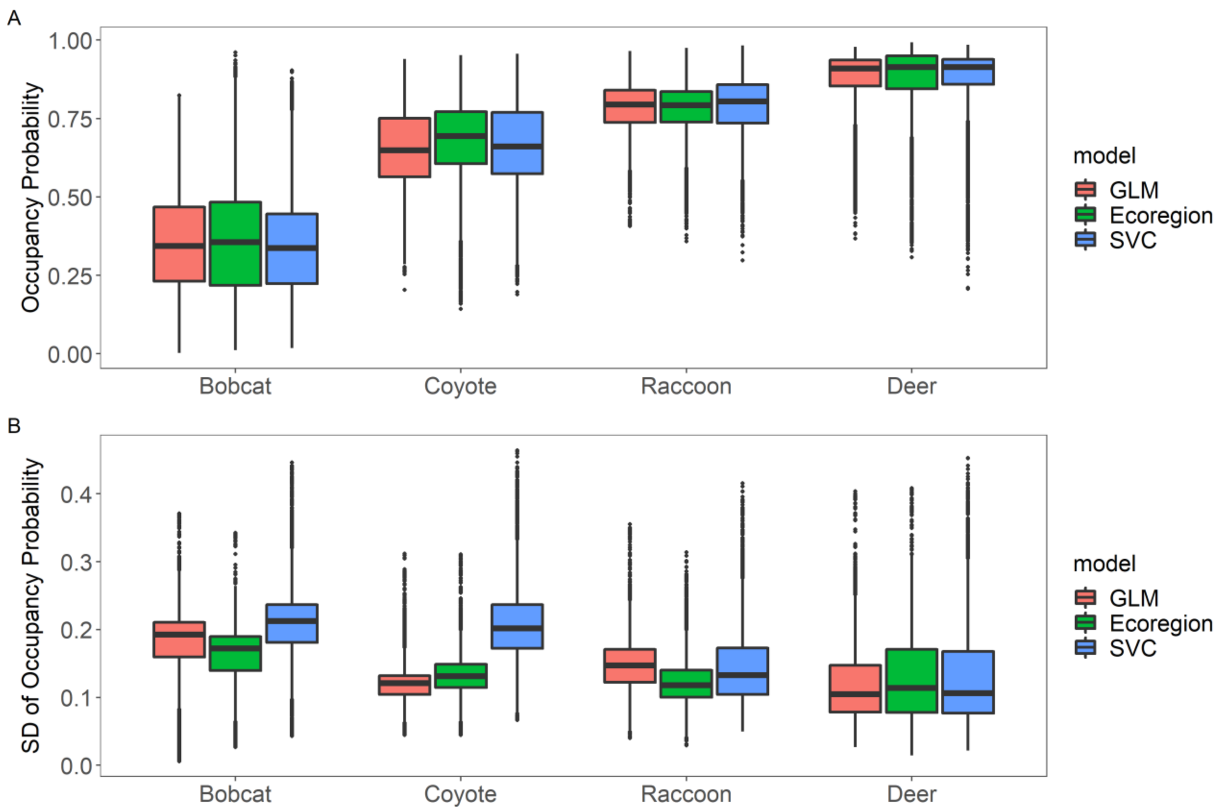


Figure 2: (A) Distribution of all estimated occupancy probabilities (n = 5497 for each model) for bobcat, coyote, raccoon, and white-tailed deer in North Carolina, USA during Oct 2015 – Dec 2019 along with the (B) distribution of uncertainty (SD = standard deviation) in those estimates.

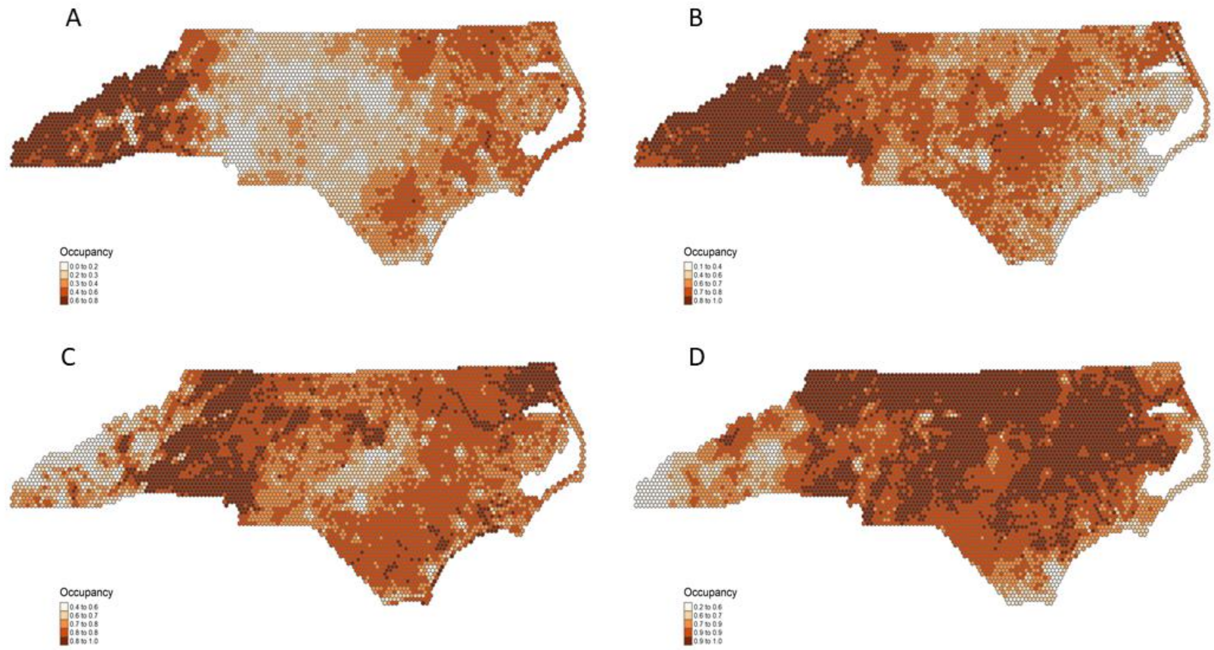


Figure 3: Estimated occupancy probability from the species-specific top model (Table 2) for (A) bobcat, (B) coyote, (C) raccoon, and (D) deer during Oct 2015 – Dec 2019 in North Carolina, USA.

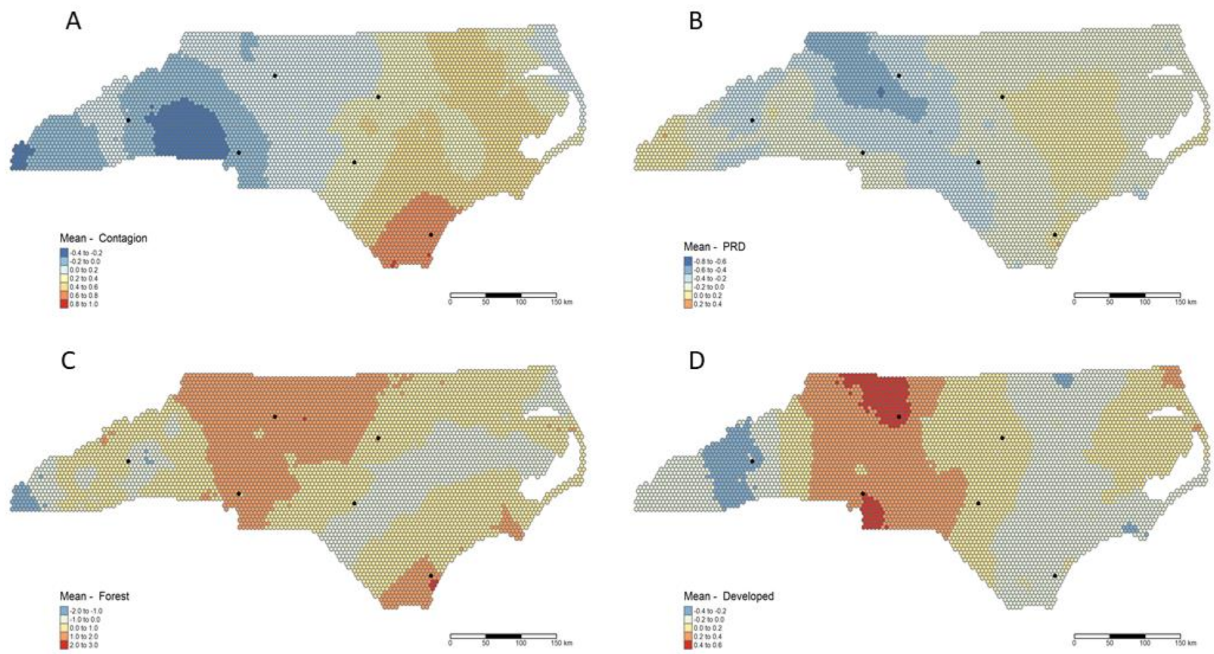


Figure 4: Spatially varying coefficient estimates (and stationary indices) of (A) contagion index (SI = 0.85), (B) patch richness density (SI = 0.47), (C) forest cover (SI = 2.13), and (D) developed (SI = 0.91) for white-tailed deer during Oct 2015 – Dec 2019 in North Carolina, USA. Black dots are for spatial reference and represent cities with populations with 100,000 people or more.

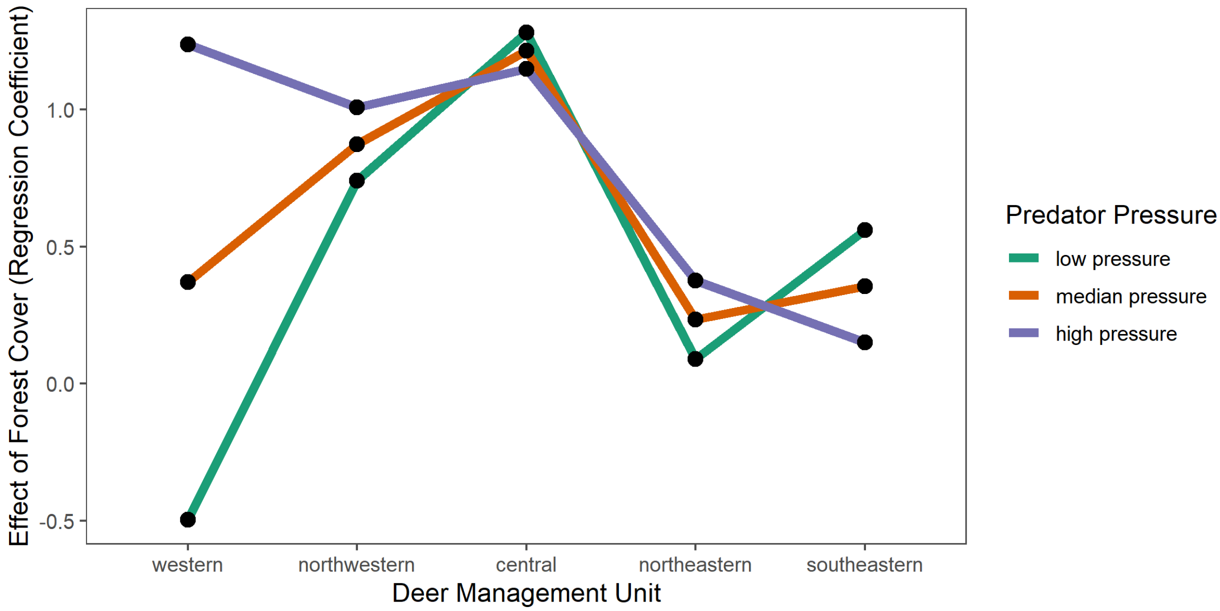


Figure 5: The interactive effect of predator pressure and deer biology best explain nonstationary effects of forest cover on white-tailed deer. The y-axis is the estimated spatially varying effect of forest cover on white-tailed deer (Figure 4C) from models described in Table 2. The x-axis lists the North Carolina Deer Management Units and serve as proxies for biological differences in deer populations across the state. Predator pressure includes effects of wild (American black bear and coyote) and human predators, as estimated by deer hunting pressure across the state.

CHAPTER 3: A GUIDE TO CHECKING DATA SOURCE COMPATABILITY IN INTEGRATED SPECIES DISTRIBUTION MODELS

Introduction

Although fundamental to ecology, understanding how species distributions change in space and over time has entered a new era in the 21st century due to massive increases in available data. Large, unstructured datasets (e.g., GBIF) as well as smaller structured monitoring programs (Kays et al. 2020b) are increasingly being made openly available, and both data types are providing the foundation to discovery while new developments in statistical ecology provide the means to new ecological insight. Recently, novel approaches to combining structured and unstructured data have been introduced as Integrated Species Distribution Models (ISDMs) - a statistical framework that fuses all available information to inform current and future animal distributions (Pacifici et al. 2017, Miller et al. 2019, Isaac et al. 2020). Typically, when *structured* data (i.e., data derived from a planned monitoring program) is integrated with *unstructured* (i.e., data arising without a formal sampling protocol, often opportunistically), practitioners find reduced uncertainty, increased spatial coverage, and potentially novel ecological insights (Pacifici et al. 2017, Schank et al. 2017, Robinson et al. 2020). Despite the great potential, the ISDM framework is in its infancy and our understanding of the known limitations nearly matches the developmental stage. Indeed, in a recent review, Isaac et al. (2020) outlined several outstanding problems in ISDMs, including how to evaluate information gained by combining datasets, the extent to which model assumption violations affect outcomes, and identifying when simpler analytical approaches are preferred over ISDMs (Simmonds et al. 2020). Depending on the exact datasets combined, it is possible that integrated models could be worse than simpler approaches with single datasets, although there is presently no guidance on identifying these scenarios.

Integrated species distribution models are challenging to specify and computationally demanding to fit because they require advanced statistical knowledge, programming capabilities, and often times enhanced computing power. Given these difficulties, it would be useful to know whether integration will be superior to simpler models *prior* to model fitting. Fortunately, some features of the individual data sources may be informative of whether integration will be beneficial for a given ecological problem. Namely, the quantity, quality, and spatial overlap of the data sources may be informative of the potential resulting improvements (e.g., reduced uncertainty) or diminishes (e.g., biased estimators) when choosing ISDMs over simpler approaches (e.g., generalized linear models) (Figure 6). For example, in cases where abundant structured data exists with minimal unstructured data, ISDMs may not be a fruitful approach. Further, ISDMs may not be useful when abundant unstructured data exists without spatial overlap among the data sources. In either case, it is unclear how these scenarios fair in an ISDM framework.

Here we propose a formal data checking process for ISDMs to address the challenges discussed, aimed to guide researchers and practitioners in the steps to data integration (Figure 7). We divided the process into three primary steps: data compatibility, preliminary data relationship checks, and evaluation of model fitting. Data compatibility includes steps to ensure preliminary data relationship checks are optimal by using a formal outlier removal procedure, filtering observations for minimum search effort, and temporally aligning data sources to ensure observations reflect similar ecological processes. Step two is a series of indicators to guide practitioners in determining whether to incorporate multiple data sources to inform species distributions. The data checking process is designed to highlight potential issues *prior* to integration. For example, the minimum sampling effort filter applied during the data compatibility step will flag data quantity imbalance scenarios and suggest alternatives to making use of the limited unstructured data. If data integration is appropriate, we provided guidance on *how* to integrate, based on the level of confidence in the unstructured data (Pacifici et al. 2017). Finally, we provided guidance on objective decision analysis to decide which integration approach best meets the objective of the analysis, where we propose the use of structured decision making to identify the optimal combination of data sources (Runge et al. 2020).

We apply our data checking procedure to estimate species distributions in North Carolina, USA, where a surplus of wildlife records is available across the state (four unique data sources). The four data sources include a state-wide volunteer-based camera trapping project and three long-term monitoring programs used by a state wildlife agency. We focus our efforts on coyote (*Canis latrans*), bobcat (*Lynx rufus*), and raccoon (*Procyon lotor*) as they are widespread across the state and are well represented across the four available data sources. The data sources arise from both structured and unstructured surveys and vary in the level of effort made in the data collection process. We used simulations to provide clarity on the importance of data features (e.g., quantity) and preliminary checks (e.g., regressed relationships) in the decision tree of data checking for ISDMs. We hypothesized that data sources with the greatest agreement, correlation, deviance explained by unstructured data sources, effort, and spatial overlap would be selected as the sources to integrate in an ISDM. we conclude the paper with general recommendations for determining the relevance of integration across ecological problems.

Methods

North Carolina Wildlife Observations

We illustrate our data checking decision tree with four empirical data sources from North Carolina, USA (Figure 8; Table 6). Our primary data source is observations from a standardized, state-wide camera trapping effort (Kays et al. In Review, Lasky et al. In Review) led by the North Carolina Museum of Natural Sciences and North Carolina State University, funded by the North Carolina Wildlife Resource Commission (NCWRC). The three additional data sources were provided by the NCWRC and include Deer Hunter Observation Surveys (Fuller et al. 2019), Hunter Harvest Records, and Trapper Harvest Records (Ryan Myers, pers. comm.). These hunter-contributed datasets are collected to inform large scale management decisions, and comparable monitoring programs are widely used across North America (Ericsson and Wallin 1999, Winchcombe and Ostfeld 2001, Manjerovic et al. 2014).

Volunteer-ran camera trap surveys

During 2015 – 2019, 4,377 camera traps were deployed by volunteers and state biologists across North Carolina. These monitoring efforts resulted in over 2.2 million photographs with 120,671 unique detections of 33 mammal and terrestrial bird species. Cameras were deployed year-round on both private (54%) and public (46%) lands in all 100 counties of North Carolina; the mean number of cameras per county was 4 (min = 1, max = 75). Camera locations were either at the volunteer's discretion or using pre-selected locations determined by stratified random sampling of land cover types in North Carolina (Kays et al. In Review). Deployment sites were unbaited, cameras were mounted 40-50 cm above ground, active 24 hours/day, and had a target deployment length of 21 days. All camera models deployed met standardized minimum requirements including trigger speeds < 0.5 seconds and an infrared flash. Cameras were set to maximum trigger sensitivity, offer no rest period between triggers, and record multiple photographs per trigger. Triggers within 60 seconds of each other were considered one detection and grouped into a sequence. Cameras continued to re-trigger if the animal was still in the detection zone of the camera. Citizen science volunteers used the eMammal software to upload pictures and identify all photographed wildlife species. Uploaded pictures were then vetted by project staff using the eMammal Expert Review Tool, where species identification was confirmed or corrected (McShea et al. 2016). We removed deployments that lasted less than 7 days or greater than 35 days, as this indicated camera malfunction or deployments that did not adhere to protocol. We also spatially thinned deployments that were deployed during the same month of a given year were separated by ≤50 m, resulting in 3,678 camera trap deployments.

Hunter harvest surveys

In North Carolina, only harvests of big game (white-tailed deer [*Odocoileus virginianus*], American black bear [*Ursus americanus*], and wild boar [*Sus scrofa*]) are required to be reported, thus harvest (and trapping) information on the remaining harvestable species is determined through hunter and/or trapper mail surveys. Following each hunting season from 2015-2019, NCWRC mailed surveys to a random sample of about 25,000 state hunting license holders (nearly 4% of the registered hunters) asking hunters to report on their annual hunting efforts. The survey asks individuals to report the counties in which they hunted, the number of days spent hunting, and the number of animals harvested in each county. Following cleaning (e.g., removing surveys with incomplete information), we used 33,488 reports from 22,535 hunters. This culminated in 781 bobcat days, 7,966 raccoon days, and 13,002 coyote days of hunting effort over the four-year period.

Trapper harvest surveys

Trapping in North Carolina is open to 14 wildlife species during 1 November to 28 February, although some local laws adjust season dates, including for red fox (*Vulpes Vulpes*) and gray fox (*Urocyon cinereoargenteus*). Regulations set forth by the NCWRC specify the type of trap (e.g., foothold, box/cage) acceptable for a given species. Trapped wildlife in North Carolina are not required to be reported to the NCWRC, and thus information on trapping effort and individuals trapped is solicited through mailed surveys to *all* registered trappers (~3,000 trappers) in the state; the NCWRC typically has an average mailed surveys response rate of ~50%. During 2015-19, the NCWRC amassed 5,170 survey responses that reported a trapping effort (number of traps * number of days set) of 406,851, 753,863, and 1,252,141 trap nights for bobcat, raccoon, and coyote, respectively, across all seasons.

Deer hunter observation surveys

The NCWRC conducts a Deer Hunter Observation Survey (DHOS) where registered deer hunters report on species observed while hunting. Unlike the Hunter Harvest and Trapper Harvest Surveys, this information is not solicited by the NCWRC through a mailed survey but rather hunters voluntarily register for the program (online and paper enrollment is available); around 1,000 – 2,000 registered hunters volunteer information each hunting season. The NCWRC does, however, advertise the program through public new releases, emailed communication to the big game harvest registration database, and personal communication with NCWRC employees. The information recorded includes the county in which the individual hunted, the date, location type (i.e. public or private land), use of bait (yes or no), the species and characteristics of the individuals (e.g., sex/age) observed while hunting, and the amount of time spent hunting (e.g., hours) as a measure of hunter effort. Hunters reported their effort regardless of whether wildlife was observed (e.g., no wildlife observed during a 2-hour hunt). Prior to cleaning (e.g., removing observations missing location or time information), 153,190 DHOS reports were available (a report constitutes a hunting event, not a unique hunter). After removing missing information, we used 148,977 surveys resulting in a total of 21,157.5 days of hunting effort (507,780.5 hours) during the four-year period. There were 1,222, 6,347, and 4,837 reported sightings of bobcats, raccoons, and coyotes, respectively.

Spatial resolution of data sources

Whereas the volunteer-deployed cameras have known, point-level latitude/longitude spatial information, the three remaining data sources from the NCWRC were available only at county-level, creating a mismatch in the spatial resolution of the data sources. Although these four data sources potentially provide useful information about the distribution of the focal species across the state, the information available is notably different in resolution and origins, and highlights a challenge to make use of all available information to inform distribution of our focal species. Fortunately, recent work on the spatial misalignment of data sources in ISDMs has shown that applying a statistical change of support can overcome this conflict, allowing inference about the underlying distribution using all observations (Pacifici et al. 2019; Figure 9). For all integrated modeling, we apply these techniques to make inference and predictions at the grid cell resolution.

Temporal coverage of data sources

The four data sources had varying temporal coverage (Figure B1). Given that animal behavior can vary seasonally due to differences in movement, diet and seasonal food resources, and breeding cycles, we believed that temporally aligning observations would be important in integration outcomes. To explore the implications of this in ISDMs, we compared the results of year-round camera trap sampling to camera sampling matching the temporal coverage of the other three data sources (i.e., used observations only from Oct – Feb).

Simulation Study

To validate findings in our empirical study, we developed simulations to test when a given data integration approach is preferred and whether empirical trends in the preliminary checks were reproducible in a controlled setting. We focused our simulation efforts on data quantity, data quality of the unstructured data source, and the spatial overlap of the two data sources (Table 7).

Simulated data overview

We started by creating a 20 × 20 rectangular grid resulting in $n = 400$ cells. We then generate an underlying intensity surface $\boldsymbol{\lambda}$ for this grid domain; the intensity surface represents the true, underlying ecological process governing the distribution and abundance of a species, and was a function of a spatially correlated covariate $\boldsymbol{X}$ and a spatial random effect $\boldsymbol{\theta}$. The spatial covariate represents some known ecological driver of distribution and abundance while the random effect acknowledges additional, unaccounted for drivers. We simulated both the covariate and spatial random effect using an improper conditional autoregressive prior, creating spatial dependence among neighboring grid cells ultimately resulting in more similar values among neighbors than distant grid cells (CAR; Banerjee et al. 2014). Specifically, for the spatial random effect $\boldsymbol{\theta} = (\theta_i, \ldots, \theta_n)$ the improper CAR is

$$p(\theta_i | \theta_{-i}, \tau) \sim N\left(\frac{1}{w_{i+}} \sum_{j \in N_i} w_{ij}\theta_j, w_{i+}\tau\right)$$

where θ_{-i} represents all elements of $\boldsymbol{\theta}$ except θ_i, the neighborhood N_i of grid cell i is the set of all j for which region j is a neighbor of region i, $w_{i+} = \sum_{j \in N_i} w_{ij}$, and the Normal distribution is parameterized in terms of precision τ. The ecological covariate $\boldsymbol{X}$ was generated in the same way. The true intensity of the spatial log-Gaussian model is then set to $\log(\lambda_i) = \beta_1 + \beta_2 * X_i + \theta_i$ with $\beta_1 = -0.5$ and $\beta_2 = 2$. We chose these parameter values to result in a mean occurrence of ~0.65. The true regression coefficients used to generate the intensity did not vary across iterations or simulation scenarios. From this intensity, we generated the true occurrence state Z of each grid cell, where $Z_i \sim Bernoulli(\psi_i)$ and ψ_i is the occupancy probability for grid cell i and was calculated using $\psi_i = 1 - e^{-\lambda_i}$. We then generated observations from simulated structured and unstructured surveys based on the true occurrences and probabilities of occupancy.

Simulations - data source one

The structured data source was designed to represent typical occupancy data where observers make repeated visits to a given site and the detection/non-detection of a focal species is recorded during each visit. The structured data were generated as $Y_{1i} \sim Binomial(n_i, pZ_i)$, where detection probability p was set to 0.5 and was conditional on the occurrence state Z in grid cell i. We let the number of binomial trials n_i vary by generating a normal random variable $n_i \sim Normal(21,2)$, to reflect a mean camera deployment of 21 days and standard deviation of 2 which was comparable to our empirical camera trapping observations.

Simulations - data source two

The unstructured data source was also simulated from the same true underlying intensity but in the form of Poisson distributed count data. That is, $Y_{2i} \sim Poisson(E\lambda_i)$ where E represents a search effort term that was held constant at 10 and intensity λ_i as described above. To represent the count data as coarse areal summaries (as in the empirical county-level NCWRC records described above), we created a coarse grid where each coarse grid cell contained four original grid cells, resulting in a total of 100 coarse grid cells. Observations and survey effort from the original grid were summed in each coarse grid cell to represent the data as a coarse aggregate summary (Figure B2).

Simulations - data quantity

We censored observations from the structured and unstructured data to reflect real-world sampling intensities and to explore the effects of data quantity on ISDM results. For the structured data, we randomly sampled the original grid domain to have observations from 10, 20, and 40 percent of the grid cells (resulting in 40, 80, and 160 grid cells with observations, respectively). We chose these values because typical structured sampling protocols tend to sample 15-20% of the study area, which reflected the sampling intensity of the empirical structured data (i.e., camera trap surveys). We also wanted to explore how greater sampling intensity diminished the benefits of integrating unstructured data. We again randomly sampled the original grid to have unstructured observations from 20, 40, and 60% of the grid (80, 160, and 240 grid cells, respectively) to reflect that unstructured observations tend to be greater in quantity than structured surveys. The difference in sampling intensity here could reflect varying detectability or probability of reporting a species in a biodiversity repository such as GBIF.

Simulations - data quality

Unstructured data are prone to measurement error (e.g., over/under counting, species misidentification) and sampling bias (i.e., spatial/temporal clustering). Here we focused on how measurement error affected integration outcomes. While there are several approaches to evaluating the effects of measurement error, such as spatially manipulating survey effort or detection probability to create an inhomogeneous observation process surface (Simmonds et al. 2020), we chose to hold the detection process constant across space but rather directly affect the intensity that generates the simulated counts. That is, we directly added a constant of +/- 0.5 to the Poisson intensity, similar to an offset, that, on average, change the simulated counts by +/- 20 percent to represent over and under counting of the species.

Simulations - spatial overlap among data sources

Finally, we explored how the spatial overlap of structured and unstructured observations affected integration outcomes. We set spatial overlap to values of 10, 25, or 50%, resulting in, for example, 10% of the grid cells where structured and unstructured observations co-occur. Again, consider the scenario where we have structured sampling intensity of 10% and unstructured sampling in 20 percent of the grid, resulting in observations in 40 and 80 grid cells, respectively; when spatial overlap is set to 10%, then unstructured observations co-occur with structured observations in 4 grid cells, and then the remaining 76 unstructured observations occur in grid cells other than those in which structured observations occur. This results in an inverse relationship with geographic coverage of the sampling grid, where when spatial overlap is highest (i.e., 50%), then geographic coverage is lowest (because observations from both data sources are occurring in the same grid cells).

Simulations - modeling

For each scenario described above, we fit a single-data source model to the structured data source to serve as a baseline comparison to the integrated models. We then fit three forms of an integrated model (see below) with the structured and unstructured data sources to three models that vary in their level of confidence in the unstructured data. All models, whether single-data source or integrated - had identical "ecological process" models (i.e., a model of the latent state variable, in our case the distribution of the focal species).

Just as in the data generation procedure, we modeled the true state of occurrence of the species as a Bernoulli random variable, $Z_i \sim Bernoulli(\psi_i)$, where ψ_i represented the probability of cell occupancy and invoked the fundamental relationship between occurrence and abundance, $\psi = 1 - \exp(-\lambda)$, to estimate the linear equation $log(\lambda_i) = \beta_1 + \beta_2 * X_i + \theta_i$. All models assumed priors $\beta_1, \beta_2 \sim Normal(0, 5)$ for the regression coefficients and $\tau \sim Gamma(0.1, 0.1)$ for the precision term in the CAR distribution.

We modeled the structured data source as a binomially distributed random variable conditional on the latent state Z_i with a known number of trials (e.g., camera deployment days) and an unknown but fixed detection probability p that assumed a prior of $p \sim Beta(1, 1)$. Single-data source models included observations from the structured data source. Integrated models included observations from both data sources, and that information was integrated in three ways to reflect the level of confidence we had in the unstructured data source. First, we modeled the structured and unstructured data in a joint-likelihood framework (herein, 'shared model') where the two observations directly *share* parameters. In the simulated model, the shared parameters included the regression coefficient for the ecological variable (i.e., β_2) and the spatial random effect term θ_i. Next, to reflect less certainty in the unstructured data, we modeled the two data sources as *correlated* observations (herein, 'correlation model'). Instead of directly sharing parameters, the correlation model estimates unique regression coefficients and spatial random effects for each data source but both data sources still inform estimates of the distribution of a species. Finally, we include unstructured observations as a covariate (herein, 'covariate model) in the linear predictor of the ecological process model described above. This final model reflects the least confidence in the unstructured data, where the unstructured data source indirectly informs estimates of distribution. To account for variable effort, we used CPUE instead of raw counts for the covariate information. All models were fit using the *r-nimble* R package with a single chain of 10,000 iterations and a burn-in period of 5,000 iterations (de Valpine et al. 2017). Model convergence was evaluated by inspecting trace plots.

Model evaluation

For all simulations, we evaluated model performance using Brier Score and comparing parameter uncertainty. Brier Score is a proper score function to evaluate predictive performance for binary outcomes (lower is better). We also monitored the uncertainty (i.e., standard deviation) of the estimated proportion of cells occupied and the regression coefficient uncertainty (lower is better).

Empirical Data

Data Checking Step One: Data Compatibility

Prior to analysis, all data sources went through an outlier removal process; we did this step after preliminary exploratory analysis indicated some sensitivity to extreme values. We considered an observation as outlier if the observation was greater or less than 1.5 × interquartile range (IQR) of the distribution of observations. We standardized each dataset to represent catch per unit effort (CPUE) where cameras were number of species-specific detections per day of camera survey, DHOS were reported observations per day of hunting summarized at the county-level, hunter harvest records were number of individuals harvested per day of hunting activity summarized at the county-level, and the trapper harvest records were number of individuals trapped per number of trap nights (number of traps × number of days) summarized at the county-level. we only used CPUEs greater than zero when calculating the IQR.

We also noticed substantial variation in sampling effort across data sources and species and wanted to explore whether imposing a minimum sampling effort would affect the outcomes of integration. For instance, considering bobcat trapper harvest data, four counties had zero trapping effort while another county had over 35,000 days of trapper effort (number of traps × number of days deployed). Further, and perhaps more problematic than zero effort, 10 counties in the hunter harvest data source have just one day of hunting effort for bobcat thereby having that single day of effort represent the abundance/distribution of bobcat in that county.

To explore these implications, we tested three approaches to specifying a minimum sampling effort: 1) no minimum sampling effort, 2) observation filtering, and 3) distribution filtering. First, we simply set no minimum sampling effort and used all available data to inform species distribution (herein, 'no filtering'). Next, we used a minimum sampling effort filter that was based on identifying an effort level needed to capture (i.e., detected on camera, harvested, or trapped) some specified threshold of individuals (e.g., 20 bobcat) (herein, 'observation filtering'). Acknowledging that the level of effort needed to detect, for example, 20 bobcat depends on the relative abundance within each county - where some counties may require 10 days of effort and another would require 100 days of effort to reach this threshold - we evaluated effort at each 10th percentile of the observed values and present median effort values across all quantiles. That is, we multiplied the mean effort made to capture the number of individuals at each quantile by the target threshold, divided by the quantile-specific number of individuals, and then took the median value across those 10 quantiles as the minimum sampling effort needed. Counties with effort less than this value were discarded and not used in analyses. The final minimum effort filter tested was based on the quantiles of the actual effort values (herein, 'distribution filter'). Here, users specify some minimum quantile (e.g., 10th percentile) of the effort distribution that the effort needs to be greater than for the observations to be used in the analysis.

We used model deviance explained from fitted binomial generalized linear models (GLMs) to decide which filtering approach was most appropriate for each data source combination. To fit the GLMs, we used the structured data source as the response, represented as the number of successes (i.e., number of days with a detection) and failures (i.e., days without a detection), and the unstructured data as the predictor, represented as CPUE. For each species, we fit nine models that included three models for each unstructured data source (e.g., hunter harvest with no filtering, hunter harvest with observation filtering, etc.). The structured data also went through the filtering process. Given the spatial resolution mismatch between the data sources, we "scaled up" the structured data to match the unstructured data's resolution by summing the number of observations (i.e., camera detections of a given species) and the deployment length (i.e., number of the days the cameras were deployed) in each county. If the maximum deviance explained was equal between the no filtering approach and the other filtering approaches, then we always selected 'no filtering' as the preferred approach; this could happen only if there were no outliers detected thus the two filtering mechanisms had identical results.

To evaluate the effect of Step One in modeling outcomes, we compared the top model from data that went through the outlier removal and filtering process to the same model fit to data without these steps. We used the same metrics and approach described in Step Three (see below) to compare these models, including parameter uncertainty and leave-one-out cross validation of the state variable, occurrence.

Data Checking Step Two: preliminary data relationship checks

A primary objective aimed to identify whether one or more preliminary data relationship check(s) was indicative of which data source combination would best improve integrated model criticisms (i.e., model fit, predictive performance). If one or more of these metrics predicts the utility of integrating a pair of datasets, then they could be useful as simple guides of which data types should be combined in more complicated ISDMs. we compared several metrics that are readily calculated without the need for advanced programming knowledge: Pearson correlations, linear regression relationships, data quantity ratio, spatial overlap, and naïve species occurrence ratios (Table 8).

We used CPUEs to generate Pearson correlation coefficients (PCCs). Correlation coefficients may be a useful indicator because they are easily calculated and indicate some relationship between two or more sets of observations. We summed the number of detections and trials of the structured data source to match the spatial resolution of the county-level data sources. To calculate PCCs when models contained more than two data sources, we took the square root of the difference between the sum of the squares and the product of the squares of the PCCs (i.e., $\sqrt{\sum PCC^2 - \prod PCC^2}$).

We also tested linear regression to evaluate how much variation in a given set of observations is explained by another data source; again, this test is readily available in statistical software and the interpretation is relatively intuitive. For the linear regression test, we regressed each data source against another - always using the structured data as the response and the unstructured data as the predictor - and used the deviance explained to evaluate relationships among the data sources. We also considered the estimated regression coefficient but found that its magnitude and interpretation did not inform the strength of the relationship between the data sources. Like the data filtering approach, we fit binomial GLMs and recorded the regression coefficient and deviance explained by the predictor. When models contained more than two data sources, we imposed an additive effect between the unstructured data sources.

Lastly, we used a suite of summary statistics to describe the spatial relationships among the data sources. Since the amount of overlap and agreement among data sources can contribute to the amount of uncertainty in parameter coefficients, these summary statistics may help determine whether integration was a useful step in the modeling process. Data quantity ratios were calculated representing the number of grid cells with structured data vs the number with unstructured data, and spatial overlap was calculated as the intersection of the grid cells with each data source compared to the total number of unique sampling locations. Lastly, naïve occurrence ratios compared the number of grid cells with detections out of the total number of cells with surveys for each data source.

Data Checking Step Three: Model Fitting, Evaluation, and Comparison

Statistical modeling of empirical data (Model Fitting)

For each species, we fit single-data source SDMs to the structured survey data (i.e., camera trap surveys) and then fit ISDMs to two or more of the NCWRC data sources described above. Each ISDM always used the camera observations and included one or more of the county-level WRC records until we fit all combinations of the data sources (i.e., up to four data sources in a single ISDM). For model evaluation, we used a weighted objective framework (described below) to determine the best combination of data sources (i.e., "alternatives" in decision analysis terminology).

Modeling true state of occurrence (ecological model)

We modeled the true but unobserved state of occurrence as a Bernoulli random variable, $Z_i \sim Bernoulli(\psi_i)$, where ψ_i represented the probability of occupancy in cell i and we used a log-link function to estimate the linear equation (Miller et al. 2019):

$$\psi_i = 1 - e^{-\lambda_i}$$

$$log(\lambda_i) = \mathbf{X}_i^T \boldsymbol{\beta} + \theta_i$$

where $\mathbf{X}_i$ is a vector of ecological covariates to describe a species' distribution across North Carolina ("Ecological Process" category in Table 9), $\boldsymbol{\beta}$ is the corresponding vector of fixed regression coefficients, and θ_i represents an improper conditional autoregressive (CAR; (Banerjee et al. 2014) random effect. All models, whether single or share, had identical "ecological process" models. The CAR random effect creates spatial dependence among neighboring grid cells. Specifically, for responses $y = (y_i, \ldots, y_N)$ the improper CAR is

$$p(Z_i | Z_{-i}, \tau) \sim N\left(\frac{1}{w_{i+}} \sum\nolimits_{j \in N_i} w_{ij} Z_j, w_{i+}\tau\right)$$

where Z_{-i} represents all elements of Z except Z_i, the neighborhood N_i of grid cell i is the set of all j for which region j is a neighbor of region i, $w_{i+} = \sum_{j \in N_i} w_{ij}$, and the Normal distribution is parameterized in terms of precision τ. Adding a CAR random effect to the linear predictor assumes that the occurrence state in grid cell i is a function of site covariates and the occurrence at neighboring grid cells.

Describing model for each data source (observation models)

We modeled the structured data as observations from a Binomial distribution:

$$y_{1ij} \sim Binomial(n_j, \mu_{ij})$$

$$\mu_{ij} = Z_i * p_{ij}$$

$$logit(p_{ij}) = \mathbf{X}_i^T \boldsymbol{\alpha}$$

where y_{1ij} indicates the number of camera days with a species detection at camera j within grid cell i, n_j is the known number of binomial trials (i.e., no. of days camera j was deployed), p_{ij} is the probability of detecting a species at camera j conditional on its occurrence state in grid cell i, and $\boldsymbol{\alpha}$ is an additional vector of regression coefficients describing effects on the detection process ("Camera trap detection process" category in Table 9).

For all county-level NCWRC records, we modeled the counts as observations from a Poisson distribution

$$y_{2k} \sim Poisson(E_k * \lambda(G2_k))$$

where y_{2k} is the counts in county k, E_k is the hunter and/or trapper effort spent searching for animals, and $G2_k$ includes all grid cells that fall within county k and effectively implements a change of support to account for the spatial misalignment among the data sources (Pacifici et al. 2019). We then used the log link to model the intensity as a function of data source specific covariates ("NCWRC hunter/trapper process" category in Table 9) and the CAR random effect described above (i.e., $\log(\lambda_i) = \mathbf{X}_i^T \boldsymbol{\gamma} + \theta_i$).

All regression coefficients in the linear predictors of the observation and ecological models assume priors $Normal(0, sd = 5)$ and $\tau \sim InvGamma(0.1, 0.1)$ for the CAR distribution precision. All models were fit using the *r-nimble* R package (de Valpine et al. 2017) with three chains of 35,000 iterations, a burn-in period of 15,000 iterations, and a thinning rate of 20 iterations. Model convergence was evaluated by inspecting trace plots using the *mcmcplots* package (Curtis 2018).

Model evaluation and comparison

Model evaluation in ISDMs is a developing area of research and currently little guidance exists (Isaac et al. 2020). We further complicated model comparison by having more than two data sources available to model a species' distribution. That said, we chose to use a structured decision making framework to evaluate and compare models fit for each species (Runge et al. 2020). We framed each parameter of interest (e.g., occupancy uncertainty) as a decision objective, normalized the parameter values to facilitate comparison (i.e., restricted values to be bounded between zero and one), assigned weights to those objectives, and then viewed each data source combination as a decision alternative. The decision alternative with the lowest cumulative normalized weights represents the best decision or optimal combination of data sources.

Because conventional approaches such as information criteria (e.g., AIC, DIC) are not applicable due to ISDMs being fit with different sets of observations, we focused on measures of parameter uncertainty and leave-one-out cross-validation (LOO) of the state variable (i.e., species occurrence). Specifically, we compared finite-sample occupancy probability (i.e., mean occupancy probability / number sampling locations) uncertainty (i.e., standard deviation), ecological process regression coefficient uncertainty, and approximate LOO of the state variable, calculated using the MCMC of the log probabilities of the latent state Z; we compared the LOO information criterion, calculated as -2 × the expected log pointwise predictive density of Z within the *loo* package in R (Vehtari et al. 2020). We assigned weights of 0.25 to the state variable measurements (i.e., occupancy probability uncertainty and LOO) and 0.125 weights to each regression coefficient, emphasizing the state variable of interest but also balancing that with interest in ecological relationships.

Results

Data Checking Step One: Data Compatibility

There was substantial variation in the number of observations considered outliers across species and data sources (Table B1). For example, no observations of bobcat were considered outliers in the hunter harvest data, whereas 284 bobcat observations were removed from the trapper harvest data source (Table B1). Similarly, 41 observations of raccoon were removed from the hunter harvest records but 849 were removed from the trapper harvest records. The magnitude of effect, measured by the change in average CPUE, varied as well and was dependent on the magnitude of the average CPUE. Removing 284 bobcat observations from the trapper harvest data source reduced the average CPUE from 0.042 to 0.017 (147% change), but removing 849 raccoon trapper harvest records reduced the average CPUE from 0.105 to 0.066 (37% change) (Table B1).

When considering whether to remove data with very small sample sizes, we determined that not filtering for minimum sampling effort tended to produce the greatest variation explained among data sources compared to distribution and observation filtering (Table B2). Comparing across the 39 combinations of three species and four datasets, no filtering resulted in the greatest deviance explained for 62% of the models, followed by observation filtering for 29%, and 9% for distribution filtering (Table B2).

We compared the top performing model for each species with and without the outlier removal and effort filtering processes to evaluate the effects of Step One on modeling outcomes (Table B3). For all species, Step One resulted in a greater mean occupancy probability but with comparable measures of uncertainty. (Table B3). Additionally, Step One always resulted in better predictive performance of the latent occurrence state (Table B3). Although the direction of the ecological effects were in agreement, Step One typically resulted differing magnitudes across the regression coefficients, most notably for coyotes (Table B3).

Data Checking Step Two: preliminary data relationship checks

There were consistencies in the preliminary data relationships among the empirical data (Figure 10). For example, we consistently documented a positive relationship between the sum of the weighted, normalized objectives (metric used to compare models) and deviance explained in the regressed relationship between the data sources (Figure 10). There was a positive relationship with Pearson correlation coefficients where the correlation coefficient increased as the model score increased (indicating poorer performance) (Figure 10). In agreement with the empirical findings, this pattern was also apparent in the simulated data except when fit using the correlation model (Figure 11).

Data Checking Step Three: Model Fitting, Evaluation, and Comparison

There was stronger support for integrated models over single-data source models but the degree of support varied across species (Table 10). For example, the single-data source bobcat model performed poorly (12th best out of 13 models), where virtually all integrated models were better than the single-data source model, but the single-data source model was the 6th ranked model out of 12 for coyote and raccoon. For all three species, the best performing model contained observations from the year-round camera surveys and DHOS records (Table 10). Notably, though, the top three raccoon models were each within 0.01 (i.e., sum scores ranged from 0.25 – 0.26) and were all models using year-round cameras and one county-level data source, suggesting nearly equal footing among the value of the county-level data sources in estimates of raccoon distribution. Importantly, these three models resulted in very similar occupancy probabilities across the state (Figure B3).

There was some support for models with three data sources (i.e., cameras integrated with two county-level data sources), where the second-best bobcat model used hunter harvest and trapper harvest records and the second-best coyote model used hunter harvest and DHOS records (Table 10). Though ranking fourth, a similar model of raccoon distribution using year-round cameras, DHOS records, and hunter harvest records was within three points of the top raccoon model (Table 10).

Compared to the most supported integrated model, the single-data source models estimated lower occurrence probabilities for all but raccoon (Figure 12). Additionally, across all species the occupancy probability estimates from the most support integrated model was more precise than the single-data source models (Figure 12). Across the four ecological covariates of interest, the estimated regression coefficients from the single-data source and best integrated model tended to be in agreement for bobcat but less so for coyote and raccoon (Figure 13). For coyote, the regression coefficients from the single-data source model generally overlapped zero indicating little to no effect of the four tested covariates, but the estimates from the best and 2nd best integrated models indicated much more decisive relationships with the covariates (Figure 13). Specifically, the magnitude of landscape composition regression coefficients (e.g., forest cover, human development) tended to be greater compared to the single data source models for all species (Figure 13).

Results from the integrated coyote models showed interesting relationships with landscape configuration metrics that were not apparent when analyzed with a single data source (Figure 13). Specifically, coyote was more affected by landscape homogenization (as measured by patch richness density) and contagion (i.e., contiguous land cover types) - where coyote occurrence was highest in contiguous, homogeneous areas - than was evident in the single-data source models (Figure 13). Additionally, the integrated models for coyote and raccoon revealed much stronger, positive associations with developed and agricultural landscapes in comparison to single-data source models (Figure 13). For bobcat, the ecological relationships found in the best integrated model were mostly in-line with the findings of the single-data source models but overall tended to more precise (Figure 13).

There were varying levels of spatial agreement between the occupancy estimates from the single-data source models and the top integrated model (Figure 14; Figure 15; Figure 16). Specifically, estimated occupancy probabilities from the single-data source coyote model agreed with the top integrated model in 28% of the grid cells, while the bobcat and raccoon models agreed in 44% and 70% of the grid cells, respectively (Figure 14; Figure 15; Figure 16).

Discussion

Data integration in species distribution models offers a promising approach to make the most of all available animal occurrence records but knowing when to integrate data sources is an important and often overlooked first step. Here we outlined a generalizable approach to data checking in ISDMs and illustrated it in an application with mammal observations in North Carolina, USA. In step one of our data checking procedure, filtering for minimum survey effort improved modeling outcomes but this depended on the species and data source characteristics. In step two we showed that Pearson correlation coefficients and deviance explained in regression analysis can indicate the relative support of each data combination, ultimately providing a preliminary check for determining which data sources to integrate *prior* to integration. Finally, in step three, we illustrated an objective approach to determining the optimal data source(s) for species distribution inference using multi-objective optimization. In North Carolina, we observed that integrated models more precisely estimated parameters, elucidated ecological relationships not apparent in single-data source SDMs, and resulted in a more robust estimate of species' distribution across the state.

Results from Step One of the procedure indicated that model predictions of the empirical data typically improved following outlier removal and filtering for minimum sampling effort. This may simply be due to the datasets containing fewer extreme values, thus less variability in the response may have made predictions more accurate. Interestingly, mean finite sample occupancy probabilities were always higher following Step One. This is likely a result of lower detection probabilities following removal of high CPUEs, which ultimately inflated estimates of occupancy probability. When integrating count data, such as DHOS or hunter harvest records, practitioners should consider filtering extreme values as this appears to affect predictive abilities and ecological inference in ISDMs.

With the support of simulations, we documented consistent, positive relationships among Pearson correlation coefficients, deviance explained in regression analysis, and the ranking of a model in Step Two of the procedure. However, the findings were in contradiction with our hypothesis, where we thought the data sources with the greatest correlation and deviance explained would result in the model with the greatest support. This result may be due to intra-grid cell correlation among the observations of each data source. This correlation potentially violates the conditional independence assumption of multivariate or joint likelihood modeling and may have affected parameter accuracy in our models. However, in the strict sense of the term, conditional independence assumes that data sources have no individuals or observations in common, which, a violation in this case, requires that individuals captured on camera were also later harvested or observed by hunters. Although we have no way of testing with the current data, we believe the probability of this occurring is very low unless the population size of the species in the area of interest is very small (none of the species tested are likely rare enough to have populations this small).

Through the use of simulations, Abadi et al., (2010) tested the effects of violating the independence assumption in integrated population models and reported little effect of non-independence on parameter estimates. However, Schaub and Abadi (2011) state that this might not be the case when integrating differing data types as we did in our simulations and empirical analyses. Interestingly, the simulations that fit the 'correlation' model – a framework that does not use a joint likelihood – were inline with our hypothesis where model performance improved as correlation among the data sources increased. We suggest ecologists that are concerned about data source independence consider alternatives to the joint likelihood approach such as the correlation model proposed in Pacifici et al., (2017).

The multi-objective optimization framework in Step Three of the procedure made comparison of many models a relatively straightforward exercise. Although the framework proposed requires practitioners to specify subjective weights to each objective, the approach provided a clear indication of which combination of data sources best met the objectives of the modeling procedure (e.g., minimizing prediction error and parameter uncertainty). Ultimately, integrated models always improved model criticisms over single-data source models, and the integrated models probably provided a more complete view of species distribution in an area. In almost all cases, the single-data source models estimated state-wide occurrence to be lower than the integrated models, and incorporating long-term monitoring results may have corrected for areas where observations of a species were lacking in the primary data source that resulted in lower estimates of occurrence. This was particularly clear in models of coyote distribution where the cameras typically underestimated their occurrence in the eastern portion of the state. This may be due to the land covers sampled in the eastern part of the state, where the camera dataset typically under sampled agricultural land cover which can be an important resource for coyotes (Hinton et al. 2015).

In general, ecological relationships with different land cover types resulted from the integrated models that were not apparent in the single-data source models. For coyotes, the single-data source model typically estimated no effect of each land cover type but integration results indicated strong attraction to developed and agricultural land cover types. The integration findings are more consistent with the literature and may reflect underrepresentation of the eastern agricultural region of NC in the camera dataset (Gehrt et al. 2011, Toomey et al. 2012, Hinton et al. 2015, Cherry et al. 2017, Parsons et al. 2019). Similarly, although the single-data source model indicated a positive effect, the shared models strongly confirmed known habits of raccoon including the use of developed and agricultural lands (Prange et al. 2004, Cove et al. 2012, Kays and Parsons 2014).

Although the results of the integration models herein support the promising future of data integration in ecology (Miller et al. 2019), the findings are primarily representative of scenarios where a spatial mismatch exists in the integrated data sources (Pacifici et al. 2019). There are, of course, many other data type-scenarios (e.g., presence-only observations) and those scenarios may result in different findings; future research should continue to explore data quantity and quality implications in these integration scenarios as these factors seem to be important determinates in integration outcomes (e.g., Simmonds et al. 2020). Additionally, it may be useful to consider species that are especially rare [e.g., long-tailed weasel (*Mustela frenata*)] or abundant [e.g., white-tailed deer (*Odocoileus virginianus*)] as integration recommendations will likely vary. Lastly, the results presented here are most applicable in situations where practitioners have access to multiple data sources that could be integrated. Although open-access data appears to be increasingly common, there may be situations where only two data sources are available and the decision is then whether integration is needed at all; the three-step procedure and structured decision framework presented here should still apply.

Overall, we provided a framework for approaching data integration in species distribution modeling when multiple data sources are available. Data integration provided a more complete picture of the distribution of wildlife in North Carolina, and we highlighted the use of a formal decision-making procedure for deciding among models. We also showed that temporal alignment of data sources and simulated bias in the unstructured data source had little impact on modeling results. This work continues to show the importance of considering multiple data sources in the modeling of wildlife and provides some direction on understanding the effects of data quantity and quality in ISDMs. Data integration will likely soon become the default approach in modeling species distribution and abundance and will surely continue to provide novel insights into wildlife ecology.

Table 6: Overview of available data sources in North Carolina, USA and potential data integration issues

Data Source	Resolution	Measurement	Proportion of Counties with Observations	Proportion of grid cells with Observations	Potential issues
Camera Traps	Point-reference	Detection/Non-detection	0.99	0.16	Low quantity/poor coverage; Clustering around developed areas; preferential sampling
Hunter Harvests	County summary	Count	1	1	Low effort and count for some species

Table 6 (continued)

Trapper Harvests	County summary	Count	1	1	Low effort and count for some species; a given county may have only one observer
Deer Hunter Observations	County summary	Count	1	1	Mismatch between species and hunter temporal activity pattern

Table 7: Simulation settings for three data source characteristics that may influence data integration outcomes.

Characteristic	Setting
Data Quantity	Unstructured: 10, 20, and 40% of grid
	Structured: 20, 40, and 80% of grid
Data Quality	-20, 0, 20 percent change in true count
Spatial Overlap	10, 25, and 50 percent overlap

Table 8: Overview of preliminary data relationship checks tested for ability to indicate integration outcomes in empirical and simulated data.

Preliminary Check	Description	Expected relationship
Correlation	Checks Pearson Correlation between the raw observations of each data source. If spatial misalignment exists, data sources are compared at the coarsest resolution by summing the finer data source.	Data sources with a positive correlation will outperform data sources exhibiting negative correlation. Data sources with a higher positive correlation are expected to outperform lower positive correlations.
Linear Regression	Checks regressed relationship between the two data sources by fitting a binomial generalized linear model. The structured data source is used as the response, represented as the number of successes (i.e., number of days with a detection) and failures (i.e., days without a detection), and the unstructured data as the predictor, represented as catch-per-unit-effort (i.e., the number of counts / effort spent counting). If spatial misalignment exists, data sources are	GLMs with a positive regression coefficient and the greatest deviance explained will outperform models indicating negative relationships and less deviance explained. Models with positive regression coefficients will be differentiated by increased deviance explained, not necessarily larger regression coefficients, and this increased

Table 8 (continued)

	compared at the coarsest resolution by summing the finer data source.	deviance will further improve model criticisms.
Data Quantity Ratio	Compares the number of locations (e.g., grid cells) that data from a given data source. If data quantities are equal, then the ratio = 1.	Models fit with a data quantity ratio (DQR) greater than or equal to one will outperform models with a DQR less than one.
Spatial Overlap	Checks the number of locations (e.g., grid cells) where observations exist from both data sources. Calculated by counting the number of locations with co-occurring observations compared to the total number of sampling locations.	Holding the total number of locations with observations constant, increasing spatial overlap will improve model criticisms.
Naïve Occupancy Ratio	Compares the naïve occurrence rate (i.e., the ratio of the number of locations with positive observations to the total number of locations with observations). If naïve occupancy calculations are equal among the data sources, then the ratio = 1.	Naïve occupancy ratios (NOR) >= 0.5 and <= 2 will outperform models with an NOR outside of that range.

Table 9: Ecological and observation covariates used in empirical integrated species distribution modeling.

Category	Covariate	Description
Ecological Process	Forest Cover†	Proportion of forest cover in 5 km x 5km grid cell
Ecological Process	Agricultural Cover‡	Proportion of cultivated land in 5 km x 5km grid cell
Ecological Process	Housing Density	Mean housing density (houses/km) in 5km x 5km grid cell
Ecological Process	Contagion Index	The degree of clumping among land cove types in 5km x 5km grid cell, or the propensity for a raster pixel of a given land cover class to be neighboring a different land cover class
Ecological Process	Patch Richness Density	Number of land cover types per 100 ha in 5km x 5km grid cell
Camera trap detection process	Distance to Road	Minimum Euclidean distance between a road and a camera trap
Camera trap detection process	Distance to Building Structure	Minimum Euclidean distance between a building structure and a camera trap
Camera trap detection process	Distance to Water Source	Minimum Euclidean distance between a water source (e.g., stream, lake) and a camera trap
Camera trap detection process	Detection Distance	Maximum field-measured detection distance by a camera trap on day of deployment

Table 9 (continued).

Camera trap survey effort	Trap Nights	Length (days) of camera trap deployment used to represent effort of camera trapping.
NCWRC* hunter/trapper process	Huntable land	Proportion of huntable land in a 5 km x 5 km grid cell. Huntable land excludes areas within a federal or state park, municipal boundary, water body, or where human density is greater than 1 person per 2 acres.

†*bobcat only;* ‡*coyote and raccoon;* **North Carolina Wildlife Resources Commission*

Table 10: Consequence table for determining the species-specific combination of data sources that minimizes parameter uncertainty and prediction error. For each species and each combination of data sources (including whether cameras were deployed year-round versus during harvest periods only), we report occupancy probability uncertainty, ecological relationship uncertainty, predictive performance (LOO), the sum of the scores, and the ranking based lowest-highest. Each objective was bounded to 0-1 for comparison across objectives, summed, then ranked.

Species	Data Source 1	Data Source 2	Data Source 3	Camera Period	Occupancy SD	Contagion SD	PRD SD	Forest-Ag SD	Developed SD	LOO	Sum	Rank
	Camera	Observation		Year	0.06	0	0	0	0	0.22	0.28	1
	Camera	Harvest	Trapper	Year	0	0.04	0.02	0.03	0.02	0.19	0.3	2
	Camera	Harvest		Year	0.02	0.05	0.03	0.03	0.01	0.18	0.33	3
	Camera	Harvest	Observation	Year	0.07	0.06	0.02	0.03	0.02	0.2	0.4	4
Bobcat	Camera	Trapper		Year	0.07	0.07	0.05	0.06	0.04	0.13	0.42	5
	Camera	Observation	Trapper	Winter	0.08	0.06	0.01	0.03	0.07	0.22	0.47	6
	Camera	Harvest		Winter	0.05	0.08	0.04	0.05	0.06	0.23	0.51	7
	Camera	Observation	Trapper	Year	0.09	0.09	0.12	0.12	0.11	0	0.54	8
	Camera	Observation		Winter	0.25	0.05	0.01	0.02	0.05	0.22	0.59	9

Table 10 (continued).

	Camera	Trapper		Winter	0.15	0.07	0.03	0.04	0.09	0.21	0.59	10
	Camera	Harvest	Observation	Winter	0.12	0.1	0.05	0.05	0.12	0.19	0.64	11
	Camera			Year	0.24	0.04	0.03	0.1	0.07	0.18	0.65	12
	Camera	Harvest	Trapper	Winter	0.04	0.12	0.6	0.05	0.12	0.25	0.65	13
	Camera	Observation		Year	0	0.03	0.03	0.03	0.06	0	0.14	1
	Camera	Harvest	Observation	Year	0.01	0.03	0.04	0.03	0.08	0.01	0.2	2
	Camera	Harvest		Year	0.06	0.02	0.02	0.02	0.05	0.08	0.25	3
	Camera	Trapper		Year	0.09	0.01	0.02	0.01	0.02	0.19	0.34	4
	Camera	Harvest	Trapper	Year	0.09	0	0	0	0	0.25	0.34	5
	Camera			Year	0.11	0	0.01	0	0	0.23	0.36	6
Coyote	Camera	Observation		Winter	0.1	0.05	0.05	0.04	0.07	0.05	0.36	7
	Camera	Harvest	Observation	Winter	0.14	0.03	0.02	0.02	0.07	0.17	0.45	8
	Camera	Harvest	Trapper	Winter	0.06	0.1	0.11	0.05	0.12	0.03	0.47	9
	Camera	Observation	Trapper	Winter	0.2	0.02	0.02	0.02	0.03	0.24	0.52	10
	Camera	Observation	Trapper	Year	0.05	0.12	0.12	0.12	0.12	0	0.55	11
	Camera	Harvest		Winter	0.25	0.02	0.02	0.02	0.06	0.2	0.57	12

Table 10 (continued).

	Camera	Trapper		Winter	0.21	0.03	0.05	0.03	0.09	0.117	0.58	13
	Camera	Observation		Year	0.06	0.02	0.02	0.03	0.05	0.07	0.25	1
	Camera	Trapper		Year	0.06	0.02	0.02	0.01	0.01	0.13	0.25	2
	Camera	Harvest		Year	0.06	0.01	0.01	0.01	0.01	0.17	0.26	3
	Camera	Harvest	Observation	Year	0.06	0.05	0.04	0.03	0.02	0.08	0.28	4
	Camera	Observation	Trapper	Year	0	0.1	0.08	0.06	0.05	0	0.28	5
	Camera			Year	0.09	0	0	0	0	0.23	0.33	6
Raccoon	Camera	Observation		Winter	0.15	0.04	0.03	0.01	0.02	0.18	0.43	7
	Camera	Observation	Trapper	Winter	0.18	0.01	0.01	0	0	0.25	0.45	8
	Camera	Harvest	Trapper	Year	0.08	0.1	0.06	0.09	0.05	0.07	0.45	9
	Camera	Trapper		Winter	0.17	0.05	0.04	0.02	0.01	0.17	0.47	10
	Camera	Harvest		Winter	0.2	0.04	0.04	0.02	0.01	0.21	0.51	11
	Camera	Harvest	Trapper	Winter	0.25	0.12	0.09	0.09	0.08	0.11	0.75	12
	Camera	Harvest	Observation	Winter	0.22	0.08	0.12	0.12	0.12	0.16	0.84	13

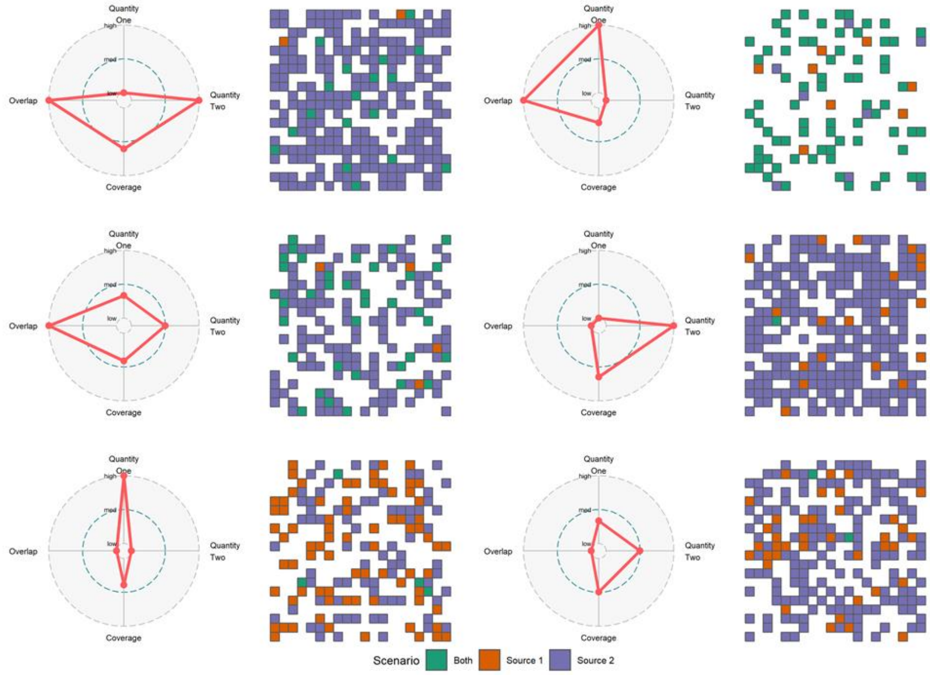
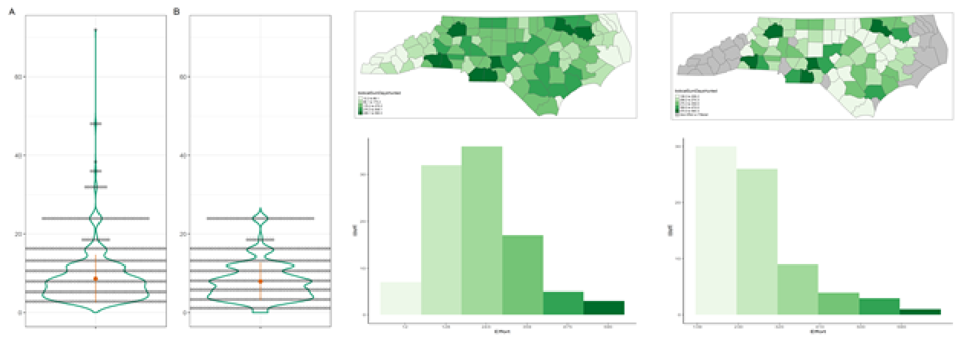
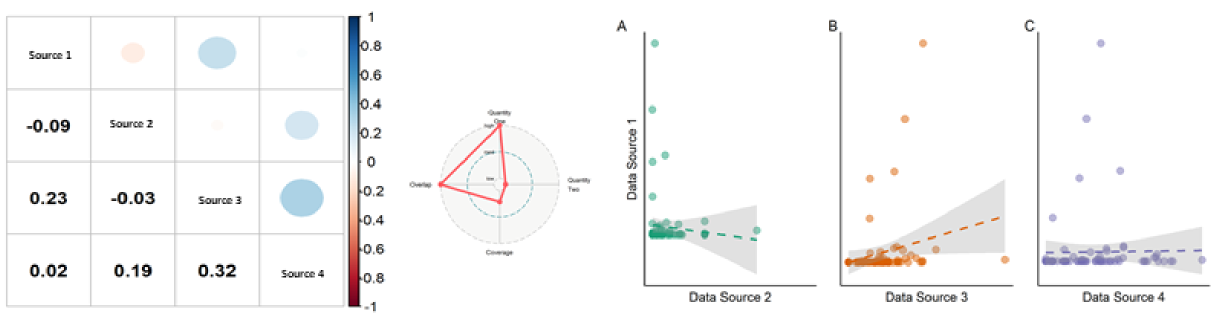


Figure 6: Different data scenarios in integrated species distribution models showing low, medium, and high amounts of each data source with varying spatial overlap (i.e., areas where both data sources occur). Included are radial plots to show a snapshot of the data scenario and then the resulting distribution of data sources and overlap among data sources; orange-colored cells represent areas where only structured data are available, purple indicates areas with only unstructured data, and green cells represent cells with information coming from both sources.

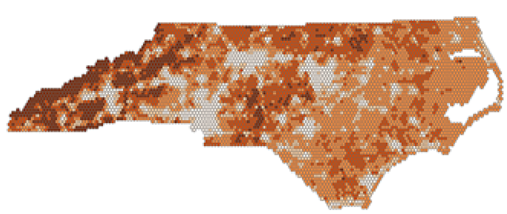
Data Compatibility



Preliminary Data Checks



Model Fitting and Evaluation



	Objective	A	B	C	
Alternative	Weights	0.65	0.2	0.15	Score
No integration		0.2	1	0	0.33
Shared Integration		0.1	0.2	0.5	0.18
Correlatation Integration		0	0	0.3	0.045
Covariate Integration		1	0.5	0.45	0.8175

Figure 7: Three primary steps to integrated species distribution models. The initial Data Compatibility step handles issues of data source-specific outliers and checks for sufficient effort in data sources. Preliminary Data Checks provide practitioners with suggestions for incorporating unstructured information and the final step, Model Fitting and Evaluation, illustrates information gained in the modeling procedure.

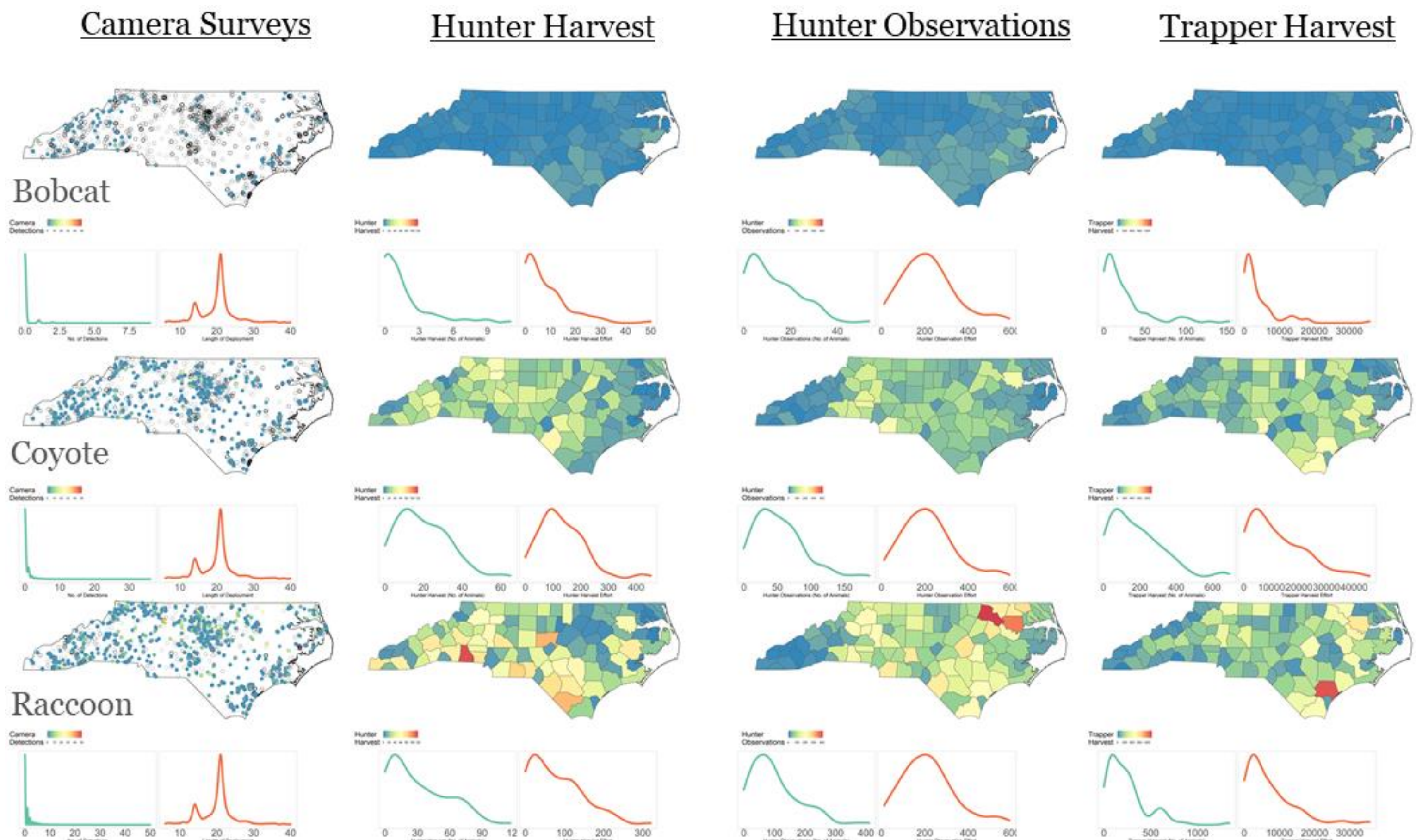


Figure 8: Overview of four data sources available for integration in North Carolina, USA. Each column has species-specific counts of animals (maps and green density plots) and effort made searching for animal (orange density plots). Maps have constant legend scales for ease of comparison while density plots have species-specific x-axes. The hollow black circles in the camera map represent camera deployments with no detections of the species. Each row represents each species: top – bobcat (Lynx rufus), middle – coyote (Canis latrans), and bottom – raccoon (Procyon lotor).

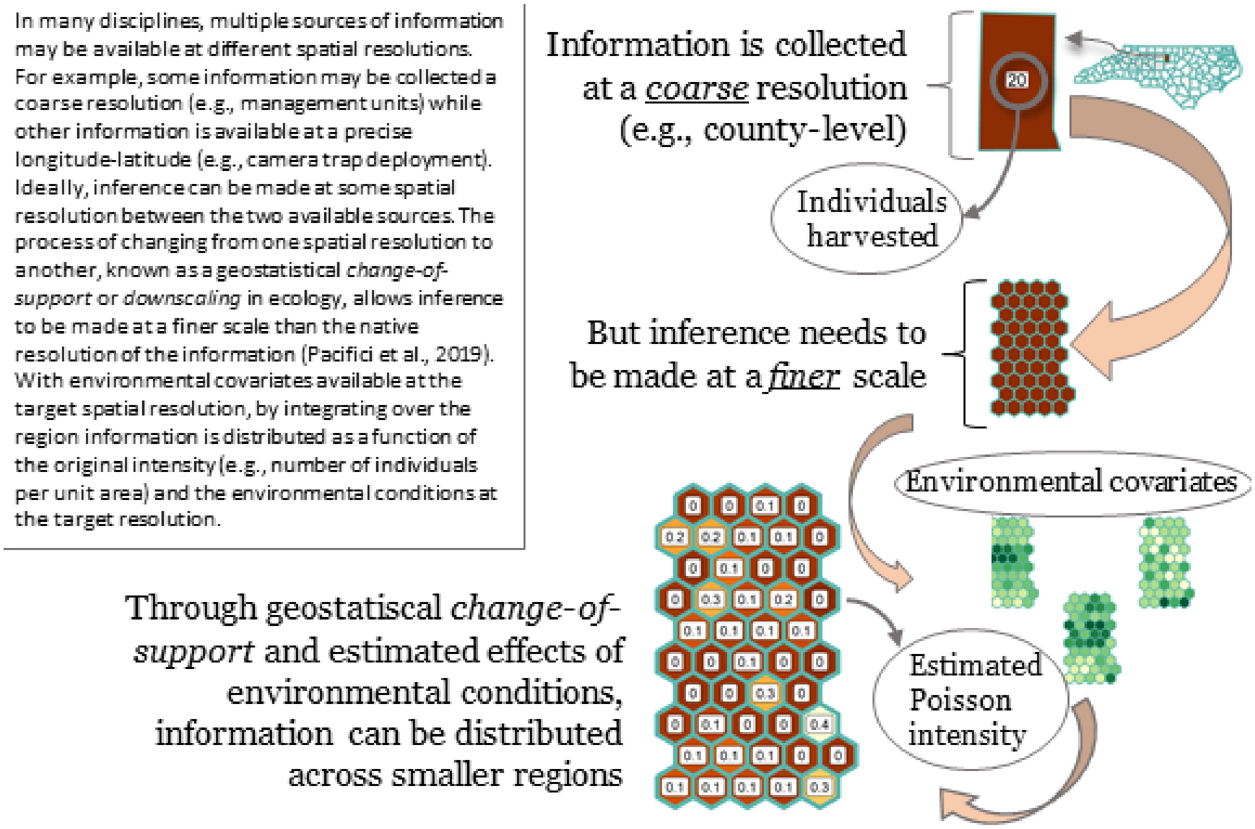


Figure 9: Overview of change-of-support in integrated species distribution models

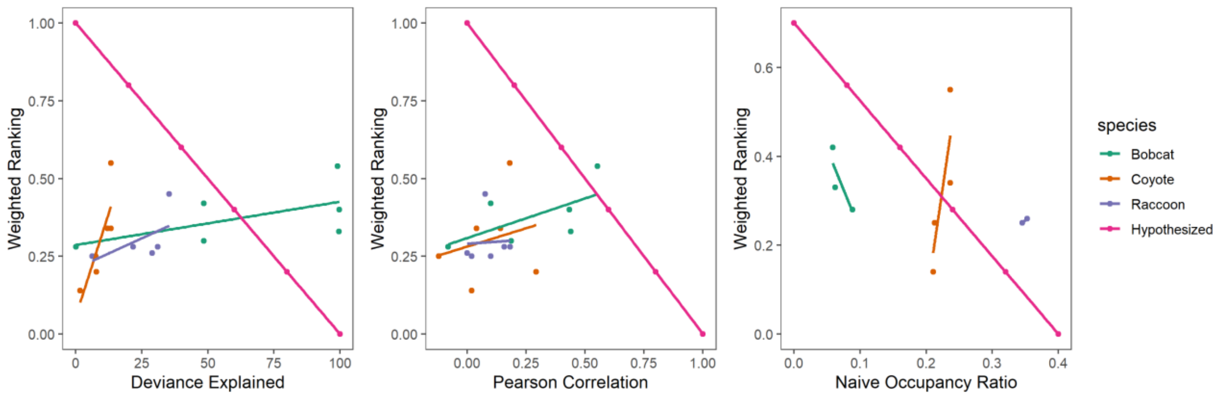


Figure 10: Preliminary data relationship checks from the empirical North Carolina data. In all figures, the normalized sum of the weights is on the y-axis while each panel represents (1) the deviance explained in the regressed relationship between the structured and unstructured data sources, (2) Pearson correlation among data sources, and (3) the naïve occupancy ratio. Colors indicate different species, including the hypothesized relationships.

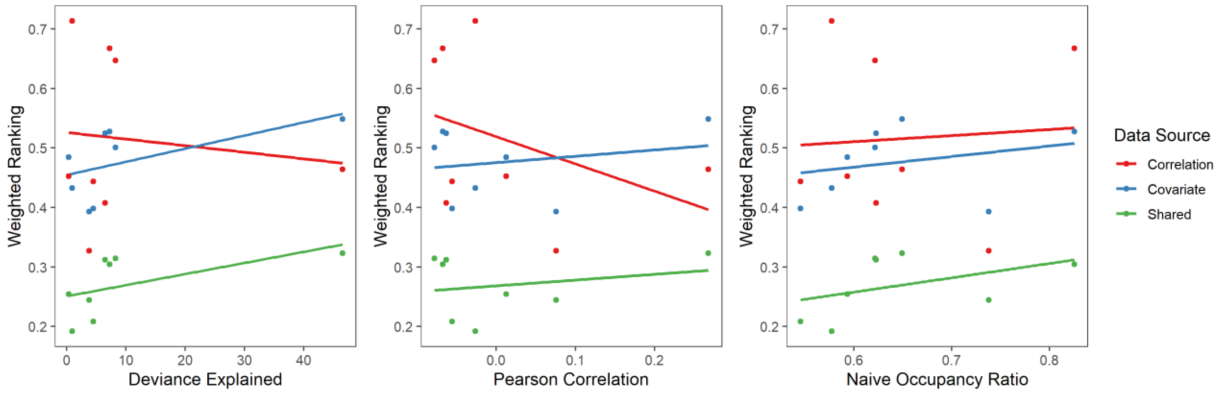


Figure 11: Preliminary data relationship checks from the simulated data using 10% percent structured data, 60% unstructured and unbiased data, and 10% overlap. In all figures, the normalized sum of the weights is on the y-axis while each panel represents (1) the deviance explained in the regressed relationship between the structured and unstructured data sources, (2) Pearson correlation among data sources, and (3) the naïve occupancy ratio. Colors represent the different modeling approaches.

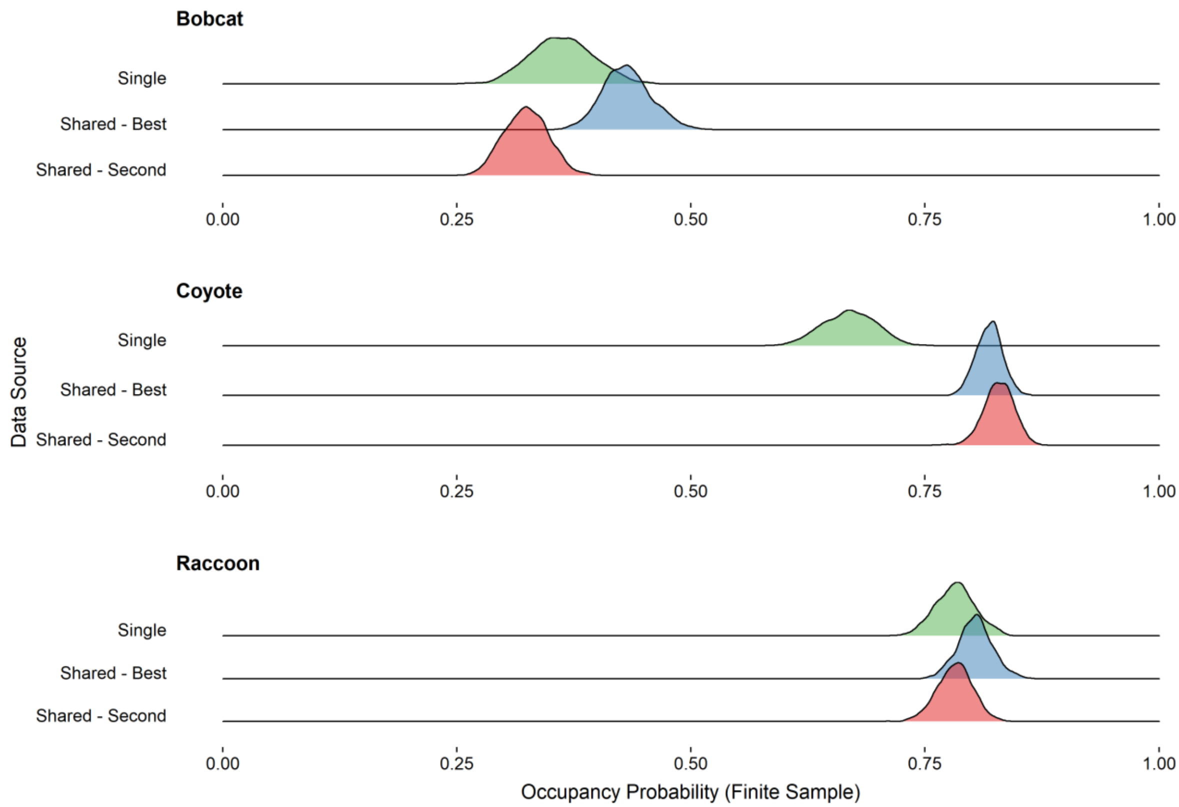


Figure 12: Comparison of the posterior distributions of finite-sample occupancy estimates among single-data source, the best performing intergrated model, and the second best integrated model across three species in North Carolina, USA.

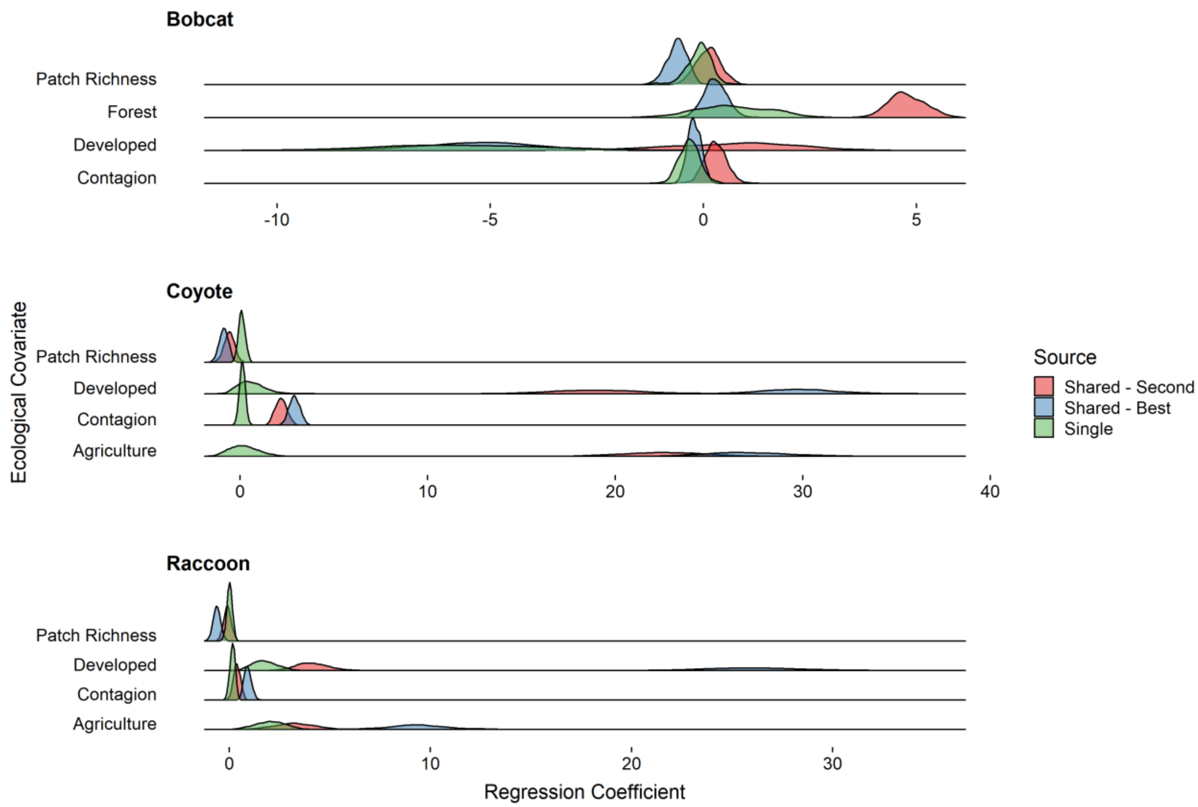


Figure 13: Comparison of the posterior distributions of the ecological regression coefficients among single-data source, the best performing intergrated model, and the second best integrated model across three species in North Carolina, USA.

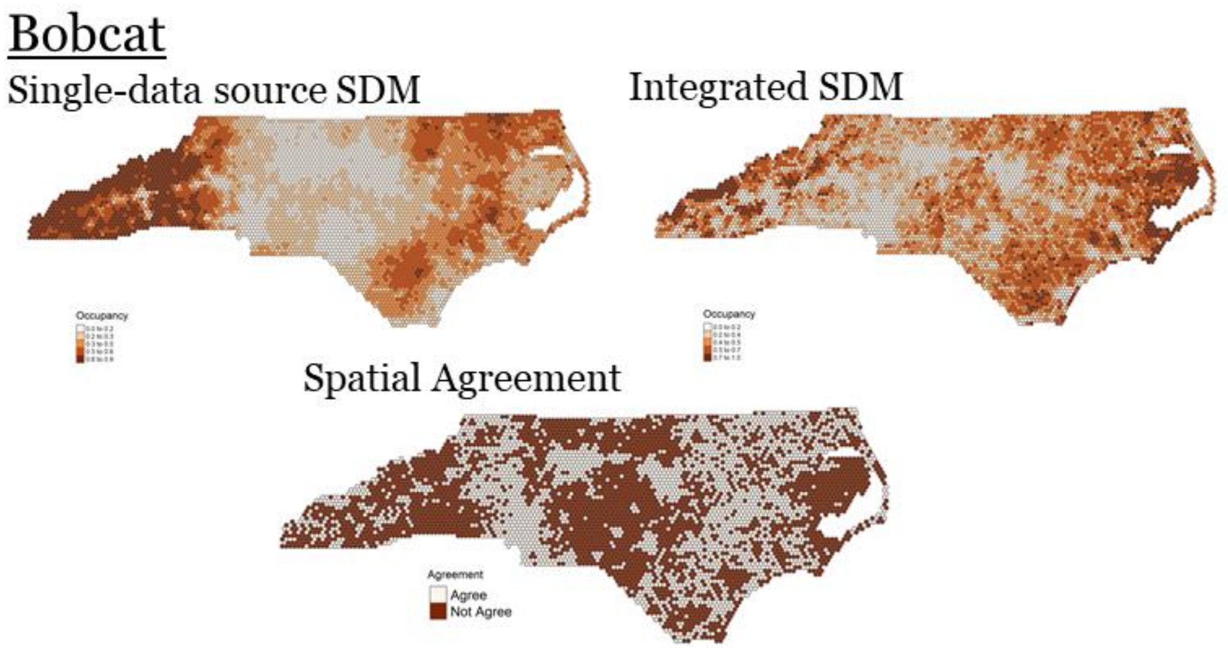


Figure 14: Estimated occupancy probability for bobcat (*Lynx rufus*) in North Carolina, USA during 2015 - 2019. *Single-data source SDM* presents estimates from the camera-only model while *Integrated SDM* shows the estimated occupancy probability from the best performing integrated model. The *Spatial Agreement* plot depicts the grid cells where the single-data source model and integrated model agreed (i.e., the estimated occupancy probabilities where comparable).

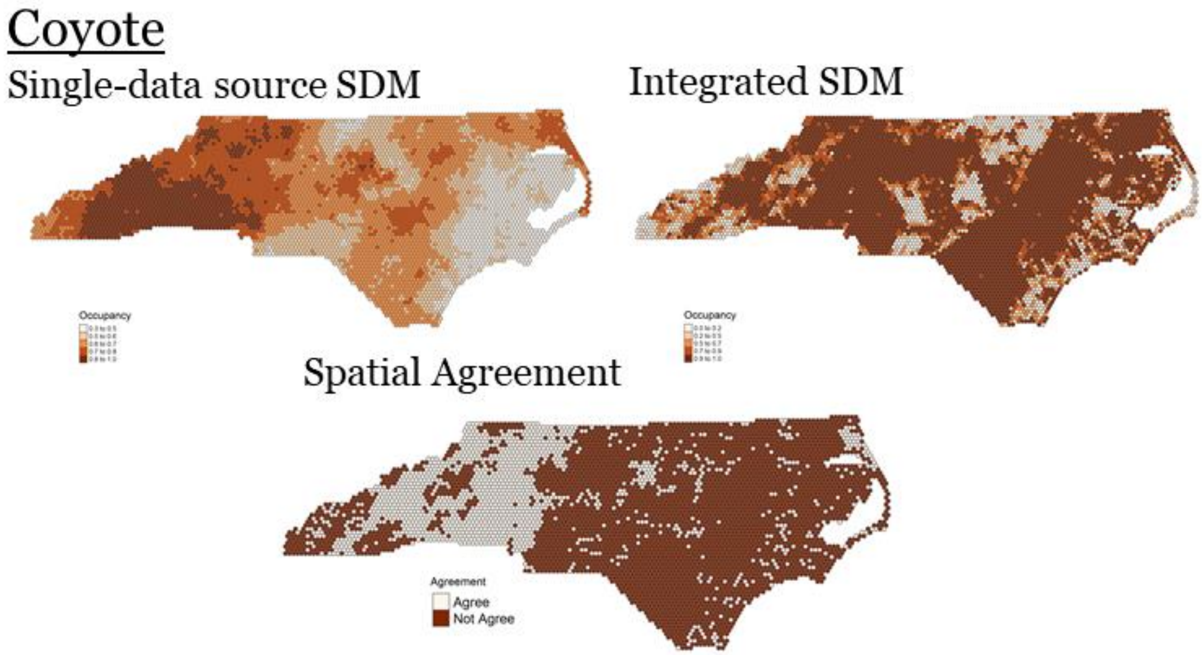


Figure 15: Estimated occupancy probability for coyote (*Canis latrans*) in North Carolina, USA during 2015 - 2019. *Single-data source SDM* presents estimates from the camera-only model while *Integrated SDM* shows the estimated occupancy probability from the best performing integrated model. The *Spatial Agreement* plot depicts the grid cells where the single-data source model and integrated model agreed (i.e., the estimated occupancy probabilities where comparable).

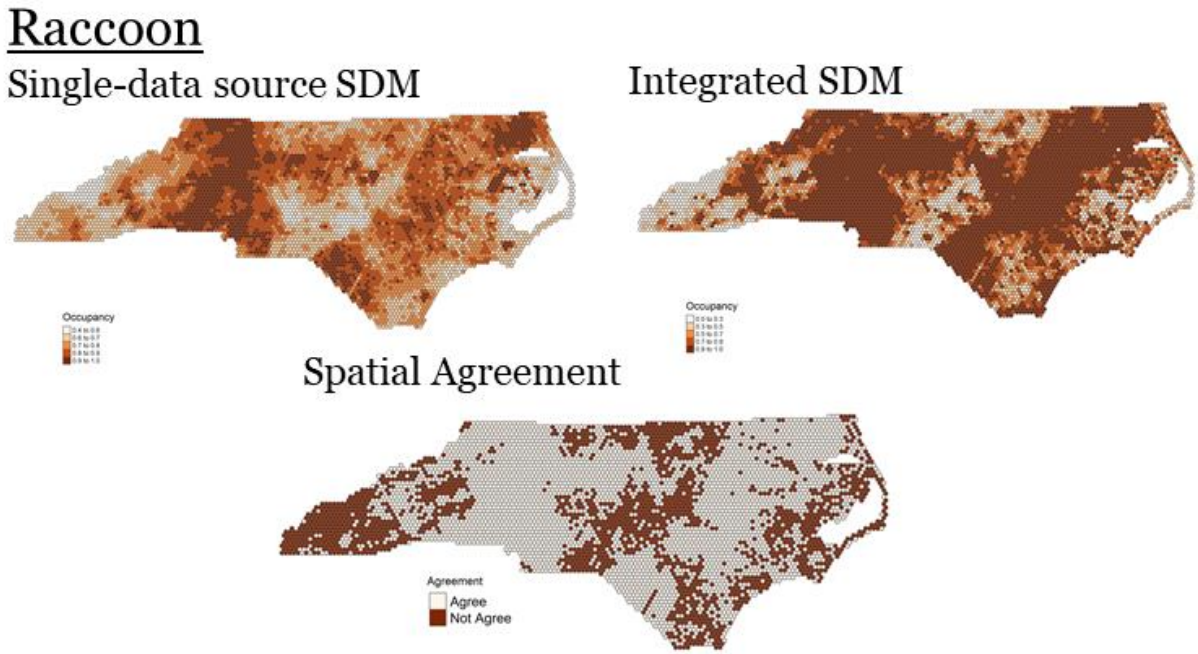


Figure 16: Estimated occupancy probability for raccoon (*Procyon lotor*) in North Carolina, USA during 2015 - 2019. *Single-data source SDM* presents estimates from the camera-only model while *Integrated SDM* shows the estimated occupancy probability from the best performing integrated model. The *Spatial Agreement* plot depicts the grid cells where the single-data source model and integrated model agreed (i.e., the estimated occupancy probabilities where comparable).

CHAPTER 4: NICHE CONSERVATISM OR LOCAL ADAPATION: DETERMINING DRIVERS OF SPATIALLY VARYING SPECIES-ENVIRONMENT RELATONSHIPS IN A WIDE-RANGING GENERALIST SPECIES

Introduction

Understanding the ecological determinants of species distributions and why they change over time is a core area of ecological research. Though the study of species distributions initially provided basic ecological knowledge about a species' relationship with the abiotic and biotic environment, recently the information has been applied to conservation planning, resource management, impact and risk assessments, and understanding the implications of climate change on global biodiversity (Loiselle et al. 2003, Elith and Leathwick 2009, Austin and Van Niel 2011, Bled et al. 2011). The increasing number of applications has in part been facilitated by larger, open-access species occurrence databases (e.g., iNaturalist, Global Biodiversity Information Facility) coupled with the increasing availability of higher-resolution environmental predictors (e.g., remote sensing products) (Kays et al. 2020b). This new era of studying species distributions has the potential to provide nuanced insight into how species use the environment across their entire geographic range (Maldonado et al. 2015).

Historically, studies of range-wide species distributions have assumed constant spatial relationships with the environment – that a species responds to a given environmental factor in the same way across their range (Koenig 2001, Osborne et al. 2007b, Miller 2012b). However, a growing body of literature has recently challenged that assertion, although the causes of within-species variation are not known and are often contested (Smith et al. 2019c). For example, McNew et al. (2013) reported that greater prairie-chicken (*Tympanuchus cupido*) females varied spatially in their nest site preferences, presumably due to individuals adapting to varying microsite conditions and quality across their three study regions. Whereas some have attributed this intraspecific variation to genetic differences (Hällfors et al. 2016, Marcer et al. 2016), others have reported ecological context as an alternative driver of within-species variation (Smith et al. 2019a, Pease et al. In Review).

If intraspecific variation in species-environment relationships is the result of genetic adaptation to local environments, then it should be strongly predicted by the evolutionary history of the Pleistocene refugia lineages (herein, 'lineage') that make up different parts of their modern range (Hällfors et al. 2016, Smith et al. 2019c). An individual lineage, however, can inhabit a wide range of environmental conditions that can contribute to local adaptation within the lineage, and finer scale delineations may be more appropriate. For example, although directly related to lineages, the varied relationships may be better described by *subspecies* boundaries, which are nested within the larger lineage subdivision.

Alternatively, there is growing support for the importance of ecological context as an explanation to varied species-environment relationships (Sexton et al. 2009, Smith et al. 2019a, Pease et al. In Review). Ecological context – the "landscape perspective" - describes the environmental characteristics of a region that often span beyond genetic subdivisions, and includes physiography, vegetation communities, and hydrology, and other abiotic and biotic environmental conditions. Although ecological context can drive speciation, ecological context transitions do not typically occur at the genetic subdivisions but rather span beyond these species-specific delineations. For species where ecological context has a greater impact than local adaptation on intraspecific variation, then boundaries such as ecoregion delineations or differing, competitive ecological communities (e.g., sympatric species) may better describe how and where a given species is using the environment differently (Omernik 1987).

Support for lineage or subspecies as an explanation for varied ecological relationships between a species and the environment suggests niche conservatism – the tendency for a species to retain ecological characteristics of their fundamental niche (Peterson et al. 1999, Wiens et al. 2010, Peterson 2011) – when the species is responding to environmental conditions in ways consistent with their known ecological traits. Greater explanatory power in ecological context, however, indicates synchronized responses that span the known the genetic boundaries of the species – where individuals or subspecies with different genetic makeup are responding similarly across space – suggesting local adaptation to environmental conditions. Identifying this distinction can provide insight into how a species may respond to continued environmental change anticipated throughout the century (Lewis and Maslin 2015). Species exhibiting high niche conservatism may experience shrinking niche space under significant environmental change, ultimately leading to declines in the species distribution and abundance, whereas patterns of local adaptation may be an indication of the species' ability to persist under altered environmental conditions.

Here we explore how American black bear (*Ursus americanus*; hereafter, black bear) – a wide-ranging, large-bodied omnivorous species – varies in its ecological relationships across North America. Black bears are an ideal species for exploring space-varying effects for a number of reasons. Foremost, the species is a generalist with a large geographic range, and these characteristics suggest a higher potential of local adaptation and plasticity. Secondly, the genetic lineage boundaries of black bear are well known, allowing us to test evolutionary history as a driver of non-constant relationships (Pelton 2003, Puckett et al. 2015). Finally, across their large range, black bears inhabit multiple distinct ecoregions, with varying levels of competition [e.g., brown bear (*Ursus arctos*)] and human disturbance, providing a means for exploring ecological context as a driver of local adaptation. Although black bear has been the subject of hundreds of local-level ecological studies, they have not been synergized at a range-wide level, and it is unclear whether black bear-environment relationships are constant across their range, and if not, which niche dimensions vary and what the primary driver for this within-species variation may be.

To evaluate the variability of black bear ecology across their range, we first tested for evidence of spatial non-stationarity – the tendency for changes in the direction and/or magnitude of relationships of bear distribution to land cover, human densities, and climate. We evaluated non-stationarity by creating hundreds of local spatial models and comparing these findings to those from a range-wide model of black bear observations (Fink et al. 2010, Lloyd 2010b). We then used the results of the local spatial models to test four hypotheses about drivers of ecological nonstationary in black bear across their range by comparing the spatial pattern of nonstationarity to the boundaries of: (1) lineage, (2) subspecies, (3) ecoregions, and (4) patterns of competing species overlap (Figure 17). We expected that, across all environmental conditions (i.e., land cover, human pressure, and climatic conditions), black bear would have the greatest response consistency within ecoregions compared to the other delineations tested. Our expectation stemmed from the rather large size of black bear lineages, the varied environmental conditions within each lineage, and that the environmental gradients within each ecoregion extend beyond the subspecies boundaries. This expectation supports local adaptation as the driver of non-stationarity rather than niche conservatism. To test these questions and hypotheses, we built models using a publicly available database of opportunistic observations of black bear across North America from the iNaturalist platform and test model results with an independent national-scale camera trap dataset, SNAPSHOT USA (Cove et al. 2021).

Methods

Animal Observations

We queried the iNaturalist database for observations of black bear across North America. iNaturalist is an online social network for sharing biodiversity information to help others learn about nature ("iNaturalist.org" 2020). Observations are submitted to the database by iNaturalist users and the iNaturalist community identifies the observations to the highest taxonomic resolution possible (e.g., species, subspecies). iNaturalist users can submit observations across all taxa, anywhere on earth, with no required sampling protocol; there are, however, some synchronized efforts at local scales (e.g., "biobiltz"). We queried the iNaturalist Application Programming Interface (API) for all research-grade observations (i.e., observations that include a photograph and have confirmed and agreed identification by more than 2 users) of the class Mammalia within North America (specific bounding box: -52 degrees northeast longitude, -170 degrees southwest longitude, 6 degrees southwest latitude, 75.5 degrees northeast latitude). We removed observations that were classified as scat, tracks, or signs, or associated with the North American Tracking Database Project on iNaturalist because they could be less that distinguishable than photographs of the animals themselves (Lonsinger et al. 2015). We also filtered observations that had coordinates obscured (uncertainty > 20 km) due to user preference or iNaturalist's protocol of obscuring the location of threatened or endangered species. From this North American mammal database, we then selected all observations of *Ursus americanus* and further refined to include observations that only occurred during our target time frame (01-Jan-2015 – 31-Dec-2019). By using a fuzzy search function (Wickham 2019), our filtering also included observations that were assigned a subspecies name (e.g., "*Ursus americanus altifrontalis"*) but we did not differentiate among subspecies in any analysis.

Study Area

We made inference across black bear's current range (IUCN ; (Garshelis et al. 2016); although only currently occupying about 65% of their historic range, black bear's range still spans Canada, Mexico, and the United States but in sometimes isolated pockets (Garshelis et al. (2016);Figure S1). We used a hexagonal grid across North America to summarize all available animal observations and environmental information. We set a target cell resolution of 50 sq km over an area of 2.1 square kilometers, resulting in 502,961 grid cells (Figure C1). We chose 50 sq km to reflect typical home range of black bears (Pelton 2003).

Ecological Covariates

We gathered environmental factors known to influence black bear distribution and abundance (Feldhamer et al. 2003, Pelton 2003, Clark et al. 2020, Penteriani and Melletti 2020), including land cover, climate, and human densities (Table 11). Black bear distribution is typically linked to areas of contiguous forest cover, but the forest types used are dependent on the geographic location (Pelton 2003). To capture this variable use of forest types, we used the Moderate Resolution Imaging Spectroradiometer (MODIS) Land Cover Type (MCD12Q1) product from 2019 to differentiate forest types, where we included the proportion of evergreen needleleaf (i.e., conifer forests), deciduous broadleaf, and mixed forest types (Table 11). We also used the MODIS product to describe the proportion of cropland available as this is a dominant land cover type across much of the United States and black bear opportunistically forage in crop fields (Jones and Pelton 2003, Ditmer et al. 2018). We used the Gridded Population of the World (GPW; Version 4) to calculate mean human population density per grid cell (Center for International Earth Science Information Network - CIESIN - Columbia University 2018). Although other products exist (e.g., human footprint index), we chose the GPW product because we wanted to isolate the effect of humans and separately account for factors often incorporated into indices such as the human footprint index (e.g., cropland). Protected areas can be an important designation for a large, wide-ranging species, so we accounted for whether a given grid cell contained protected areas using information from the World Database of Protected Areas (UNEP-WCMC and IUCN 2020). Here, "protected areas" includes a range of designations including national forests, national parks, wildlife management areas, and wildlife sanctuaries with varying regulatory practices (e.g., hunting vs no hunting). For climate variables, we used the Bioclimatic variables of annual average temperature and annual average precipitation from the WorldClim database (Fick and Hijmans 2017). We selected these climatic variables as they reflected variation in climate conditions across the continent but also are correlated with productivity (e.g., gross primary productivity). We considered including additional climatic variables to describe nuanced differences across space but found other WorldClim variables to be highly correlated with the existing set of predictors; this was also true for topographic characteristics such as elevation.

Opportunistic Reporting

Opportunistic observations of biodiversity, such as those reported in iNaturalist, may provide valuable insights into ecological relationships but the data typically prove challenging for unbiased statistical inference (Yackulic et al. 2013, Bird et al. 2014, Isaac et al. 2014, Clare et al. 2019). Observations are often clustered and positively associated with human population density and the lack of standardized sampling protocols presents complications when evaluating survey search effort. Although human population density and wild spaces such as National Parks are typically good proxies to account for the inherent spatial sampling biases in opportunistic observations, we chose to the use the iNaturalist database itself as a direct measure of iNaturalist user behavior to evaluate the tendency for a species to be reported. We used the total number of iNaturalist mammal observations in each grid cell to reflect the relative search effort in that geographic space (Table 11; Kéry et al. 2010). By using this measure of effort we assume that the probability of reporting a black bear was related to the general probability of reporting *any* mammal species in the area.

Model Description

There is a long history of fitting species distribution models using presence-only observations, which has become increasingly common with widespread availability of biodiversity occurrence records (Elith and Leathwick 2009, Elith et al. 2011, Guillera-Arroita et al. 2015, Kays et al. 2020b). Typically, presence-only observations are modeled by generating so-called "background points" - locations where an observation did not occur – and then using logistic regression (1 for locations with observations, 0 otherwise) to make inference about ecological relationships (Renner and Warton 2013, Warton et al. 2013, Renner et al. 2015). This approach is conceptually equivalent to a resource selection function under a used-available framework, where iNaturalist observations are considered "used" locations and the unused grid cells become the "available" background environmental conditions (Matthiopoulos et al. 2020). Research suggests, however, that this typical approach of logistic regression may erroneously estimate model parameters and alternative approaches, such as infinitely weighted logistic regression (IWLG), are recommended (Hastie and Fithian 2013, Renner et al. 2015). Recently, however, Renner et al. (2015) suggested the use of downweighted Poisson regression (DWPR) as a modification to IWLG because of biologically unrealistic intercept estimates often found using IWLG. Downweighted Poisson regression involves setting a weight of some small value (e.g., 1e-06) at presence locations while background points are assigned a weight equal to the area of the study region, A, (in our case, 2.1 million sq km) divided by the number of background points (here, the grid cells that did not contain animal observations). Given this, we consider black bear observations, y_s, as an inhomogeneous Poisson point-process (IPP) where the predicted response (the Poisson intensity at grid cell s, λ_s) is scaled to represent the expected number of black bear reportings per-unit area (Renner et al. 2015). We fit IPP models using the *glm* function in R version 4.0.3 (R Core Team 2020b) with Poisson distributed errors and a log link. The linear predictor of the model included all eight ecological covariates listed in Table 11. We set weights, w_s, equal to 10^{-6} at presence points and (A/*no. of background points) at* non-observation grid cells and used $z_s = y_s/w_s$ as the response. We additionally included an offset of the log of the total number of iNaturalist mammal observations, which represents the search effort, in a given grid cell. Variation inflation factors were checked to be below 5 using the R package *car* in program R (v 4.0.3) (Fox and Iisberg 2019). We used the single set of regression coefficients estimated using this model, fit to all black bear observations simultaneously, to serve as the baseline, "global" model to evaluate non-stationarity.

Non-stationarity in black bear-environment relationships

To explore non-stationarity in black bear ecological relationships we used an ensemble modeling procedure that used local spatial models within user-defined "spatial support windows" to parameterize local distribution models. A "spatial support window" is a user-defined geographically restricted sampling area (i.e., a square smaller than the total area of interest) where only observations and background points within each window are used to estimate regression coefficients. Estimates from each spatial support window were then averaged to effectively "scale up" and create an ensemble prediction at each grid cell across the entire landscape (Fink et al. 2010). We required a minimum number of 100 black bear observations in a given spatial support window for that window to be included in the ensemble model. This resulted in 1,239 equal-sized, 10.5 by 7.5 degree spatial windows with at least 100 black bear observations (range = 100 – 1,617; mean = 512) (Figure 18). We determined the spatial window sizes to balance the geographic coverage of current black bear range with iNaturalist observation density. We then fit an IPP identical to the global model, with the same linear predictor that used the environmental covariates described in Table 11, but the black bear observations and background points were those only found in each window. That is, the ensemble and global models were identical in their specification (i.e., same use ecological predictors, response, weights, and offset), but differed in their extent. We compared absolute value of residual errors to quantify performance between the global and ensemble models.

Stationarity Index

Following parameter estimation of the ensemble models, we calculated a stationary index (SI) for each ecological covariate to evaluate whether evidence of non-stationarity existed and, if so, the magnitude of the variation. In the ensemble modeling approach, each ecological covariate is estimated within each spatial window (n = 1,239), so an interquartile range of the 1,239 parameter estimates can be calculated; this represents the range or magnitude of variation in the parameter estimates across space. The SI is calculated by dividing the interquartile range of a given ecological covariate by two times the standard error of the same ecological parameter estimated in the global model. Index values greater than one indicate evidence of non-stationarity, with higher values suggesting more spatial variation in a given ecological relationship (Osborne et al. 2007b).

Evaluating drivers of non-stationarity

We used each ecological covariate with evidence of non-stationarity (based on SI values; SI value > 1) to test the four hypotheses (ecoregions, lineages, subspecies, or competing species) concerning the spatial pattern of non-stationarity (Figure 17). We did this for each covariate because the drivers of non-stationarity may be different across land use and climatic responses. For the ecoregion hypothesis, we used the EPA Level we ecoregion delineations which are based on similar vegetation, climate, hydrology, land use, geology, landforms, and soils (Omernik 1987, Commission for Environmental Cooperation 1997, Omernik and Griffith 2014). Lineages were delineated following phylogeographic analyses by Puckett et al. (2015) and subspecies delineations described by Larivière (2001). The competition hypothesis describes whether a given spatial support centroid is within the current range of top North American predators [brown bears, gray wolf (*Canis lupus*), red wolf (*Canis rufus*), or cougar (*Puma concolor*)], which were determined using each species' IUCN range map (McLellan et al. 2017, Boitani et al. 2018); overlap with each species resulted in a categorical predictor of 1 – 4 (depending on the number of overlapped ranges). We included these four species because they are typically competitive with black bear, either dominantly (e.g., brown bear) or indirectly through kleptoparasitism (e.g., cougar) (Ballard et al. 2003, Apps et al. 2006, Latham and Boutin 2011, McLellan 2011).

We first generated centroids of each spatial support window and then assigned each centroid to the hypothesis subdivisions with each centroid, such that each centroid was assigned to one subdivision of each hypothesis (Figure 17). The hypotheses were then tested by fitting linear models with the response (i.e., dependent variable) as the ensemble estimated ecological covariate (n = 1,239 estimates for each covariate, one from each spatial window) and the subdivision of each hypothesis as a categorical predictor variable. Thus, for each covariate, we fit four linear models with different groupings of observations and determined hypothesis support using Akaike's Information Criterion (AIC) (Akaike 1998).

Model Assumptions and Verification

Assumptions

There are many assumptions inherent to wildlife sampling that are typically not met in opportunistic, presence-only databases (Hochachka et al. 2012, Yackulic et al. 2013, Bird et al. 2014). One key assumption is that sampling is either random or representative throughout the area of interest (Yackulic et al. 2013). We explored whether the spatial distribution of iNaturalist mammal observations were representative of land cover conditions across North America using the MODIS land cover product. Using a used-available framework (akin to resource selection functions), we created a categorical classification indicating whether an iNaturalist observation occurred in a grid cell; cell values were 1 if an iNaturalist observation occurred and 0 otherwise. We then computed the most common land cover type within each grid cell and assigned that land cover class to the grid cell. Using this information, we compared use vs available across all 16 land cover classes described by the MODIS product using a chi-square test of homogeneity, testing the null hypothesis that *different populations (used vs available)* have the same proportions of land cover classifications (Thomas and Taylor 1990).

Verification

An important step in modeling animal occurrence records is determining how well a model describes where a species occurs. That is, *predictive performance* of such models is often an essential summary when evaluating model diagnostics (Elith et al. 2006). Typically, this is done by withholding a set of occurrence records, fitting a model without those observations, and then evaluating the model predictions to those locations ("testing points"). However, for the most reliable validation, the testing points should be independent of the observations used to train the model ("training points"); this standard is rarely met because an additional, independent data source is usually not available (Hijmans 2012).

Here, however, we take advantage of a recent, large-scale animal monitoring effort – SNAPSHOT USA – to use as an independent data source to verify parameter estimates from our modeling efforts (Cove et al. 2021). SNAPSHOT USA is a coordinated camera trapping network across the United States, with an overall project goal of sampling terrestrial wildlife in all 50 states during the same time each year. This project required participants to follow a standardized sampling protocol (e.g., camera spacing, length of deployment) along with guidance on where cameras were deployed. We used camera trap observations from 2019 SNAPSHOT USA as a structured data source to verify parameters estimated using iNaturalist observations of black bear and to test whether the model fit using iNaturalist observations could predict black bear occurrences in the SNAPSHOT USA database.

The SNAPSHOT USA database includes camera trap observations from 1509 camera deployments within 110 sampling arrays (non-independent camera clusters) from all 50 U.S. states during 17-Aug-2019 through 24-Nov-2019. We fit single-species, single-season occupancy models to the SNAPSHOT USA camera observations using the same ecological covariates as used in the iNaturalist model. Given repeated observations at a camera site, we was able to explicitly account for imperfect animal detection through occupancy models (MacKenzie et al. 2017b). We used four categorical detection variables that covaried across the camera deployments, including the height a camera was deployed, whether a camera was deployed on a trail, and indicators for obstructed view and understory presence in a camera image.

We assumed a significant difference in regression coefficients between the DWPR iNaturalist model and the SNAPSHOT USA occupancy model if the 95% confidence intervals did not overlap zero. We then used the regression coefficients from global DWPR estimated using iNaturalist observations to predict the occurrence of black bear at the SNAPSHOT USA locations. We evaluated the predictions using the area under the Receiver Operating Characteristics (ROC) curve (AUC) and considered sufficient predictive performance if the AUC was greater than 0.7; we calculated the AUC using the *verification* package (v 1.42) in program R (v 4.0.3) (Laboratory 2015).

Results

iNaturalist Observations

The iNaturalist query resulted in 456,323 terrestrial Mammalia observations during 2015 – 2019, with 9,991 observations of black bear. We removed all iNaturalist observations that were flagged as animal tracks or scat, had an accuracy uncertainty greater than 20,000 m, or were missing coordinates. We additionally buffered the current IUCN black bear range map by 100 km and removed observations outside of this area (n = 5 observations). This brought the total number of Mammalia observations to 370,983 with 6,383 black bear observations. Black bear observations occurred in 3,025 grid cells across North America (min = 1, max = 242, mean = 2 observations per grid cell). Mammalia observations occurred in 45,592 grid cells (min = 1, max = 33,126, mean = 8). Overall, iNaturalist mammal observations were representative of the available land cover types across North America ($\chi_{14}^2 = 0.101$, p-value = 1; Figure C2).

Modeling results

Range-wide iNaturalist model

Results from the global, range-wide model of black bear iNaturalist observations indicated that protected area status had the strongest positive effect on the number of observations across their range (Figure 19; Table C1). The global model indicated greater black bear occurrence with greater forest cover, regardless of forest cover type (conifer, deciduous, and mixed) but bear occurrence was lower where human population density and cropland cover were greater (Figure 19; Table C1). At the continental scale, black bear observations were more common in wetter conditions, but there was no apparent effect of temperature (Figure 19; Table C1).

The global regression coefficients from iNaturalist agreed directionally with SNAPSHOT USA, which explicitly accounted for detection probability, in 75% of the covariates (Figure 19; Table C1). The greatest discrepancy between the iNaturalist and SNAPSHOT USA model occurred in the estimated effect of mean annual temperature, where the iNaturalist model estimated a null effect (i.e., point estimate centered on zero) but the SNAPSHOT USA modeled estimated a negative effect, with the two 95% confidence intervals not overlapping (Figure 19; Table C1). Despite this discrepancy, there was a prediction accuracy of 73% when using the iNaturalist model to predict the SNAPSHOT USA black bear occurrences (AUC = 0.73; Brier Score = 0.11).

Local Ensemble Model

We fit 1,239 DWPR models identical to the global DRWR in the predictor variables used but geographically restricted to observations and environmental conditions within each spatial support window (Figure 20). Although all models converged, of the 1,239 models, 161 models had at least one standard error estimate greater than 20, which we deemed as erroneous and discarded from the ensemble predictions, bringing the total number of spatial support windows to 1,078. The ensemble model had marginally lower absolute value of residual errors compared to the global regression model (error reduction < 1 percent). Nonetheless, the local ensemble modeling approach indicated strong support for spatially varying ecological relationships (Figure 5). Each of the eight ecological covariates had stationary indices greater than one (range 10 – 80), with responses to climatic conditions (i.e., mean annual temperature and precipitation) being the most variable across black bear's range (Figure 21).

The ensemble model results showed notable variation in the spatial patterns of the eight coefficient estimates (Figure 22). The greatest variation in response to climatic conditions – the two ecological covariates with the greatest SI – tended to occur along the Midwest divide in the species' range, particularly in and around Arkansas, Oklahoma, Texas, and Louisiana (Figure 22). The spatial pattern hypothesis comparison indicated the greatest support for the subspecies hypothesis in six of the eight ecological covariates; responses to climatic conditions (i.e., mean annual temperature and precipitation), however, were best explained by the ecoregion hypothesis (Table 12; Figure 23).

Discussion

Despite the mounting evidence that species-environment relationships can vary across space, the key drivers of the variation and the characteristics of the species exhibiting non-stationarity are often unknown (Miller 2012b, Smith et al. 2019a, Pease et al. In Review). Identifying these drivers is important because it may provide nuanced insight into a species' ecology, potentially informing how a species could respond to future environmental change, but may also aid in understanding the types of species that have varied relationships. Our results show that black bear-environmental relationships vary strongly across their range, with the most varied response in climatic conditions (i.e., mean annual temperature and precipitation). To explain this variation across each ecological covariate, we tested four hypotheses (lineage, subspecies, ecoregion, or competition) and found the best support for subspecies boundaries as an explanation in six of the eight covariates but ecoregion boundaries for the remaining two covariates which were measures of climatic conditions. Results demonstrate the complexity of spatiotemporal variation in ecological relationships and point to drivers of genetic adaptation and ecological context as an explanation of non-stationarity in a wide-ranging, generalist species.

For black bear, subspecies boundaries most often described intraspecific variation in the response to land cover and human pressure. This suggests that genetic affinity plays an important role in the ecological relationships of one of the most opportunistic, generalist mammals of North America. There are 16 subspecies of black bear, and most are similar morphologically, with minor differences in cranial shape (Larivière 2001, Puckett et al. 2015). However, the subspecies do differ substantially in their relative geographic position within the species' range, resulting in considerable among-subdivision variation in land cover and climatic conditions, suggesting ecological context may also be an important driver of non-stationarity in land cover and human pressure responses but may be masked by the subspecies delineations. That said, although the majority of the ecological associations were best described by subspecies, the nonstationarity exhibited by black bear provides further ecological support for the importance of considering alternatives to genetic similarity for explaining intraspecific variation, such as ecological context, which was the best supported hypothesis for climatic associations (Sexton et al. 2009, Smith et al. 2019a, Pease et al. In Review). Ecological relationships are linked tightly to environment conditions, and continued changes in land use, human densities, and climatic conditions will likely continue to modify species-environment relationships over space, regardless of whether subspecies boundaries remain static through time, potentially resulting in notable differences of black bear distribution in the current century (Beckmann and Berger 2003).

The strongest nonstationarity in bear observations was linked to mean annual temperature and mean annual precipitation, indicating local adaptation to climatic conditions. The strongest signals with these climate variables occurred on the boundaries of black bear's range (e.g., Missouri, southern Minnesota), and the magnitude of the effects suggested strong avoidance and attraction to climatic conditions, depending on the geographic region. For example, black bears were attracted to cooler conditions in Arkansas, Oklahoma, and east Texas, but warmer conditions at the northern boundary of the species' range. Black bear's extensive geographical range and the inherent broad climatic niche associated with this range, coupled with evidence of local adaptation to climatic conditions, suggests a lack of vulnerability to forecasted climatic variation and thus a weaker or neutral relationship with climate variables should follow.

We also found an interesting spatial pattern in the relationship with protected area status. Black bears are typically associated with protected areas, particularly in the presence of high human disturbance, but in the Ouachita mountain range of Oklahoma, Arkansas, and northern Louisiana, bears were less common in protected areas. Several large national forests exist in this region but many iNaturalist observations occurred outside the protected areas. This may be due to reporting of black bear observations to iNaturalist in the area. The Arkansas Game and Fish Commission put out a public call for citizens to report all black bear observations to iNaturalist, potentially causing a higher rate of reporting in residential areas outside of the protected area boundaries. Alternatively, unlike some western regions with strong, positive effects of protected areas, the high forest cover of the Ouachita region extends beyond protected area boundaries, supporting black bear occurrence beyond protected areas, resulting in an apparent negative relationship of protected areas.

Although interspecific competition can drive differential species-environment relationships (Durrett and Levin 1998), and has been found to be important for sympatric bear species (Schwartz and Franzmann 1991, Apps et al. 2006, Belant et al. 2010, McLellan 2011), . we found little support for this hypothesis as the driver of nonstationarity. This may be due to the geographic scale at which we studied black bear-environment relationships. Studies of competition in large mammals are typically conducted over restricted geographic areas, and intensity of competition can vary at fine-scales (e.g., 4th-order selection; Johnson 1980). Given this, differentiating only by whether the competing species' ranges overlap may be lacking the resolution needed to identify strong signals of competition-mediated species-environment relationships. Similarly, black bear Pleistocene refugia lineages are also quite coarse subdivisions and an individual lineage can occupy broad spectrums of environmental conditions that may lead to local adaptation within the lineage (Illborn and Langerhans 2015). The lack of support for these hypotheses – competition and lineage – suggest the resolution at which these subdivisions are capturing variation may be too coarse to describe space-varying relationships for black bear.

Testing Assumptions of iNaturalist data

The use of iNaturalist observations to inform wildlife ecology is in its infancy, and there is much debate over the validity of using such opportunistic, presence-only, citizen science observations to describe ecological relationships (Hochachka et al. 2012, Bird et al. 2014). To address this, we assessed (1) the accuracy of observations, (2) how representatively these observations sampled land cover classes across North America and (3) how ecological findings match results from an independent but systematic camera trap survey. our verification efforts resulted in three interesting findings. First, the community identified records were very accurate in identifying black bears to species level in photographs (100% accuracy of black bear photographs). Second, mammal observations from the iNaturalist database were representative of the available land cover conditions across North America. Third, the modeled ecological relationships closely matched those found using the SNAPSHOT USA database. Together, these results show that iNaturalist records are suitable for studying ecological relationships of black bears across their range with spatial resolution and geographic coverage unparalleled by other available data.

To account for additional sources of bias (e.g., inhomogeneous sampling effort), we used a unique approach of leveraging the iNaturalist database itself to describe the sampling intensity across North America (Kéry et al. 2010). We chose to restrict the database to mammal species only under the assumption that individuals using iNaturalist may tend to report species of a given taxa. Given that we used a taxa-specific search intensity as an offset in our models, we may have underestimated the actual intensity of the Poisson process. That is, using the complete iNaturalist database (i.e., across all taxa) may have been a better approach to account for sampling intensity because it more thoroughly describes the reporting probability associated with this database. This latter approach would relax our assumption by simply assuming that if an iNaturalist user was in the vicinity of a black bear, then the user would report it (rather than ignoring because the species was a mammal instead of, for example, an insect).

Nonetheless, the parameter estimates from the global DWPR model fit using iNaturalist records agreed with those estimates from an occupancy model fit using SNAPSHOT USA observations. The one clear exception was the estimated effect of mean annual temperature on black bear distribution. iNaturalist observations suggested a significant, positive relationship with mean annual temperature while SNAPSHOT USA suggested a significant, negative effect. We believe this may be due to the restricted sampling extent of SNAPSHOT USA, where the observations were sampling a comparatively small portion of the climatic gradients across North America; iNaturalist observations occurred almost completely across the entire latitudinal gradient of black bear's current range. Despite this conflict, estimated effects of land cover conditions and human population density agreed, lending further verification to iNaturalist as a potentially important source of ecological information for wildlife.

Using the presence-only observations of iNaturalist, we was unable to account for an important source of sampling bias – imperfect detection of black bear (Kéry and Schmidt 2008). Imperfect detection of a species – that is, the tendency to not detect a species given it is present – can lead to biased parameter estimates and potentially conflate ecological effects with the observation process. For black bear, we believe this source of bias is reduced because its' large body size and diurnal habits make it conspicuous. We attempted to account for reporting probability through the use of sampling effort and occurrence probability using a suite of ecological covariates known to affect the species, thus we believe our ecological findings are reflective of the species' ecology. Further, we were able to explicitly account for imperfect detection in the SNAPSHOT USA through hierarchical modeling, and parameter estimates from those modeling efforts mirrored the ecological findings of our iNaturalist model, suggesting that the bias of imperfect detection may be minimal for iNaturalist black bear observations taken at the 50-km^2 scale.

Our research contributes to the increasing evidence that species-environment relationships are not constant over space, and whether niche conservatism or local adaptation to ecological context explains the variation depends on the environmental conditions being evaluated (e.g., land use vs climatic) (Smith et al. 2019a, Pease et al. In Review). Although we documented support for subspecies delineations, local adaptation may be occurring at a finer scale than tested here and future research should look to fine-scale drivers such as human pressure or species-specific management regimes as drivers of local adaptation. Further, we are still lacking a consensus of the characteristics associated with species exhibiting nonstationarity. We believe that lifespan, body size, and dispersal capabilities of a species may be strong indicators and warrant future research. We suspect that long-lived species will have a greater tendency for local adaptation compared to short-lived species. We am also inclined to think larger bodied species with greater dispersal capabilities will be more likely to show non-constant ecological relationships as these species encounter novel environments. Future work should look to explore this across taxa over large geographic space; iNaturalist observations may be a promising path forward.

The strong ties we found to subspecies and ecoregions suggest that nonstationarity is taking place over a relatively fine spatial scale, especially in the context of wide-ranging species such as black bear. Our findings have implications for studying how species-environment relationships may change over time, and we encourage the acknowledgement of such variation in correlative assessments of species distribution, abundance, and richness. We also believe that persistent research on the validity and use of opportunistic, presence-only databases such as iNaturalist will continue to lead to interesting, nuanced ecological findings not apparent in studies conducted over restricted geographic areas.

Table 11: List of covariates used in the analysis of American black bear (*Ursus americanus*) iNaturalist observations during 2015 - 2019 across North America.

Submodel	Covariate	Description	Source
Ecological	Deciduous broadleaf forest	Proportion of deciduous broadleaf forest (MODIS class 4)	MODIS* Land Cover Type (MCD12Q1)
	Coniferous forest	Proportion of evergreen needleleaf forest (MODIS class 1)	MODIS* Land Cover Type (MCD12Q1)
	Mixed forest	Proportion of mixed forest (MODIS class 5)	MODIS* Land Cover Type (MCD12Q1)
	Cropland	Proportion of cropland (MODIS class 12)	MODIS* Land Cover Type (MCD12Q1)
	Temperature	Annual mean temperature	WorldClim
	Precipitation	Annual mean precipitation	WorldClim
	Human population density	Mean human population density	GPW^ Version 4
	Protected Areas	Proportion of designated protected area	WPDA†

Table 11 (continued)

	Trail (SNAPSHOT USA)	Indicator for whether camera was deployed on trail	Deployment metadata
	Camera height (SNAPSHOT USA)	Categorical indicating deployment height (1, 2, or 3)	Deployment metadata
Observation	Understory vegetation (SNAPSHOT USA)	Indicator for whether understory vegetation was present in camera image	Camera observations
	Obstructed view (SNAPSHOT USA)	Indicator for whether camera view was obstructed by natural or unnatural features	Camera observations
	Total no. of reported mammal observations (iNaturalist)	The total number of reported mammal observations on iNaturalist (search restricted to *Mammalia* class). Used as a measure of search effort or tendency to report observations in an area	iNaturalist API

Table 12: Model comparisons testing which non-stationarity hypothesis best explains variation in each ecological covariate's spatial pattern. R-squared column indicates the coefficient of variation for the top model. Delta AIC values indicate the change in AIC from the top model.

			Delta AIC			
Covariate	Top Model	R-Squared	Lineage	Subspecies	Ecoregion	Competition
Coniferous Forests	Subspecies	0.22	251	0	129	179
Deciduous Forests	Subspecies	0.24	260	0	14	203
Mixed Forests	Subspecies	0.14	148	0	105	146
Cropland	Subspecies	0.22	114	0	77	143
Human Population Density	Subspecies	0.24	173	0	128	134
Protected Area Status	Subspecies	0.30	150	0	11	84
Mean Annual Temperature	Ecoregion	0.19	209	16	0	99
Mean Annual Precipitation	Ecoregion	0.13	19	14	0	58

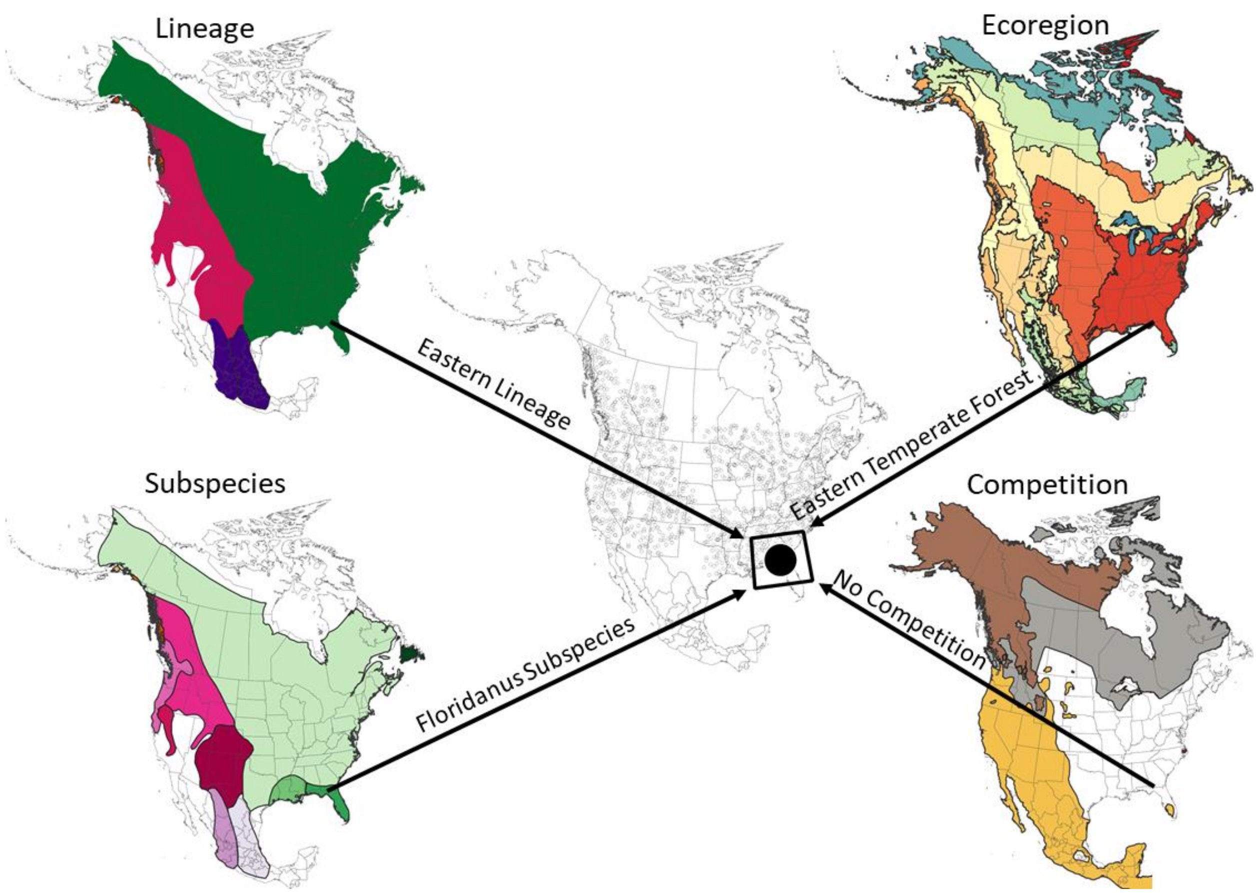


Figure 17: Visualization of four hypotheses driving non-stationarity in American Black Bear (Ursus americanus) ecological relationships across the species' current range, with an example of how a given spatial support window (gray points) is assigned to a single subdivision of each hypothesis. Each gray point represents the centroid of a spatial support window (n = 1,239). The spatial support windows were used to build local models across the species' range. Each local model estimates ecological relationships based on observations within the window. Each centroid was assigned one subdivision of each hypothesis. The example here shows a box centroid in the panhandle of Florida, USA and is assigned to the: (1) "Eastern Temperate Forest" subdivision of the North America Ecoregions; (2) "Eastern" black bear lineage; (3) "Floridanus" subspecies; and (4) No overlap with competing predators [brown bear (Ursus arctos, green), gray wolf (Canis lupus, brown), red wolf (Canis rufus), and cougar (Puma concolor)].

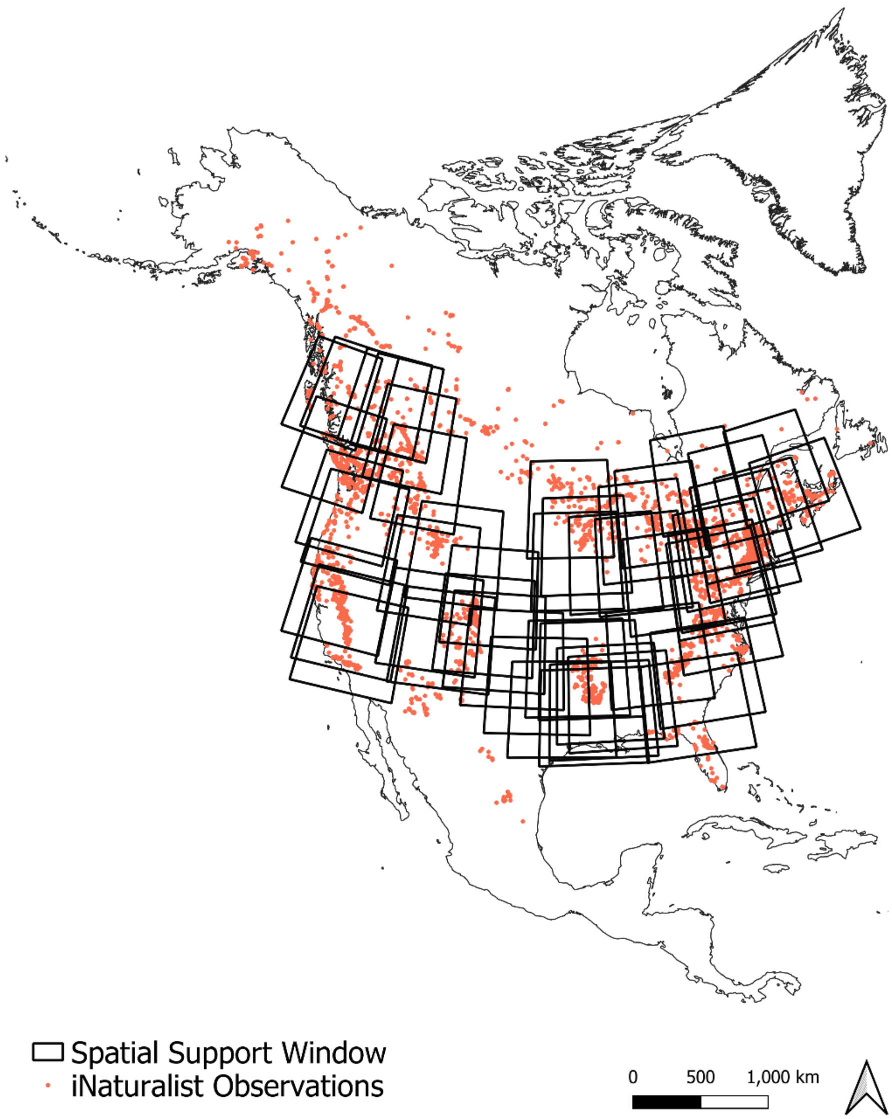


Figure 18: Local ensemble models were developed using restricted spatial support windows across observations of American Black Bear (Ursus americanus) reported on iNaturalist during 2015 – 2019 (orange dots). Spatial support windows (black-outlined boxes) (total = 1,239; n = 50 presented here for illustration) serve as bounds for observations and background points in the ensemble modeling approach. Results from overlapping boxes are averaged to effectively "scale up" to the range of black bear.

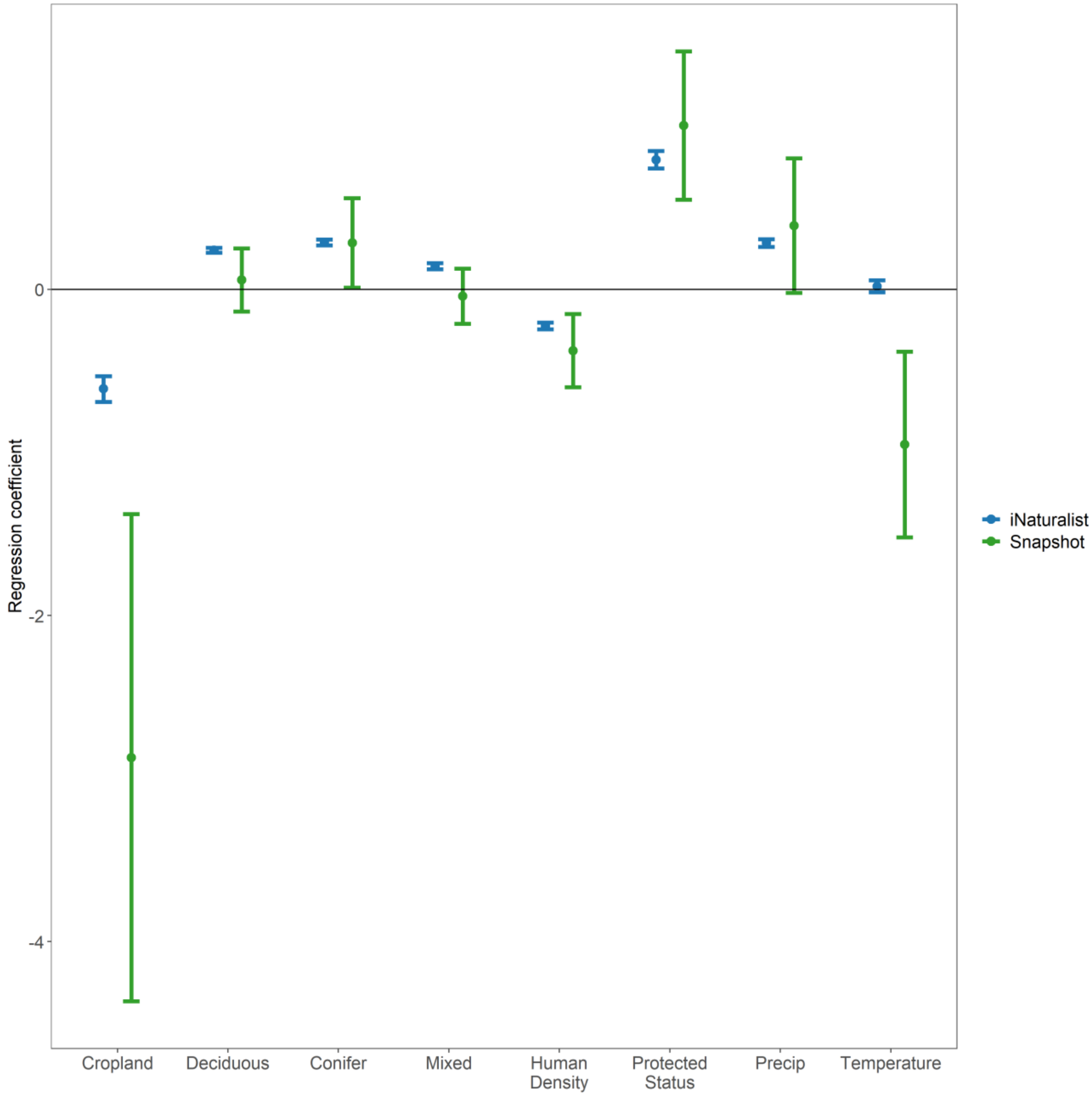


Figure 19: A comparison of regression coefficient estimated using a global Downweighted Poisson Regression model fit using American Black Bear (Ursus americanus) observations from the iNaturalist database during 2015-2019 (blue dots and lines) and a single-species, single-season occupancy model fit using camera observations from the 2019 SNAPSHOT USA monitoring initiative (green dots and lines). Dots represent point estimates and bars indicate 95% confidence intervals.

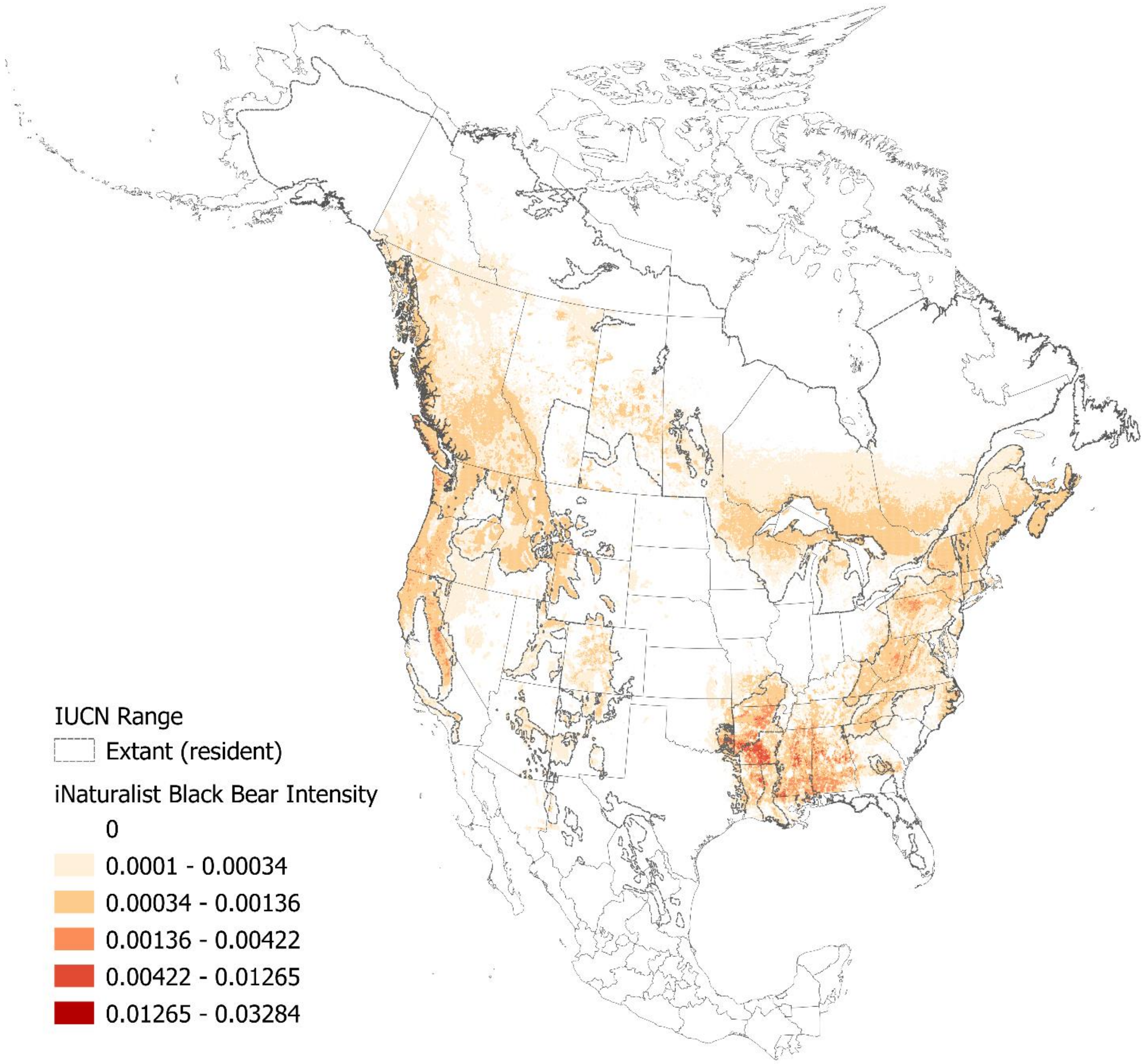


Figure 20: Predicted intensity from an ensemble modeling approach estimated using 1,078 geographically restricted spatial support windows and observations of American Black Bear (Ursus americanus) reported on iNaturalist during 2015 – 2019. Dotted outlines represent current known black bear species range. White colors represent areas beyond the range of black bear or regions of the species range where windows had less than 100 black bear observations and were omitted from analysis.

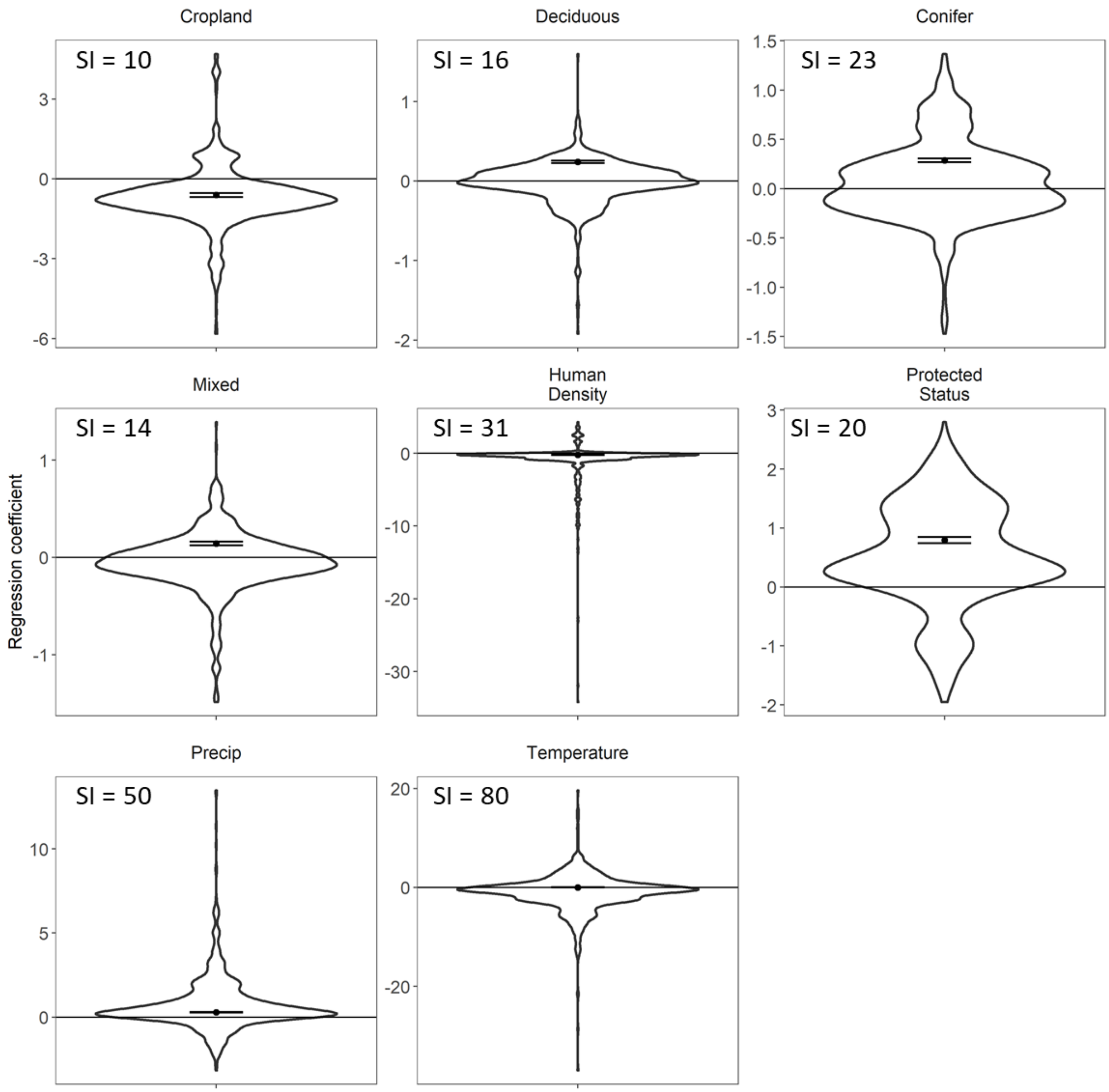


Figure 21: Regression coefficients estimates (violins; n = 1,239) from the regional Downweighted Poisson Regression (DWPR) models fit to iNaturalist black bear observations during 2015 - 19 across North America. Violins represent the coefficient estimates across all spatial support windows. Also included for comparison is the 95% confidence intervals (horizontal lines) from the global model. All covariates were scaled and centered prior to analysis. Stationary indices (SI) are reported in top left corner of each panel; a SI > 1 indicates support for non-stationarity of that covariate across their range.

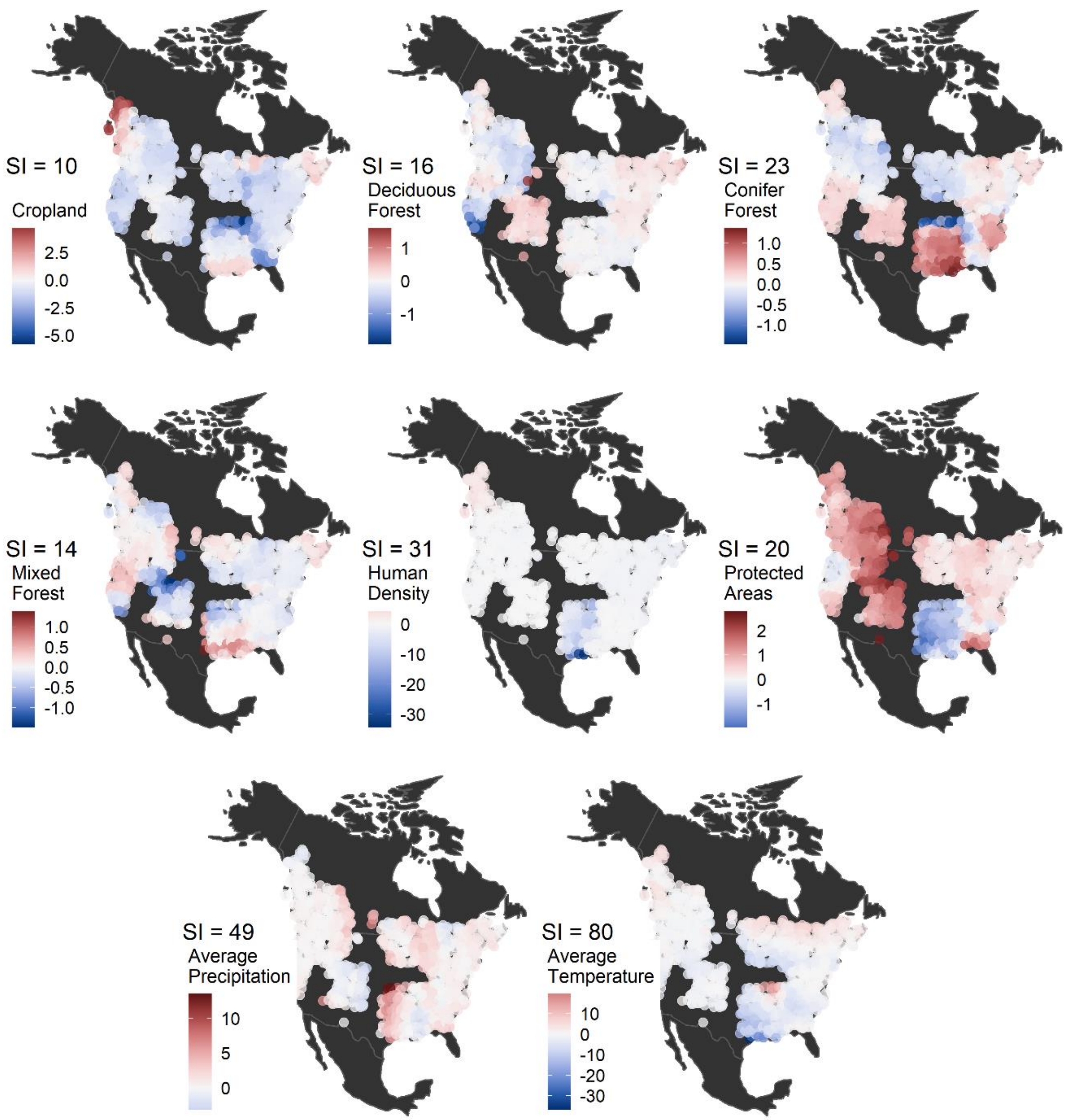


Figure 22: Estimated regression coefficients from the local ensemble models (n = 1,239) using American Black Bear (Ursus americanus) iNaturalist observations during 2015 – 2019 across North America. Each subfigure shows the spatial variation for a given ecological relationship (name listed near legend), where the legend represents the range of the estimated coefficients for each covariate, along with the Stationary Index (SI); SI values greater than one indicate support for non-stationarity.

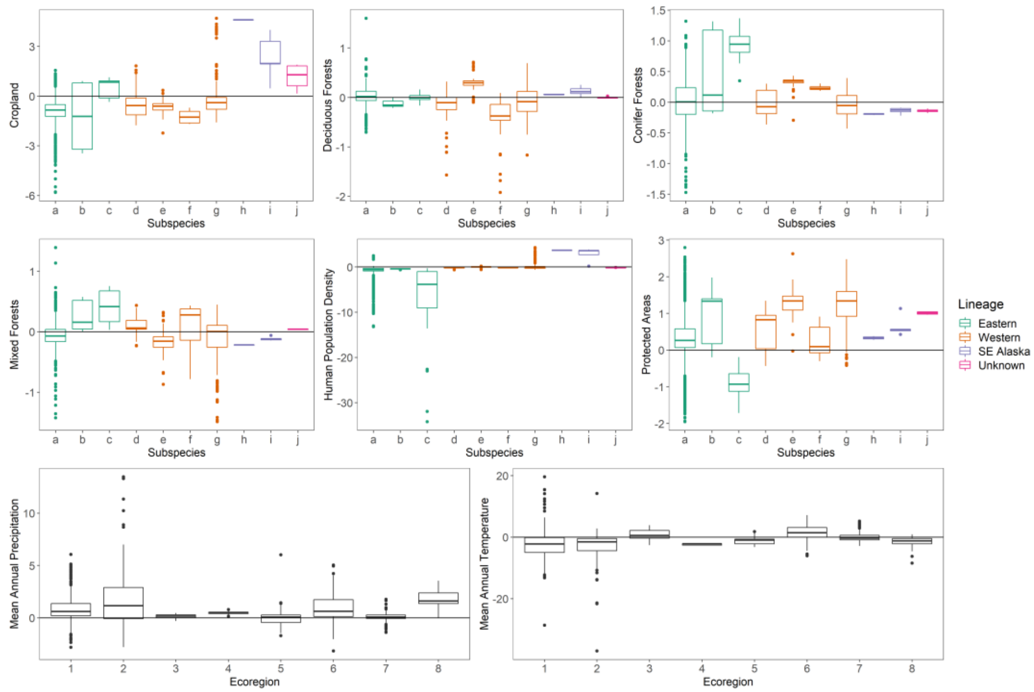


Figure 23: Comparison of estimated regression coefficients from local ensemble modeling approach across each of the eight focal ecological covariates, grouped by the best supported spatial pattern hypothesis (lineage, subspecies, ecoregion, or competition). For covariates best supported by the subspecies hypothesis, color groupings indicate lineages associated with each subspecies. The two climatic responses (bottom row) were best explained by ecoregions. Subspecies subdivisions: (a) americanus, (b) floridanus, (c) luteolus, (d) altifrontalis, (e) amblyceps, (f) californiensis, (g) cinnamonum, (h) carlotte, (i) kermodei, and (j) vancouveri. Ecoregion subdivisions: (1) eastern temperate forests, (2) great plains, (3) marine west coast forest, (4) Mediterranean California, (5) north American deserts, (6) northern forests, (7) northwestern forested mountains, and (8) water dominated.

CHAPTER 5: CONCLUSION

Understanding the distribution of wildlife, or where a species occurs in a geographic region during a specific time, is a fundamental state variable in ecology and essential to effectively manage and conserve wildlife populations (Elith and Leathwick 2009, Kéry and Royle 2015). In this dissertation, I documented widespread support of spatial non-stationarity in species-environment relationships, both at the state- and national-level, suggesting the drivers of current species distributions vary across their range. This has important implications for the way wildlife populations are managed across large geographic space and for the way ecologists think about species-environment relationships. Specifically, though largely recognized as static, species-environmental relationships are likely continually changing through time and space, and consistent monitoring of these changes is likely needed to effectively manage and conserve wildlife.

As I illustrated in this dissertation, several methods to estimate spatially varying coefficients exist, including eigenvector spatial filtering (Griffith 1996, 2008), Generalized Additive Models (Wood 2017), Geographic Weighted Regression (GWR) (Fotheringham et al. 2003), and Bayesian Spatially-Varying Coefficients Models (SVCs) (Gelfand et al. 2003, Banerjee et al. 2014). As concluding remarks, I will focus on a review of GWR and SVCs to guide future implementation of such methods. Spatially-Varying Coefficients Models are fully parametric hierarchal models that place spatial Gaussian processes on the coefficient parameters to allow their effect to vary over the geographic extent of interest. In contrast, GWR is an exploratory modeling approach intended to indicate if and where non-stationarity in ecological relationships is occurring over an area of interest (McMillen 2004). More specifically, GWR assumes there is some continuous spatial random field of parameter values and there parameters are measured at discrete sampling locations in the random field to describe the spatial variability, with the coefficients being deterministic functions of locations in space (Fotheringham et al. 2003). So when estimating a covariate effect at some location, a subset of the sampling locations surrounding the focal point are used to estimate the parameter at this location, and the importance of the surrounding locations in estimating this effect is assumed to decay with distance from the focal site. That is, the response and covariate values at a nearby neighboring site are assumed to be more influential on the focal site than another sampling location that is farther away. Given the name of this model, the neighboring sites are *weighted* by their distance to the focal site, with nearby sites receiving more weight.

GWR constructs a unique regression model for every location in the dataset that occurs within a specified bandwidth of a given data point to identify where locally weighted regression coefficients deviate from the global regression coefficients (Bivand 2017). In other words, GWR uses a moving window with a specified bandwidth (i.e., shape and extent of ’buffer’ surrounding a given point) over each point and all points occurring within the bandwidth are used in a ’local’ model, where increased weights are given to points closer to the focal data point. A global coefficient is calculated from the complete dataset, and then each data point can also be assigned a local coefficient to see and how relationships vary over space. If a dataset has, for example, has 100 data points, 100 local regression models will be fit.

GWR methods are contrasted with SVC in that the latter is a valid probabilistic model affording full posterior distributions while GWR tends to be considered valid for exploratory analysis only. In GWR, the joint distribution linking the parameters and the data is not a valid probability distribution, and thus probabilistic statements cannot be made from the results (Finley 2011). Other issues with GWR include the problem of local collinearity among covariates and issues handling complicated spatial structuring in the data (Wheeler and Tiefelsdorf 2005, Wheeler and Waller 2009), but Fotheringham attempted to discredit these claims in recent years (Fotheringham and Oshan 2016). Another consideration is that some results suggest that GWR is less precise than SVC, especially as spatial extent increases (Ma et al. 2012).

Future monitoring of wildlife distributions should work to account for space-varying effects of the environment. If possible, we recommend implementing an SVC model to evaluate these effects, but each modeling approach will be a balancing act between the number of observations, the extent of the area of interest, and the desired resolution of inference. Finally, as illustrated in chapter 3, we recommend researchers work to include multiple streams of information when possible to ensure a more robust inference of wildlife occurrence.

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 91:7–14.

Akaike, H. 1998. Information theory and an extension of the maximum likelihood principle. Pages 199–213 Selected papers of hirotugu akaike. Springer.

Apps, C. D., B. N. McLellan, and J. G. Woods. 2006. Landscape partitioning and spatial inferences of competition between black and grizzly bears. ECOGRAPHY 29:561–572.

Austin, M. P., and K. P. Van Niel. 2011. Improving species distribution models for climate change studies: variable selection and scale. Journal of Biogeography 38:1–8.

Ballard, W. B., L. N. Carbyn, and D. W. Smith. 2003. Wolf interactions with non-prey.

Banerjee, S., B. P. Carlin, and A. E. Gelfand. 2014. Hierarchical modeling and analysis for spatial data. Chapman and Hall/CRC.

Beckmann, J., and J. Berger. 2003. Rapid ecological and behavioural changes in carnivores: the responses of black bears (Ursus americanus) to altered food. JOURNAL OF ZOOLOGY 261:207–212.

Belant, J. L., B. Griffith, Y. Zhang, E. H. Follmann, and L. G. Adams. 2010. Population-level resource selection by sympatric brown and American black bears in Alaska. Polar Biology 33:31–40.

Berger, J., P. B. Stacey, L. Bellis, and M. P. Johnson. 2001. A mammalian predator–prey imbalance: grizzly bear and wolf extinction affect avian neotropical migrants. Ecological Applications 11:947–960.

Bini, L. M., J. A. F. Diniz-Filho, T. F. L. V. B. Rangel, T. S. B. Akre, R. G. Albaladejo, F. S. Albuquerque, A. Aparicio, M. B. Araújo, A. Baselga, J. Beck, M. I. Bellocq, K. Böhning-Gaese, P. A. V. Borges, I. Castro-Parga, V. K. Chey, S. L. Chown, P. De Marco, D. S. Dobkin, D. Ferrer-Castán, R. Field, J. Filloy, E. Fleishman, J. F. Gómez, J. Hortal, J. B. Iverson, J. T. Kerr, W. D. Kissling, I. J. Kitching, J. L. León-Cortés, J. M. Lobo, D. Montoya, I. Morales-Castilla, J. C. Moreno, T. Oberdorff, M. Á. Olalla-Tárraga, J. G. Pausas, H. Qian, C. Rahbek, M. Á. Rodríguez, M. Rueda, A. Ruggiero, P. Sackmann, N. J. Sanders, L. C. Terribile, O. R. Vetaas, and B. A. Hawkins. 2009. Coefficient shifts in geographical ecology: An empirical evaluation of spatial and non-spatial regression. Ecography 32:193–204.

Bird, T. J., A. E. Bates, J. S. Lefcheck, N. A. Hill, R. J. Thomson, G. J. Edgar, R. D. Stuart-Smith, S. Wotherspoon, M. Krkosek, and J. F. Stuart-Smith. 2014. Statistical solutions for error and bias in global citizen science datasets. Biological Conservation 173:144–154.

Bivand, R. 2017. Geographically weighted regression.

Bled, F., J. A. Royle, and E. Cam. 2011. Hierarchical modeling of an invasive spread: the Eurasian Collared-Dove Streptopelia decaocto in the United States. Ecological Applications 21:290–302.

Boitani, L., M. Phillips, and Y. Jhala. 2018. Canis lupus (errata version published in 2020). The IUCN Red List of Threatened Species e.T3746A163508960.

Bolker, B. M., M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J.-S. S. White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in ecology & evolution 24:127–135.

Bureau, U. S. C. 2010. TIGER/Line shapefiles. US Census Bureau.

Byrne, M. E., J. J. Vaudo, G. C. M. Harvey, M. W. Johnston, B. M. Wetherbee, and M. Shivji. 2019. Behavioral response of a mobile marine predator to environmental variables differs across ecoregions. Ecography.

Center for International Earth Science Information Network - CIESIN - Columbia University. 2018. Gridded Population of the World, Version 4 (GPWv4): Population Density, Revision 11. NASA Socioeconomic Data and Applications Center (SEDAC), Palisades, NY.

Chase, J. M., and M. A. Leibold. 2003. Ecological niches: linking classical and contemporary approaches.

Cherry, M. J., P. E. Howell, C. D. Seagraves, R. J. Warren, and L. M. Conner. 2017. Effects of land cover on coyote abundance. Wildlife Research 43:662–670.

Clare, J. D. J., E. M. Anderson, and D. M. MacFarland. 2015. Predicting bobcat abundance at a landscape scale and evaluating occupancy as a density index in central Wisconsin. The Journal of Wildlife Management 79:469–480.

Clare, J. D., P. A. Townsend, C. Anhalt-Depies, C. Locke, J. L. Stenglein, S. Frett, K. J. Martin, A. Singh, T. R. Van Deelen, and B. Zuckerberg. 2019. Making inference with messy (citizen science) data: when are data accurate enough and how can they be improved? Ecological Applications 29:e01849.

Clark, J. D., J. P. Beckmann, M. S. Boyce, B. D. Leopold, A. E. Loosen, and M. R. Pelton. 2020. American Black Bear (Ursus americanus). Pages 122–138 *in* M. Melletti and V. Penteriani, editors. Bears of the World: Ecology, Conservation and Management. Cambridge University Press, Cambridge.

Commission for Environmental Cooperation. 1997. Ecological regions of North America: toward a common perspective. Montreal, Quebec, Canada.

Constible, J. M., M. J. Chamberlain, and B. D. Leopold. 2006. Relationships between landscape pattern and space use of three mammalian carnivores in central Mississippi. The American midland naturalist 155:352–363.

Cove, M. V., B. M. Jones, A. J. Bossert, D. R. Clever Jr, R. K. Dunwoody, B. C. White, and V. L. Jackson. 2012. Use of camera traps to examine the mesopredator release hypothesis in a fragmented Midwestern landscape. The American Midland Naturalist:456–465.

Cove, M. V., R. Kays, and H. Bontrager. 2021. Snapshot USA 2019: A coordinated national camera trap survey of the United States. Ecology.

Curtis, S. M. 2018. mcmcplots: Create Plots from MCMC Output.

Ditmer, M. A., K. V. Noyce, J. R. Fieberg, and D. L. Garshelisa. 2018. Delineating the ecological and geographic edge of an opportunist: The American black bear exploiting an agricultural landscape. ECOLOGICAL MODELLING 387:205–219.

Dowle, M., and A. Srinivasan. 2019. data.table: Extension of `data.frame`.

Durrett, R., and S. Levin. 1998. Spatial aspects of interspecific competition. Theoretical population biology 53:30–43.

Elith, J., C. H. Graham*, R. P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, and others. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129–151.

Elith, J., and J. R. Leathwick. 2009. Species distribution models: ecological explanation and prediction across space and time. Annual review of ecology, evolution, and systematics 40:677–697.

Elith, J., S. J. Phillips, T. Hastie, M. Dudík, Y. E. Chee, and C. J. Yates. 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17:43–57.

Elton, C. 1946. Competition and the structure of ecological communities. The Journal of Animal Ecology:54–68.

Ericsson, G., and K. Wallin. 1999. Hunter observations as an index of moose Alces alces population parameters. Wildlife Biology 5:177–185.

Faraggi, D., and B. Reiser. 2002. Estimation of the area under the ROC curve. Statistics in medicine 21:3093–3106.

Feldhamer, G. A., B. C. Thompson, and J. A. Chapman. 2003. Wild mammals of North America: biology, management, and conservation. JHU Press.

Fick, S. E., and R. J. Hijmans. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. International journal of climatology 37:4302–4315.

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 20:2131–2147.

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

Foody, G. M. 2004. Spatial nonstationarity and scale-dependency in the relationship between species richness and environmental determinants for the sub-Saharan endemic avifauna. Global Ecology and Biogeography 13:315–320.

Fortin, D., R. Courtois, P. Etcheverry, C. Dussault, and A. Gingras. 2008. Winter Selection of Landscapes by Woodland Caribou: Behavioural Response to Geographical Gradients in Habitat Attributes . Blackwell Publishing .

Fotheringham, A. S., C. Brunsdon, and M. Charlton. 2003. Geographically weighted regression: the analysis of spatially varying relationships. John Wiley & Sons.

Fotheringham, A. S., and T. M. Oshan. 2016. Geographically weighted regression and multicollinearity: dispelling the myth. Journal of Geographical Systems 18:303–329.

Fox, J., and S. Weisberg. 2019. An R Companion to Applied Regression. Third. Sage, Thousand Oaks CA.

Franklin, J. 2010. Mapping species distributions: spatial inference and prediction. Cambridge University Press.

Fuller, J., C. Kreh, R. Myers, C. Olfenbuttel, J. Shaw, and A. Shipley. 2019. Deer Hunter Observation Survey 2014 - 2019. Page 37. North Carolina Wildife Resources Commission, Raleigh, North Carolina.

Garshelis, D. L., B. Scheick, D. L. Doan-Crider, J. Beecham, and M. E. Obbard. 2016. Ursus americanus (errata version published in 2017). e.T41687A114251609. https://dx.doi.org/10.2305/IUCN.UK.2016-3.RLTS.T41687A45034604.en.

Gehrt, S. D., J. L. Brown, and C. Anchor. 2011. Is the urban coyote a misanthropic synanthrope? The case from Chicago. Cities and the Environment (CATE) 4:3.

Gelfand, A. E., H.-J. Kim, C. F. Sirmans, and S. Banerjee. 2003. Spatial modeling with spatially varying coefficient processes. Journal of the American Statistical Association 98:387–396.

Gelman, A., D. K. Park, S. Ansolabehere, P. N. Price, and L. C. Minnite. 2001. Models, assumptions and model checking in ecological regressions. Journal of the Royal Statistical Society: Series A (Statistics in Society) 164:101–118.

Gorini, L., J. D. Linnell, R. May, M. Panzacchi, L. Boitani, M. Odden, and E. B. Nilsen. 2012. Habitat heterogeneity and mammalian predator–prey interactions. Mammal review 42:55–77.

Green, S. B. 1991. How many subjects does it take to do a regression analysis. Multivariate behavioral research 26:499–510.

Griffith, D. A. 1996. Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data. Canadian Geographer/Le Géographe canadien 40:351–367.

Griffith, D. A. 2008. Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR). Environment and Planning A 40:2751–2769.

Guillera-Arroita, G., J. J. Lahoz-Monfort, J. Elith, A. Gordon, H. Kujala, P. E. Lentini, M. A. McCarthy, R. Tingley, and B. A. Wintle. 2015. Is my species distribution model fit for purpose? Matching data and models to applications. Global Ecology and Biogeography 24:276–292.

Hällfors, M. H., J. Liao, J. Dzurisin, R. Grundel, M. Hyvärinen, K. Towle, G. C. Wu, and J. J. Hellmann. 2016. Addressing potential local adaptation in species distribution models: implications for conservation under climate change. Ecological Applications 26:1154–1169.

Halls, L. K. 1984. White-tailed deer: ecology and management. Stackpole Books.

Hastie, T., and W. Fithian. 2013. Inference from presence-only data; the ongoing controversy. Ecography 36:864–867.

Hesselbarth, M. H. K., M. Sciaini, K. A. With, K. Wiegand, and J. Nowosad. 2019. landscapemetrics: an open-source R tool to calculate landscape metrics. Ecography 42:1648–1657.

Hijmans, R. J. 2012. Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. Ecology 93:679–688.

Hinton, J. W., F. T. van Manen, and M. J. Chamberlain. 2015. Space use and habitat selection by resident and transient coyotes (Canis latrans). PLoS One 10:e0132203.

Hochachka, W. M., D. Fink, R. A. Hutchinson, D. Sheldon, W.-K. Wong, and S. Kelling. 2012. Data-intensive science applied to broad-scale citizen science. Trends in ecology & evolution 27:130–137.

Homer, C., J. Dewitz, L. Yang, S. Jin, P. Danielson, G. Xian, J. Coulston, N. Herold, J. Wickham, and K. Megown. 2015. Completion of the 2011 National Land Cover Database for the conterminous United States–representing a decade of land cover change information. Photogrammetric Engineering & Remote Sensing 81:345–354.

Howard, B., D. Sawyer, D. Cobb, E. Stanford, J. Shaw, and R. Myers. 2015. Evaluation of Deer Hunting Season Structures and Deer Management Units In North Carolina. North Carolina Wildife Resources Commission, Raleigh, North Carolina.

Hutchinson, G. E. 1957. Concluding remarks. Pages 415–427 Cold Spring Harbor symposia on quantitative biology. Cold Spring Harbor Laboratory Press.

iNaturalist.org. 2020. . https://www.inaturalist.org/pages/what+is+it.

Isaac, N. J. B., A. J. van Strien, T. A. August, M. P. de Zeeuw, and D. B. Roy. 2014. Statistics for citizen science: extracting signals of change from noisy ecological data. Methods in Ecology and Evolution 5:1052–1060.

Isaac, N. J., M. A. Jarzyna, P. Keil, L. I. Dambly, P. H. Boersch-Supan, E. Browning, S. N. Freeman, N. Golding, G. Guillera-Arroita, and P. A. Henrys. 2020. Data integration for large-scale models of species distributions. Trends in ecology & evolution 35:56–67.

Jarzyna, M. A., A. O. Finley, W. F. Porter, B. A. Maurer, C. M. Beier, and B. Zuckerberg. 2014. Accounting for the space-varying nature of the relationships between temporal community turnover and the environment. Ecography 37:1073–1083.

Johnson, D. H. 1980. The comparison of usage and availability measurements for evaluating resource preference. Ecology 61:65–71.

Jones, M. D., and M. R. Pelton. 2003. Female American black bear use of managed forest and agricultural lands in coastal North Carolina. Ursus:188–197.

Jones, M. E., and L. A. Barmuta. 2000. Niche differentiation among sympatric Australian dasyurid carnivores. Journal of Mammalogy 81:434–447.

Kays, R., M. Lasky, B. S. Pease, A. W. Parsons, and K. Pacifici. In Review. Evaluation of the spatial biases and sample size of a statewide citizen science project. Citizen Science: Theory and Practice.

Kays, R., W. J. McShea, and M. Wikelski. 2020a. Born digital biodiversity data: millions and billions. Diversity and Distributions 26:644–648.

Kays, R., W. J. McShea, and M. Wikelski. 2020b. Born-digital biodiversity data: Millions and billions. Diversity and Distributions 26:644–648.

Kays, R., A. Parsons, M. C. Baker, E. L. Kalies, T. Forrester, R. Costello, C. T. Rota, J. J. Millspaugh, and W. J. McShea. 2017. Does hunting or hiking affect wildlife communities in protected areas? Journal of Applied Ecology 54:242–252.

Kays, R., and A. W. Parsons. 2014. Mammals in and around suburban yards, and the attraction of chicken coops. Urban Ecosystems 17:691–705.

Kéry, M., B. Gardner, and C. Monnerat. 2010. Predicting species distributions from checklist data using site-occupancy models. Journal of Biogeography 37:1851–1862.

Kéry, M., and J. A. Royle. 2015. Applied Hierarchical Modeling in Ecology: Analysis of distribution, abundance and species richness in R and BUGS: Volume 1: Prelude and Static Models. Academic Press.

Kéry, M., and B. Schmidt. 2008. Imperfect detection and its consequences for monitoring for conservation. Community Ecology 9:207–216.

Kie, J. G., R. T. Bowyer, M. C. Nicholson, B. B. Boroski, and E. R. Loft. 2002. Landscape heterogeneity at differing scales: effects on spatial distribution of mule deer. Ecology 83:530–544.

Koenig, W. D. 2001. Spatial autocorrelation and local disappearances in wintering North American birds. Ecology 82:2636–2644.

Krebs, C. J. 1994. Ecology: the experimental analysis of distribution and abundance. HarperCollins College Publishers New York.

Laboratory, N.-R. A. 2015. verification: Weather Forecast Verification Utilities.

Larivière, S. 2001. Ursus americanus. Mammalian species 2001:1–11.

Lasky, M., A. W. Parsons, S. Schuttler, A. Mash, L. Larson, B. Norton, B. S. Pease, H. Boone, L. Gatens, and R. Kays. In Review. Studying Large-Scale Ecological Questions with Camera Traps and Citizen Scientists. Citizen Science: Theory and Practice.

Latham, A. D. M., and S. Boutin. 2011. Wolf, Canis lupus, pup mortality: interspecific predation or non-parental infanticide? The Canadian Field-Naturalist 125:158–161.

Lewis, S. L., and M. A. Maslin. 2015. Defining the anthropocene. Nature 519:171–180.

Lister, A. M., and A. J. Stuart. 2008. The impact of climate change on large mammal distribution and extinction: evidence from the last glacial/interglacial transition. Comptes Rendus Geoscience 340:615–620.

Lloyd, C. D. 2010a. Local models for spatial analysis. CRC press.

Lloyd, C. D. 2010b. Local models for spatial analysis. CRC press.

Loiselle, B. A., C. A. Howell, C. H. Graham, J. M. Goerck, T. Brooks, K. G. Smith, and P. H. Williams. 2003. Avoiding pitfalls of using species distribution models in conservation planning. Conservation biology 17:1591–1600.

Lombardi, J. V, C. E. Comer, D. G. Scognamillo, and W. C. Conway. 2017. Coyote, fox, and bobcat response to anthropogenic and natural landscape features in a small urban area. Urban Ecosystems 20:1239–1248.

Long, R. A., T. M. Donovan, P. MacKay, W. J. Zielinski, and J. S. Buzas. 2011. Predicting carnivore occurrence with noninvasive surveys and occupancy modeling. Landscape Ecology 26:327–340.

Lonsinger, R. C., E. M. Gese, and L. P. Waits. 2015. Evaluating the reliability of field identification and morphometric classifications for carnivore scats confirmed with genetic analysis. Wildlife Society Bulletin 39:593–602.

Ma, Z., B. Zuckerberg, W. F. Porter, and L. Zhang. 2012. Spatial Poisson models for examining the influence of climate and land cover pattern on bird species richness. Forest Science 58:61–74.

MacKenzie, D. I., J. D. Nichols, J. A. Royle, K. H. Pollock, L. Bailey, and J. E. Hines. 2017a. Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence. Elsevier.

MacKenzie, D. I., J. D. Nichols, J. A. Royle, K. H. Pollock, L. Bailey, and J. E. Hines. 2017b. Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence.

Maldonado, C., C. I. Molina, A. Zizka, C. Persson, C. M. Taylor, J. Albán, E. Chilquillo, N. Rønsted, and A. Antonelli. 2015. Estimating species diversity and distribution in the era of B ig D ata: to what extent can we trust public databases? Global Ecology and Biogeography 24:973–984.

Manjerovic, M. B., M. L. Green, N. Mateus-Pinilla, and J. Novakofski. 2014. The importance of localized culling in stabilizing chronic wasting disease prevalence in white-tailed deer populations. Preventive Veterinary Medicine 113:139–145.

Marcer, A., B. Méndez-Vigo, C. Alonso-Blanco, and F. X. Picó. 2016. Tackling intraspecific genetic structure in distribution models better reflects species geographical range. Ecology and Evolution 6:2084–2097.

Matthiopoulos, J., J. Fieberg, and G. Aarts. 2020. Species-Habitat Associations: Spatial data, predictive models, and ecological insights.

McLellan, B. N. 2011. Implications of a high-energy and low-protein diet on the body composition, fitness, and competitive abilities of black (Ursus americanus) and grizzly (Ursus arctos) bears. Canadian Journal of Zoology 89:546–558.

McLellan, B., M. Proctor, and S. Michel. 2017. Ursus arctos (amended version of 2017 assessment). The IUCN Red List of Threatened Species e.T41688A121229971.

McMillen, D. P. 2004. Geographically weighted regression: the analysis of spatially varying relationships. Oxford University Press.

McNew, L. B., A. J. Gregory, and B. K. Sandercock. 2013. Spatial heterogeneity in habitat selection: Nest site selection by greater prairie-chickens. The Journal of wildlife management 77:791–801.

McShea, W. J., T. Forrester, R. Costello, Z. He, and R. Kays. 2016. Volunteer-run cameras as distributed sensors for macrosystem mammal research. Landscape Ecology 31:55–66.

Meehan, T. D., N. L. Michel, and H. Rue. 2019. Spatial modeling of Audubon Christmas Bird Counts reveals fine-scale patterns and drivers of relative abundance trends. Ecosphere 10:e02707.

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

Miller, J. 2010. Species distribution modeling. Geography Compass 4:490–509.

Miller, J. A. 2012a. Species distribution models: Spatial autocorrelation and non-stationarity. Progress in Physical Geography 36:681–692.

Miller, J. A. 2012b. Species distribution models: Spatial autocorrelation and non-stationarity. Progress in Physical Geography 36:681–692.

Moritz, C., J. L. Patton, C. J. Conroy, J. L. Parra, G. C. White, and S. R. Beissinger. 2008. Impact of a century of climate change on small-mammal communities in Yosemite National Park, USA. Science 322:261–264.

Müller, K. 2017. here: A Simpler Way to Find Your Files.

Myers, E. A., A. D. McKelvy, and F. T. Burbrink. 2020. Biogeographic barriers, Pleistocene refugia, and climatic gradients in the southeastern Nearctic drive diversification in cornsnakes (Pantherophis guttatus complex). Molecular Ecology 29:797–811.

Neuwirth, E. 2014. RColorBrewer: ColorBrewer Palettes.

Newbold, T. 2010. Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. Progress in Physical Geography 34:3–22.

Nielsen, C. K., and A. Woolf. 2002. Habitat-relative abundance relationship for bobcats in southern Illinois. Wildlife Society Bulletin:222–230.

Odum, E. P., and G. W. Barrett. 1971. Fundamentals of ecology. Saunders Philadelphia.

Omernik, J. M. 1987. Ecoregions of the conterminous United States. Annals of the Association of American geographers 77:118–125.

Omernik, J. M., and G. E. Griffith. 2014. Ecoregions of the conterminous United States: evolution of a hierarchical spatial framework. Environmental management 54:1249–1266.

Osborne, P. E., G. M. Foody, and S. Suárez-Seoane. 2007a. Non-stationarity and local approaches to modelling the distributions of wildlife. Diversity and Distributions 13:313–323.

Osborne, P. E., G. M. Foody, and S. Suárez-Seoane. 2007b. Non-stationarity and local approaches to modelling the distributions of wildlife. Diversity and Distributions 13:313–323.

Osborne, P. E., and S. Sua. 2002. Should data be partitioned spatially before building large-scale distribution models? 157.

Pacifici, K., B. J. Reich, R. M. Dorazio, and M. J. Conroy. 2016. Occupancy estimation for rare species using a spatially-adaptive sampling design. Methods in Ecology and Evolution 7:285–293.

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

Pacifici, K., B. J. Reich, D. A. W. Miller, and B. S. Pease. 2019. Resolving misaligned spatial data with integrated species distribution models. Ecology:e02709.

Parsons, A. W., T. Forrester, M. C. Baker-Whatton, W. J. McShea, C. T. Rota, S. G. Schuttler, J. J. Millspaugh, and R. Kays. 2018. Mammal communities are larger and more diverse in moderately developed areas. eLife 7:e38012.

Parsons, A. W., C. T. Rota, T. Forrester, M. C. Baker-Whatton, W. J. McShea, S. G. Schuttler, J. J. Millspaugh, and R. Kays. 2019. Urbanization focuses carnivore activity in remaining natural habitats, increasing species interactions. Journal of Applied Ecology 56:1894–1904.

Pease, B. S., R. Kays, and K. Pacifici. In Review. Do wildlife-habitat relationships vary across space?

Pebesma, E. 2018. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 10:439–446.

Pelton, M. R. 2003. Black bear. Wild Mammals of North America: Biology, Management and Conversation:547–555.

Penteriani, V., and M. Melletti. 2020. Bears of the World: Ecology, Conservation and Management. Cambridge University Press.

Peterson, A. T. 2011. Ecological niche conservatism: A time-structured review of evidence. Journal of Biogeography 38:817–827.

Peterson, A. T., J. Soberón, and V. Sánchez-Cordero. 1999. Conservatism of ecological niches in evolutionary time. Science 285:1265–1267.

Prange, S., S. D. Gehrt, and E. P. Wiggers. 2004. Influences of anthropogenic resources on raccoon (Procyon lotor) movements and spatial distribution. Journal of Mammalogy 85:483–490.

Puckett, E. E., P. D. Etter, E. A. Johnson, and L. S. Eggert. 2015. Phylogeographic analyses of American black bears (Ursus americanus) suggest four glacial refugia and complex patterns of postglacial admixture. Molecular biology and evolution 32:2338–2350.

R Core Team. 2020a. R: A Language and Environment for Statistical Comuputing. R Foundation for Statistical Computing, Vienna, Austria.

R Core Team. 2020b. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Reichman, O. J., M. B. Jones, and M. P. Schildhauer. 2011. Challenges and opportunities of open data in ecology. Science 331:703–705.

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

Renner, I. W., and D. I. Warton. 2013. Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. Biometrics 69:274–281.

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

Roemer, G. W., M. E. Gompper, and B. Van Valkenburgh. 2009. The ecological role of the mammalian mesocarnivore. BioScience 59:165–173.

Roughgarden, J. 1972. Evolution of niche width. The American Naturalist 106:683–718.

Runge, M. C., S. J. Converse, J. E. Lyons, and D. R. Smith, editors. 2020. Structured Decision Making: Case Studies in Natural Resource Management. Johns Hopkins University Press.

Saïd, S., and S. Servanty. 2005. The influence of landscape structure on female roe deer home-range size. Landscape ecology 20:1003–1012.

Sargeant, B. L. 2007. Individual foraging specialization: niche width versus niche overlap. Oikos 116:1431–1437.

Schank, C. J., M. V. Cove, M. J. Kelly, E. Mendoza, G. O'Farrill, R. Reyna-Hurtado, N. Meyer, C. A. Jordan, J. F. González-Maya, and D. J. Lizcano. 2017. Using a novel model approach to assess the distribution and conservation status of the endangered Baird's tapir. Diversity and Distributions 23:1459–1471.

Schaub, M., and F. Abadi. 2011. Integrated population models: a novel analysis framework for deeper insights into population dynamics. Journal of Ornithology 152:227–237.

Schimel, D. 2011. The era of continental-scale ecology. Frontiers in Ecology and the Environment 9:311.

Schoener, T. W. 1986. Resource partitioning. Community ecology pattern and process.

Schwartz, C. C., and A. W. Franzmann. 1991. Interrelationship of black bears to moose and forest succession in the northern coniferous forest. Wildlife Monographs:3–58.

Serenari, C., J. Shaw, R. Myers, and D. T. Cobb. 2019. Explaining deer hunter preferences for regulatory changes using choice experiments. The Journal of Wildlife Management 83:446–456.

Sexton, J. P., P. J. McIntyre, A. L. Angert, and K. J. Rice. 2009. Evolution and ecology of species range limits. Annu. Rev. Ecol. Evol. Syst. 40:415–436.

Shaw, J., E. Stanford, R. Myers, B. Howard, D. Sawyer, and D. T. Cobb. 2016. Delineating biological deer management units in North Carolina. Raleigh, North Carolina.

Shi, H., E. J. Laurent, J. LeBouton, L. Racevskis, K. R. Hall, M. Donovan, R. V Doepker, M. B. Walters, F. Lupi, and J. Liu. 2006. Local spatial modeling of white-tailed deer distribution. Ecological Modelling 190:171–189.

Simmonds, E. G., S. G. Jarvis, P. A. Henrys, N. J. Isaac, and R. B. O'Hara. 2020. Is more data always better? A simulation study of benefits and limitations of integrated distribution models. Ecography.

Smith, A. B., E. A. Beever, A. E. Kessler, A. N. Johnston, C. Ray, C. W. Epps, H. C. Lanier, R. C. Klinger, T. J. Rodhouse, and J. Varner. 2019a. Alternatives to genetic affinity as a context for within-species response to climate. Nature Climate Change 9:787–794.

Smith, A. B., E. A. Beever, A. E. Kessler, A. N. Johnston, C. Ray, C. W. Epps, H. C. Lanier, R. C. Klinger, T. J. Rodhouse, J. Varner, J. D. Perrine, A. Seglund, L. E. Hall, K. Galbreath,

C. MacGlover, P. Billman, G. Blatz, J. Brewer, J. Castillo Vardaro, A. D. Chalfoun, G. Collins, A. Craighead, C. Curlis, C. Daly, D. F. Doak, M. East, M. Edwards, L. Erb, K. A. Ernest, B. Fauver, K. R. Foresman, K. Goehring, J. Hagar, C. L. Hayes, P. Henry, K. Hersey, S. L. Hilty, J. Jacobson, M. R. Jeffress, T. Manning, A. Masching, B. Maxell, R. McCollough, C. McFarland, E. Miskow, T. L. Morelli, L. Moyer-Horner, M. Mueller, M. Nugent, B. Pratt, M. Rasmussen-Flores, T. H. Rickman, H. Robison, A. Rodriguez, K. Rowe, K. Rowe, M. A. Russello, V. Saab, A. Schmidt, J. A. E. Stewart, J. N. Stuart, L. K. Svancara, W. Thompson, J. Timmins, G. Treinish, M. D. Waterhouse, M. L. Westover, J. Wilkening, and L. Yandow. 2019b. Alternatives to genetic affinity as a context for within-species response to climate .

Smith, A. B., W. Godsoe, F. Rodríguez-Sánchez, H.-H. Wang, and D. Warren. 2019c. Niche estimation above and below the species level. Trends in ecology & evolution 34:260–273.

Sulla-Menashe, D., and M. A. Friedl. 2018. User guide to collection 6 MODIS land cover (MCD12Q1 and MCD12C1) product.

Tennekes, M. 2018. {tmap}: Thematic Maps in {R}. Journal of Statistical Software 84:1–39.

Thomas, D. L., and E. J. Taylor. 1990. Study designs and tests for comparing resource use and availability. The Journal of wildlife management:322–330.

Tigas, L. A., D. H. Van Vuren, and R. M. Sauvajot. 2002. Behavioral responses of bobcats and coyotes to habitat fragmentation and corridors in an urban environment. Biological Conservation 108:299–306.

Toomey, A., M. Weckel, C. Nagy, L. Gormezano, and S. Silver. 2012. The last frontier: eastern coyotes in New York City. Wildlife Professional 6:54–57.

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 26:403–413.

Vehtari, A., J. Gabry, M. Magnusson, Y. Yao, P. Burkner, T. Paananen, and A. Gelman. 2020. loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models.

Walter, W. D., K. C. VerCauteren, H. Campa, W. R. Clark, J. W. Fischer, S. E. Hygnstrom, N. E. Mathews, C. K. Nielsen, E. M. Schauber, T. R. Van Deelen, and others. 2009. Regional assessment on influence of landscape configuration and connectivity on range size of white-tailed deer. Landscape Ecology 24:1405–1420.

Warton, D. I., I. W. Renner, and D. Ramp. 2013. Model-based control of observer bias for the analysis of presence-only data in ecology. PloS one 8:e79168.

Wellborn, G. A., and R. B. Langerhans. 2015. Ecological opportunity and the adaptive diversification of lineages. Ecology and evolution 5:176–195.

Wheeler, D. C., and L. A. Waller. 2009. Comparing spatially varying coefficient models: a case study examining violent crime rates and their relationships to alcohol outlets and illegal drug arrests. Journal of Geographical Systems 11:1–22.

Wheeler, D., and M. Tiefelsdorf. 2005. Multicollinearity and correlation among local regression coefficients in geographically weighted regression. Journal of Geographical Systems 7:161–187.

Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wickham, H. 2019. stringr: Simple, Consistent Wrappers for Common String Operations.

Wickham, H., M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller,

J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani. 2019. Welcome to the {tidyverse}. Journal of Open Source Software 4:1686.

Wiens, J. J., D. D. Ackerly, A. P. Allen, B. L. Anacker, L. B. Buckley, H. V. Cornell, E. I. Damschen, T. Jonathan Davies, J. Grytnes, and S. P. Harrison. 2010. Niche conservatism as an emerging principle in ecology and conservation biology. Ecology letters 13:1310–1324.

Wiens, J. J., and C. H. Graham. 2005. Niche conservatism: integrating evolution, ecology, and conservation biology. Annu. Rev. Ecol. Evol. Syst. 36:519–539.

Winchcombe, R. J., and R. S. Ostfeld. 2001. Bowhunter observations versus spotlighting as an index to deer abundance. Northeast Wildlife 56:39–48.

Windle, M. J. S., G. A. Rose, R. Devillers, and M.-J. Fortin. 2009. Exploring spatial non-stationarity of fisheries survey data using geographically weighted regression (GWR): an example from the Northwest Atlantic. ICES Journal of Marine Science 67:145–154.

Wood, S. N. 2017. Generalized additive models: an introduction with R. Chapman and Hall/CRC.

Woolf, A., C. K. Nielsen, T. Weber, and T. J. Gibbs-Kieninger. 2002. Statewide modeling of bobcat, Lynx rufus, habitat in Illinois, USA. Biological Conservation 104:191–198.

Yackulic, C. B., R. Chandler, E. F. Zipkin, J. A. Royle, J. D. Nichols, E. H. Campbell Grant, and S. Veran. 2013. Presence-only modelling using MAXENT: when can we trust the inferences? Methods in Ecology and Evolution 4:236–243.

Young, J. K., J. M. Golla, D. Broman, T. Blankenship, and R. Heilbrun. 2019. Estimating density of an elusive carnivore in urban areas: use of spatially explicit capture-recapture models for city-dwelling bobcats. Urban Ecosystems 22:507–512.

Zipkin, E. F., E. H. C. Grant, and W. F. Fagan. 2012. Evaluating the predictive abilities of community occupancy models using AUC while accounting for imperfect detection. Ecological Applications 22:1962–1972.

APPENDICES

Appendix A: Chapter 1 Supplementary Information

Table A1: Estimated coefficients across each species, model, and ecological covariate. SD = standard deviation; Lower 95 CI = lower bound of 95% credible interval; Upper 95 CI = upper bound of 95% credible interval.

Species	Covariate	Model	Region	Mean	SD	Lower 95 CI	Upper 95 CI
Bobcat	Contagion	Ecoregion	Coastal	0.077	0.384	-0.649	0.840
Bobcat	Contagion	Ecoregion	Mountains	0.266	0.457	-0.609	1.193
Bobcat	Contagion	Ecoregion	Piedmont	-0.343	0.460	-1.284	0.533
Bobcat	PRD	Ecoregion	Coastal	0.808	0.426	-0.053	1.630
Bobcat	PRD	Ecoregion	Mountains	-0.740	0.514	-1.774	0.208
Bobcat	PRD	Ecoregion	Piedmont	-0.068	0.521	-1.069	0.927
Bobcat	Forest	Ecoregion	Coastal	-0.071	0.431	-0.952	0.766
Bobcat	Forest	Ecoregion	Mountains	0.924	0.361	0.226	1.669
Bobcat	Forest	Ecoregion	Piedmont	-0.852	0.410	-1.697	-0.103
Bobcat	Developed	Ecoregion	Coastal	-0.512	0.528	-1.532	0.506
Bobcat	Developed	Ecoregion	Mountains	1.488	0.656	0.360	2.881
Bobcat	Developed	Ecoregion	Piedmont	-0.976	0.400	-1.890	-0.287
Bobcat	Contagion	GLM	Global	-0.294	0.350	-1.034	0.282
Bobcat	PRD	GLM	Global	-0.159	0.304	-0.756	0.404
Bobcat	Forest	GLM	Global	0.295	0.245	-0.234	0.766
Bobcat	Developed	GLM	Global	-0.996	0.378	-1.837	-0.360
Coyote	Contagion	Ecoregion	Coastal	-0.380	0.280	-0.882	0.221

Table A1 (continued).

Coyote	Contagion	Ecoregion	Mountains	0.208	0.389	-0.557	1.030
Coyote	Contagion	Ecoregion	Piedmont	0.172	0.373	-0.572	0.850
Coyote	PRD	Ecoregion	Coastal	0.449	0.313	-0.107	1.121
Coyote	PRD	Ecoregion	Mountains	-0.587	0.415	-1.390	0.187
Coyote	PRD	Ecoregion	Piedmont	0.138	0.440	-0.752	0.995
Coyote	Forest	Ecoregion	Coastal	-0.573	0.346	-1.227	0.096
Coyote	Forest	Ecoregion	Mountains	-0.214	0.302	-0.834	0.369
Coyote	Forest	Ecoregion	Piedmont	0.787	0.338	0.148	1.460
Coyote	Developed	Ecoregion	Coastal	-0.675	0.307	-1.314	-0.093
Coyote	Developed	Ecoregion	Mountains	0.289	0.308	-0.304	0.904
Coyote	Developed	Ecoregion	Piedmont	0.385	0.170	0.072	0.727
Coyote	Contagion	GLM	Global	0.340	0.196	-0.027	0.733
Coyote	PRD	GLM	Global	0.199	0.237	-0.235	0.678
Coyote	Forest	GLM	Global	0.687	0.222	0.286	1.156
Coyote	Developed	GLM	Global	0.282	0.126	0.053	0.547
Raccoon	Contagion	Ecoregion	Coastal	0.298	0.262	-0.233	0.809
Raccoon	Contagion	Ecoregion	Mountains	-0.414	0.292	-0.998	0.153
Raccoon	Contagion	Ecoregion	Piedmont	0.117	0.305	-0.470	0.732
Raccoon	PRD	Ecoregion	Coastal	0.076	0.296	-0.508	0.644
Raccoon	PRD	Ecoregion	Mountains	0.191	0.265	-0.319	0.724
Raccoon	PRD	Ecoregion	Piedmont	-0.266	0.330	-0.914	0.353
Raccoon	Forest	Ecoregion	Coastal	0.012	0.306	-0.564	0.644

Table A1 (continued).

Raccoon	Forest	Ecoregion	Mountains	-0.463	0.257	-0.981	0.032
Raccoon	Forest	Ecoregion	Piedmont	0.452	0.299	-0.108	1.040
Raccoon	Developed	Ecoregion	Coastal	-0.260	0.234	-0.706	0.206
Raccoon	Developed	Ecoregion	Mountains	-0.260	0.228	-0.706	0.175
Raccoon	Developed	Ecoregion	Piedmont	0.519	0.177	0.187	0.879
Raccoon	Contagion	GLM	Global	0.166	0.247	-0.276	0.747
Raccoon	PRD	GLM	Global	-0.018	0.243	-0.475	0.472
Raccoon	Forest	GLM	Global	-0.600	0.283	-1.284	-0.141
Raccoon	Developed	GLM	Global	0.128	0.174	-0.185	0.503
Deer	Contagion	Ecoregion	Coastal	0.365	0.346	-0.298	1.046
Deer	Contagion	Ecoregion	Mountains	0.409	0.371	-0.314	1.136
Deer	Contagion	Ecoregion	Piedmont	-0.774	0.435	-1.645	0.035
Deer	PRD	Ecoregion	Coastal	0.194	0.343	-0.485	0.859
Deer	PRD	Ecoregion	Mountains	-0.114	0.311	-0.740	0.455
Deer	PRD	Ecoregion	Piedmont	-0.080	0.404	-0.898	0.720
Deer	Forest	Ecoregion	Coastal	-0.191	0.483	-1.236	0.630
Deer	Forest	Ecoregion	Mountains	-1.257	0.491	-2.335	-0.452
Deer	Forest	Ecoregion	Piedmont	1.448	0.631	0.542	3.081
Deer	Developed	Ecoregion	Coastal	0.301	0.341	-0.387	0.955
Deer	Developed	Ecoregion	Mountains	-0.454	0.330	-1.122	0.126
Deer	Developed	Ecoregion	Piedmont	0.153	0.225	-0.255	0.653
Deer	Contagion	GLM	Global	0.113	0.230	-0.321	0.597

Table A1 (continued).

Deer	PRD	GLM	Global	0.185	0.218	-0.259	0.604
Deer	Forest	GLM	Global	0.315	0.240	-0.121	0.830
Deer	Developed	GLM	Global	-0.188	0.143	-0.468	0.105

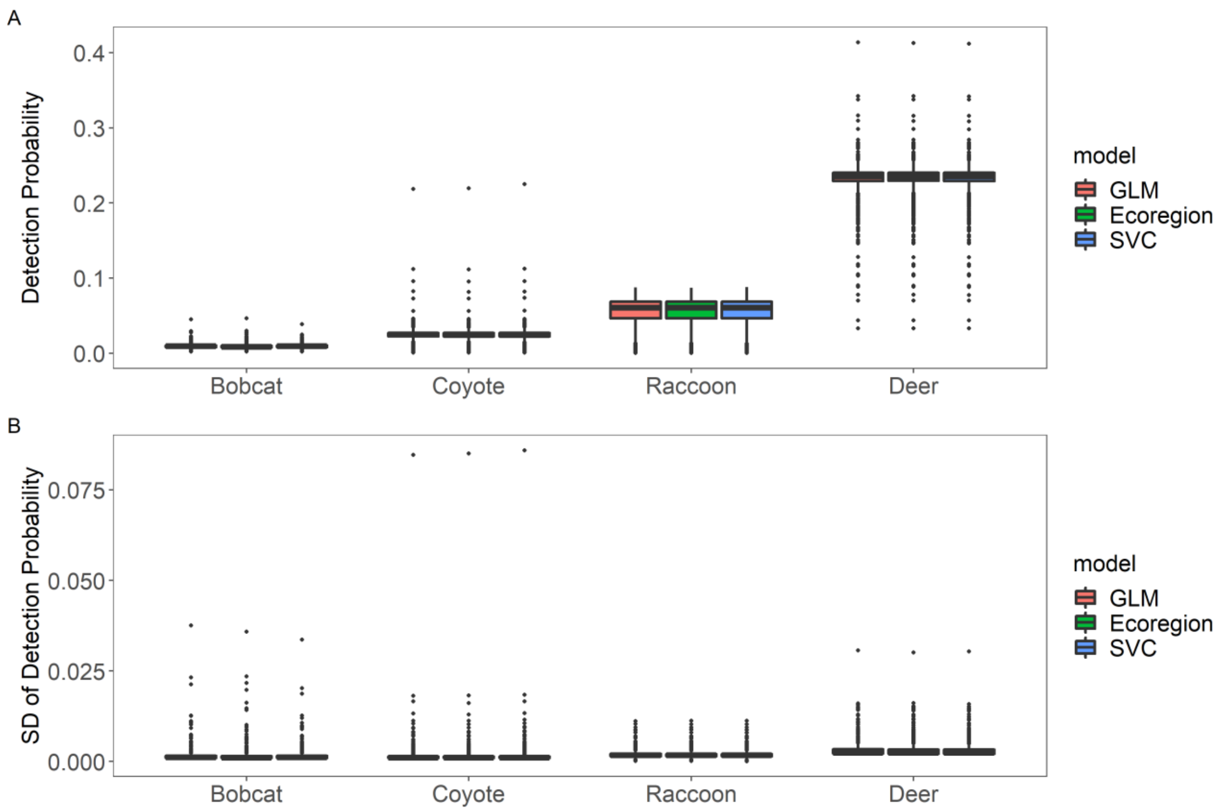


Figure A1: Grid cell-level estimated detection probability for each species across three modeling approaches.

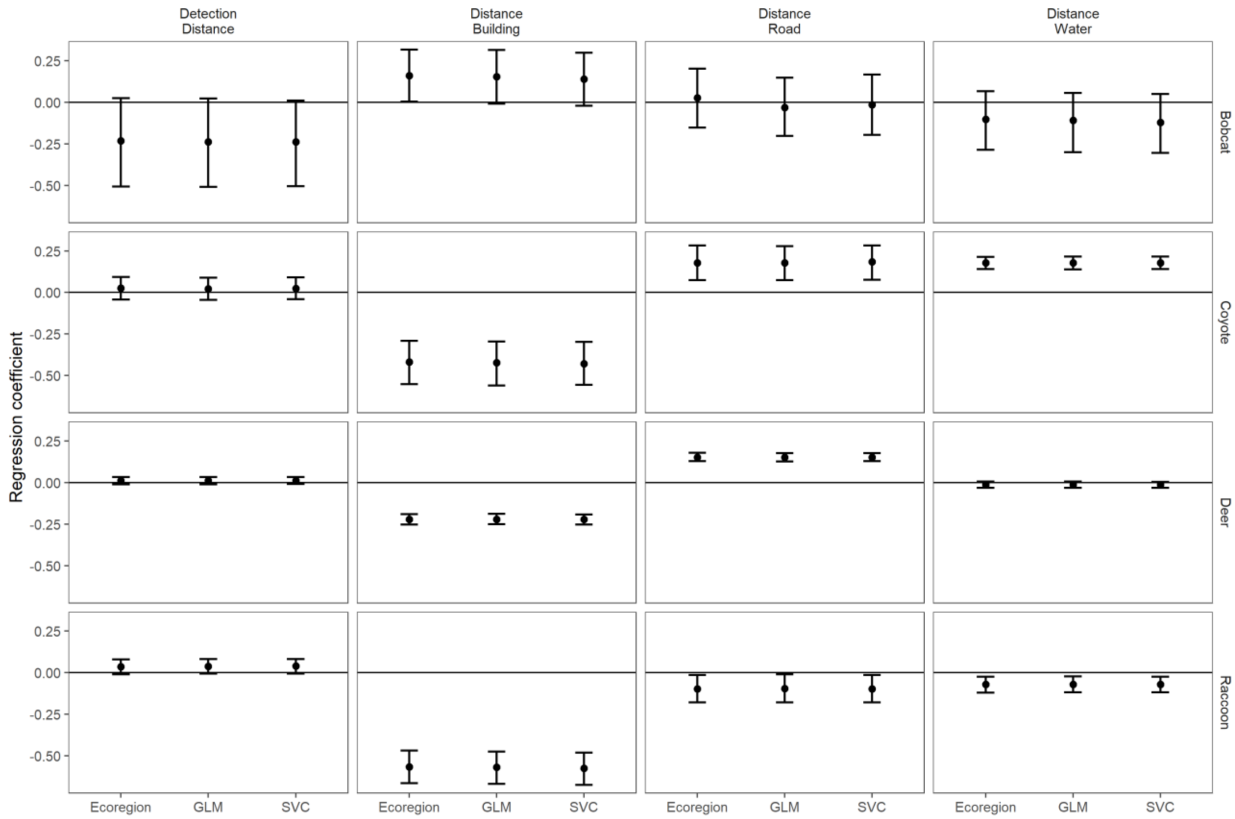


Figure A2: Estimated effects of observation covariates on the detection probably across each focal species and three modeling approaches. Solid line indicates zero (no effect), dots are point estimates and bars represent 95% credible intervals.

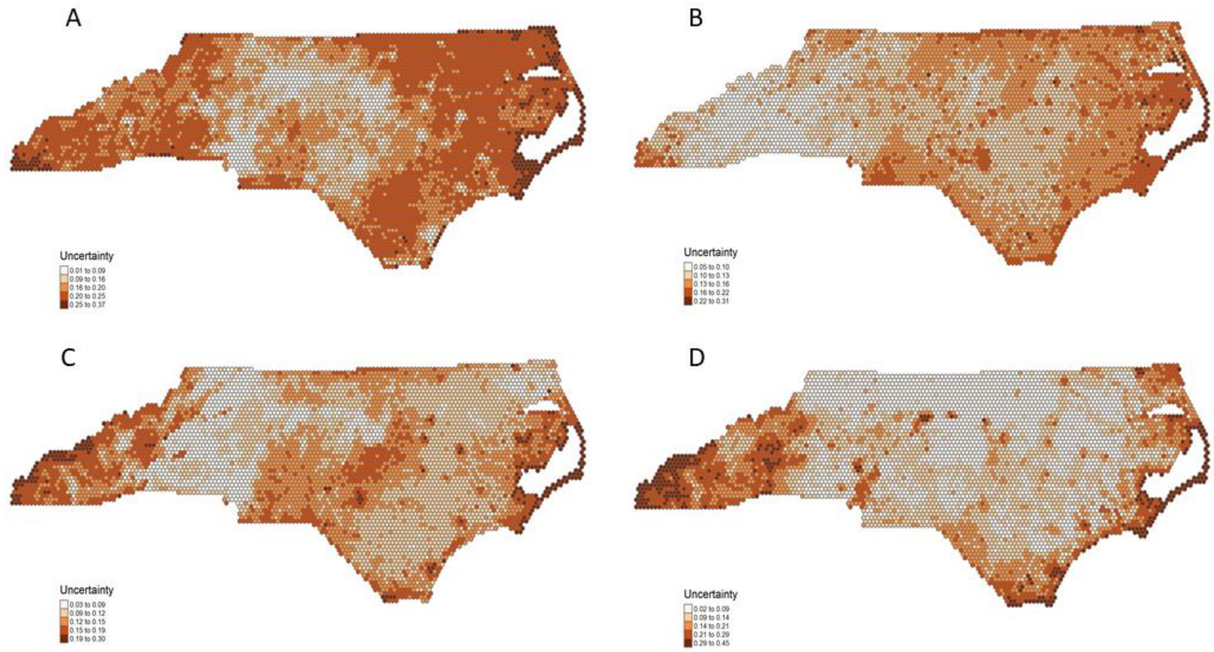


Figure A3: Occupancy probability uncertainty (standard deviation) for A) bobcat, B) coyote, C) raccoon, and D) white-tailed deer during Oct 2015 – Dec 2019 in North Carolina, USA.

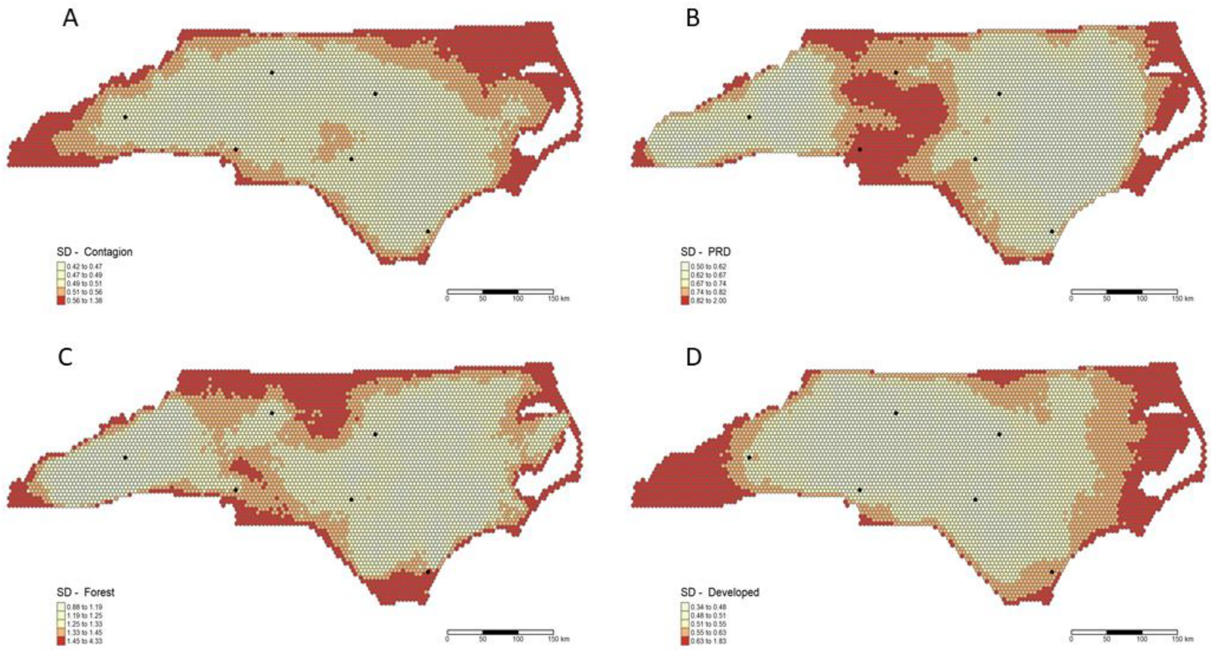


Figure A4: Uncertainty (standard deviation) in spatially varying coefficient estimates of (A) contagion index (SI = 0.85), (B) patch richness density (SI = 0.47), (C) forest cover (SI = 2.13), and (D) developed (SI = 0.91) for white-tailed deer during Oct 2015 – Dec 2019 in North Carolina, USA. Black dots are for spatial reference and represent cities with populations with 100,000 people or more.

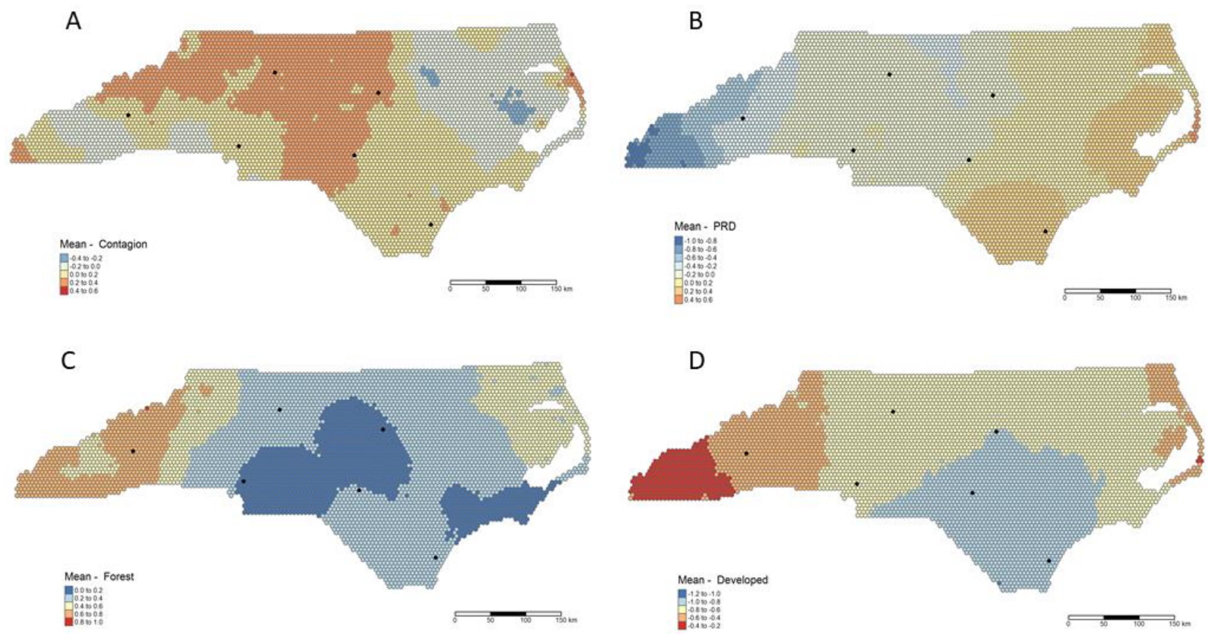


Figure A5: Spatially varying coefficient estimates (and stationary indices) of (A) contagion index (SI = 0.85), (B) patch richness density (SI = 0.47), (C) forest cover (SI = 2.13), and (D) developed (SI = 0.91) for bobcat (*Lynx rufus*) during Oct 2015 – Dec 2019 in North Carolina, USA. Black dots are for spatial reference and represent cities with populations with 100,000 people or more.

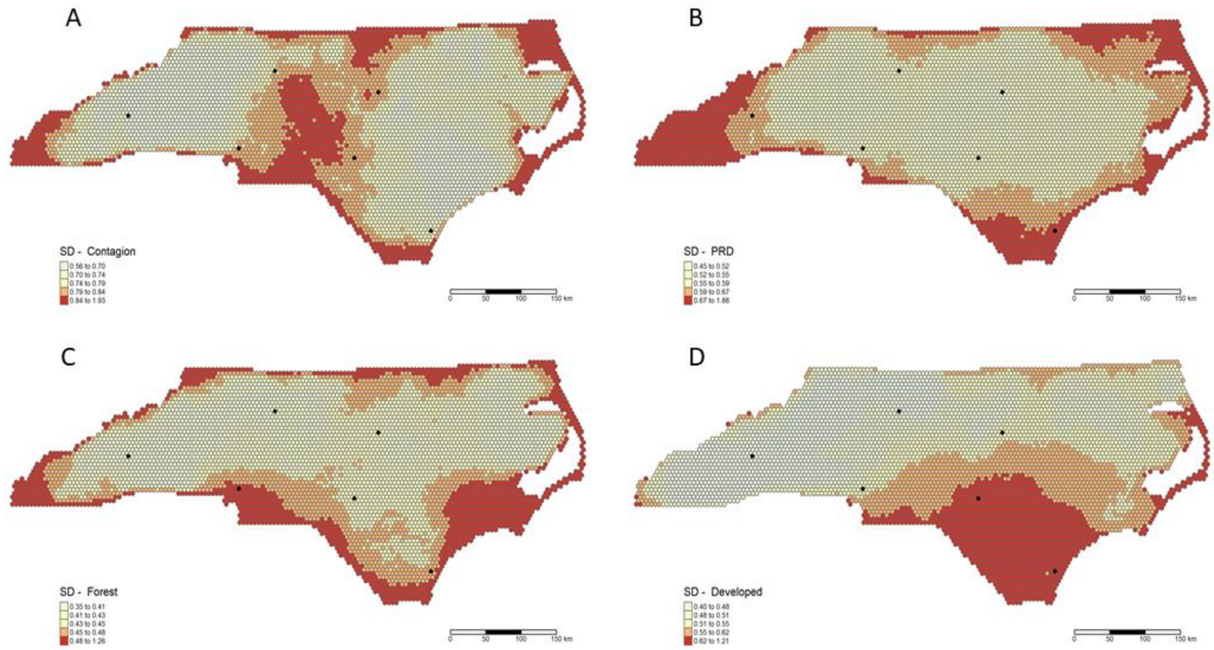


Figure A6: Uncertainty (standard deviation) in spatially varying coefficient estimates of (A) contagion index (SI = 0.85), (B) patch richness density (SI = 0.47), (C) forest cover (SI = 2.13), and (D) developed (SI = 0.91) for bobcat (*Lynx rufus*) during Oct 2015 – Dec 2019 in North Carolina, USA. Black dots are for spatial reference and represent cities with populations with 100,000 people or more.

Appendix B: Chapter 2 Supplementary Information

Table B1: Details of outlier removal process for each species/data source combination. For each data source, we report on the minimum catch per unit effort (CPUE) value that was considered an outlier based on 1.5 × the interquartile range (IQR) of the observed values (herein, "outlier value"). We also report the number of observations that was greater than the outlier value alongside a comparison of the mean CPUE before and after outlier removal. *No outliers detected.

Data Source	Value	Bobcat	Coyote	Raccoon
Cameras	Outlier Value	0.158	0.172	0.364
	No. of Observations Removed	11	90	93
	Mean CPUE Before Removal	0.070	0.088	0.137
	Mean CPUE After Removal	0.063	0.066	0.107
Hunter Harvest	Outlier Value	-*	2.333	2.250
	No. of Observations Removed	0	248	41
	Mean CPUE Before Removal	0.724	0.577	0.680
	Mean CPUE After Removal	0.724	0.537	0.636
Trapper Harvest	Outlier Value	0.078	0.082	0.256
	No. of Observations Removed	284	199	849
	Mean CPUE Before Removal	0.042	0.037	0.105
	Mean CPUE After Removal	0.017	0.023	0.066
Hunter Observations	Outlier Value	24.000	21.333	36.923
	No. of Observations Removed	18	560	507
	Mean CPUE Before Removal	8.647	11.247	16.467

Table B1 (continued)

	Mean CPUE After Removal	8.193	8.380	13.022

Table B2: Comparison of deviance explained (more is better) and regression relationships between each data source combination and filtering mechanism. The filtering type with the maximum deviance explained in each Species/Unstructured Data Source combination was selected as the preferred filtering mechanism.

Species	Data Source 1	Data Source 2	Data Source 3	Data Source 4	Filtering Type	Deviance Explained	Regression Coefficient 1	Regression Coefficient 2	Regression Coefficient 3
Bobcat	Cameras	Hunter Harvest	-	-	Distribution Filtering	0.07	0.05	-	-
Bobcat	Cameras	Hunter Harvest	-	-	No Filtering	0.07	0.05	-	-
Bobcat	Cameras	Hunter Harvest	-	-	Observation Filtering	0	-	-	-
Bobcat	Cameras	Hunter Harvest	Hunter Observations	-	Distribution Filtering	24.59	-0.06	8.99	-
Bobcat	Cameras	Hunter Harvest	Hunter Observations	-	No Filtering	48.49	-0.05	7.28	-
Bobcat	Cameras	Hunter Harvest	Hunter Observations	-	Observation Filtering	0	-	-	-
Bobcat	Cameras	Hunter Harvest	Hunter Observations	Trapper Harvest	Distribution Filtering	41.75	-0.16	4.06	117.52
Bobcat	Cameras	Hunter Harvest	Hunter Observations	Trapper Harvest	No Filtering	99.22	0.05	1.88	104.71
Bobcat	Cameras	Hunter Harvest	Hunter Observations	Trapper Harvest	Observation Filtering	0	-	-	-
Bobcat	Cameras	Hunter Harvest	Trapper Harvest	-	Distribution Filtering	99.8	0.08	118.57	-
Bobcat	Cameras	Hunter Harvest	Trapper Harvest	-	No Filtering	96.58	0.07	115.62	-
Bobcat	Cameras	Hunter Harvest	Trapper Harvest	-	Observation Filtering	0	-	-	-

Table B2 (continued).

Bobcat	Cameras	Hunter Observations	-	-	Distribution Filtering	24.53	8.85	-	-
Bobcat	Cameras	Hunter Observations	-	-	No Filtering	48.42	7.29	-	-
Bobcat	Cameras	Hunter Observations	-	-	Observation Filtering	33.14	8.26	-	-
Bobcat	Cameras	Hunter Observations	Trapper Harvest	-	Distribution Filtering	41.36	3.69	117	-
Bobcat	Cameras	Hunter Observations	Trapper Harvest	-	No Filtering	99.14	1.9	104.21	-
Bobcat	Cameras	Hunter Observations	Trapper Harvest	-	Observation Filtering	37.22	0.32	114.56	-
Bobcat	Cameras	Trapper Harvest	-	-	Distribution Filtering	99.64	118.02	-	-
Bobcat	Cameras	Trapper Harvest	-	-	No Filtering	96.44	115.12	-	-
Bobcat	Cameras	Trapper Harvest	-	-	Observation Filtering	58.5	112.32	-	-
Coyote	Cameras	Hunter Harvest	-	-	Distribution Filtering	11.71	1.36	-	-
Coyote	Cameras	Hunter Harvest	-	-	No Filtering	2.46	0.54	-	-
Coyote	Cameras	Hunter Harvest	-	-	Observation Filtering	13.28	1.43	-	-
Coyote	Cameras	Hunter Harvest	Hunter Observations	-	Distribution Filtering	2.01	0.79	0.05	-
Coyote	Cameras	Hunter Harvest	Hunter Observations	-	No Filtering	10.41	0.58	-0.66	-
Coyote	Cameras	Hunter Harvest	Hunter Observations	-	Observation Filtering	11.99	1.31	-0.3	-
Coyote	Cameras	Hunter Harvest	Hunter Observations	Trapper Harvest	Distribution Filtering	2.04	0.77	0.04	1.05

Table B2 (continued).

Coyote	Cameras	Hunter Harvest	Hunter Observations	Trapper Harvest	No Filtering	11.34	0.67	-0.59	-4.98
Coyote	Cameras	Hunter Harvest	Hunter Observations	Trapper Harvest	Observation Filtering	12.06	1.29	-0.31	1.43
Coyote	Cameras	Hunter Harvest	Trapper Harvest	-	Distribution Filtering	11.73	1.36	-0.62	-
Coyote	Cameras	Hunter Harvest	Trapper Harvest	-	No Filtering	5.65	0.71	-8.77	-
Coyote	Cameras	Hunter Harvest	Trapper Harvest	-	Observation Filtering	13.28	1.43	-0.23	-
Coyote	Cameras	Hunter Observations	-	-	Distribution Filtering	0.21	0.14	-	-
Coyote	Cameras	Hunter Observations	-	-	No Filtering	7.6	-0.65	-	-
Coyote	Cameras	Hunter Observations	-	-	Observation Filtering	1.14	-0.28	-	-
Coyote	Cameras	Hunter Observations	Trapper Harvest	-	Distribution Filtering	0.63	0.09	3.88	-
Coyote	Cameras	Hunter Observations	Trapper Harvest	-	No Filtering	7.82	-0.62	-2.37	-
Coyote	Cameras	Hunter Observations	Trapper Harvest	-	Observation Filtering	2.34	-0.35	5.78	-
Coyote	Cameras	Trapper Harvest	-	-	Distribution Filtering	1.69	-6.16	-	-
Coyote	Cameras	Trapper Harvest	-	-	No Filtering	1.69	-6.16	-	-
Coyote	Cameras	Trapper Harvest	-	-	Observation Filtering	0.97	-4.73	-	-
Raccoon	Cameras	Hunter Harvest	-	-	Distribution Filtering	0	-0.01	-	-
Raccoon	Cameras	Hunter Harvest	-	-	No Filtering	28.91	-0.44	-	-

Table B2 (continued).

Raccoon	Cameras	Hunter Harvest	-	-	Observation Filtering	16.57	0.72	-	-
Raccoon	Cameras	Hunter Harvest	Hunter Observations	-	Distribution Filtering	21.94	0.83	-0.03	-
Raccoon	Cameras	Hunter Harvest	Hunter Observations	-	No Filtering	31.04	-0.42	-0.07	-
Raccoon	Cameras	Hunter Harvest	Hunter Observations	-	Observation Filtering	17.33	0.75	0.02	-
Raccoon	Cameras	Hunter Harvest	Hunter Observations	Trapper Harvest	Distribution Filtering	40.89	0.81	-0.08	7.47
Raccoon	Cameras	Hunter Harvest	Hunter Observations	Trapper Harvest	No Filtering	39.49	-0.4	-0.1	3.65
Raccoon	Cameras	Hunter Harvest	Hunter Observations	Trapper Harvest	Observation Filtering	56.67	0.58	-0.02	10.44
Raccoon	Cameras	Hunter Harvest	Trapper Harvest	-	Distribution Filtering	12	-0.01	4.82	-
Raccoon	Cameras	Hunter Harvest	Trapper Harvest	-	No Filtering	35.34	-0.44	3.11	-
Raccoon	Cameras	Hunter Harvest	Trapper Harvest	-	Observation Filtering	27.52	0.62	5.34	-
Raccoon	Cameras	Hunter Observations	-	-	Distribution Filtering	0.93	0.07	-	-
Raccoon	Cameras	Hunter Observations	-	-	No Filtering	6.3	-0.12	-	-
Raccoon	Cameras	Hunter Observations	-	-	Observation Filtering	0.75	0.05	-	-
Raccoon	Cameras	Hunter Observations	Trapper Harvest	-	Distribution Filtering	14.36	0.07	5.34	-
Raccoon	Cameras	Hunter Observations	Trapper Harvest	-	No Filtering	15.86	-0.15	3.91	-
Raccoon	Cameras	Hunter Observations	Trapper Harvest	-	Observation Filtering	21.79	0.05	6.16	-

Table B2 (continued).

Raccoon	Cameras	Trapper Harvest	-	-	Distribution Filtering	4.99	2.79	-	-
Raccoon	Cameras	Trapper Harvest	-	-	No Filtering	6.24	3.09	-	-
Raccoon	Cameras	Trapper Harvest	-	-	Observation Filtering	6.22	3.13	-	-

Table B3: Effect of data checking Step One, outlier removal and minimum effort filtering, on modeling outcomes.

Species	Model	Mean Occupancy Probability	Occupancy SD	Contagion Mean	Contagion SD	PRD Mean	PRD SD	Forest-Ag Mean	Forest-Ag SD	Develop Mean	Develop SD	LOO
Bobcat	Top Model	0.43	0.03	-0.21	0.16	-0.62	0.24	0.27	0.28	-5.33	1.34	7533
Bobcat	No Outlier Removal /Effort Filtering	0.33	0.02	0.02	0.20	-0.27	0.23	2.76	0.30	-1.31	1.32	7583
Coyote	Top Model	0.82	0.01	2.90	0.28	-0.87	0.24	27.13	2.21	29.80	2.02	3071
Coyote	No Outlier Removal /Effort Filtering	0.70	0.03	0.49	0.14	-0.28	0.14	1.31	0.48	3.36	0.62	7526
Raccoon	Top Model	0.80	0.02	0.92	0.18	-0.62	0.18	9.55	1.66	25.99	2.68	4363
Raccoon	No Outlier Removal /Effort Filtering	0.79	0.02	0.62	0.15	-0.30	0.15	4.76	0.92	7.16	1.68	5772

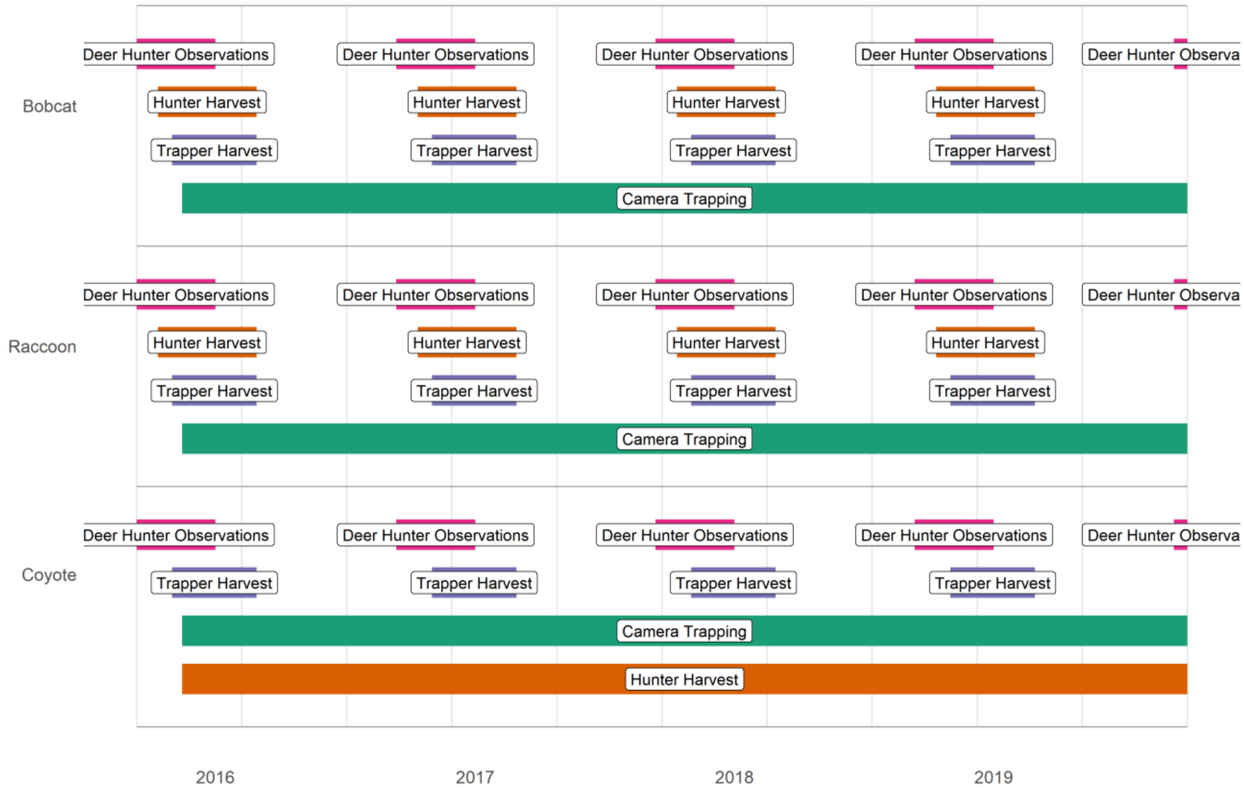


Figure B1: Visualization of the species-specific temporal coverage of data sources available during 2015 - 2019 in North Carolina, USA. Data were provided by the North Carolina Wildlife Resources Commission, the North Carolina Museum of Natural Sciences, and North Carolina State University.

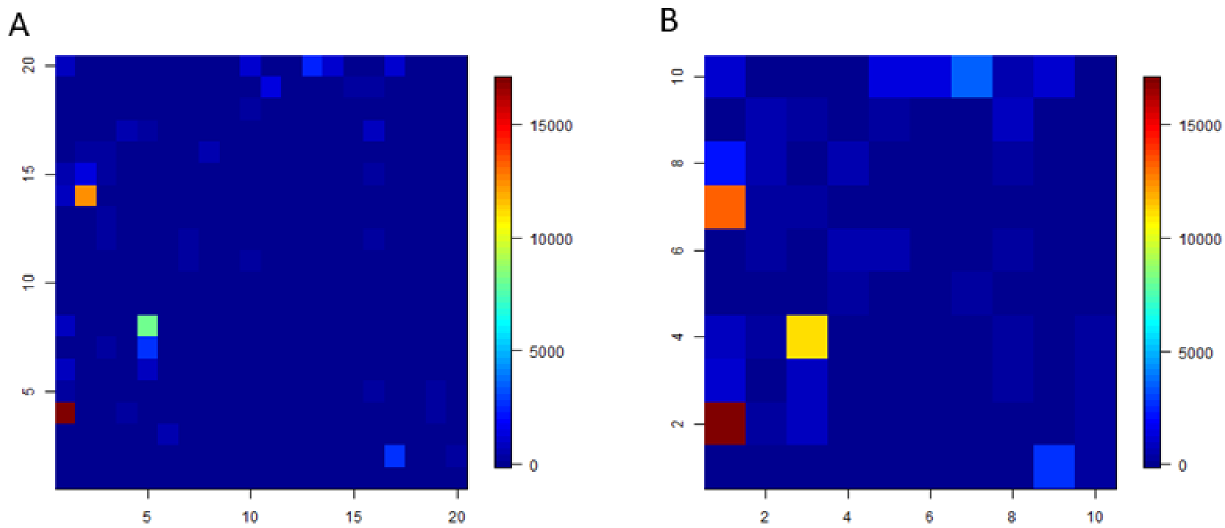


Figure B2: Illustration of (A) original unstructured data used in the simulation study to (B) an aggregation of the data to a coarse cell resolution which mimics empirical county-level records in North Carolina, USA.

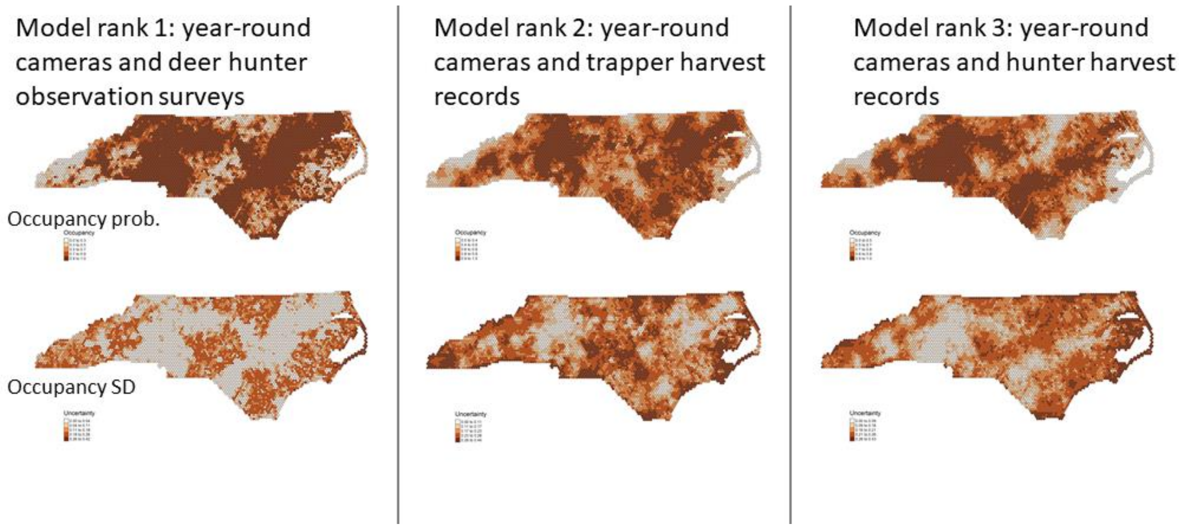


Figure B3: Top three integrated species distribution models for raccoon (Procyon lotor) in North Carolina, USA during 2015 - 2019. Top row of figures is the occupancy probability and bottom row represented occupancy standard deviation (SD).

Appendix C: Chapter 3 Supplementary Information

Table C1: Global regression coefficient estimates from a Downweighted Poisson Regression model fit using American Black Bear (Ursus americanus) observations during 2015-2019 from the iNaturalist database (Global DWPR) compared to (1) a single-species, single-season occupancy model fit using camera observations from the 2019 SNAPSHOT USA monitoring initiative, and (2) an integrated species distribution model using SNAPSHOT USA as the primary observations with iNaturalist included as a covariate effect.

Covariate	Global DWPR (iNaturalist)		Occupancy Model (SNAPSHOT USA)	
	Mean	SE	Mean	SE
Intercept	-9.35	0.02	-2.88	0.38
Coniferous Forests	0.29	0.01	0.29	0.14
Deciduous Forests	0.24	0.01	0.06	0.10
Mixed Forests	0.14	0.01	-0.04	0.09
Cropland	-0.61	0.04	-2.87	0.76
Human Population Density	-0.22	0.01	-0.38	0.11
Protected Area Status	0.80	0.03	1.00	0.23

Table C1 (continued)

Mean Annual Temperature	0.02	0.02	-0.95	0.29
Mean Annual Precipitation	0.28	0.01	0.39	0.21

Table C2: Model comparisons testing which non-stationarity hypothesis best explains variation in each ecological covariate's spatial pattern. Numeric values are absolute AIC values.

Covariate	Lineage	Subspecies	Ecoregion	Competition
Coniferous Forests	1021.9	**748.45**	925.44	976.63
Deciduous Forests	136	-8.66	**-66.57**	124.28
Mixed Forests	456.11	**260.92**	414.03	454.03
Cropland	3207.72	**2783.42**	3074.69	3141.15
Human Population Density	5031.33	**4879.316**	4997.2	5002.85
Protected Area Status	2318.67	**2089.12**	2462.35	2535.3
Mean Annual Temperature	5910.34	5784.65	**5707.28**	5806.37
Mean Annual Precipitation	3965.81	3948.56	**3907.56**	3965.74

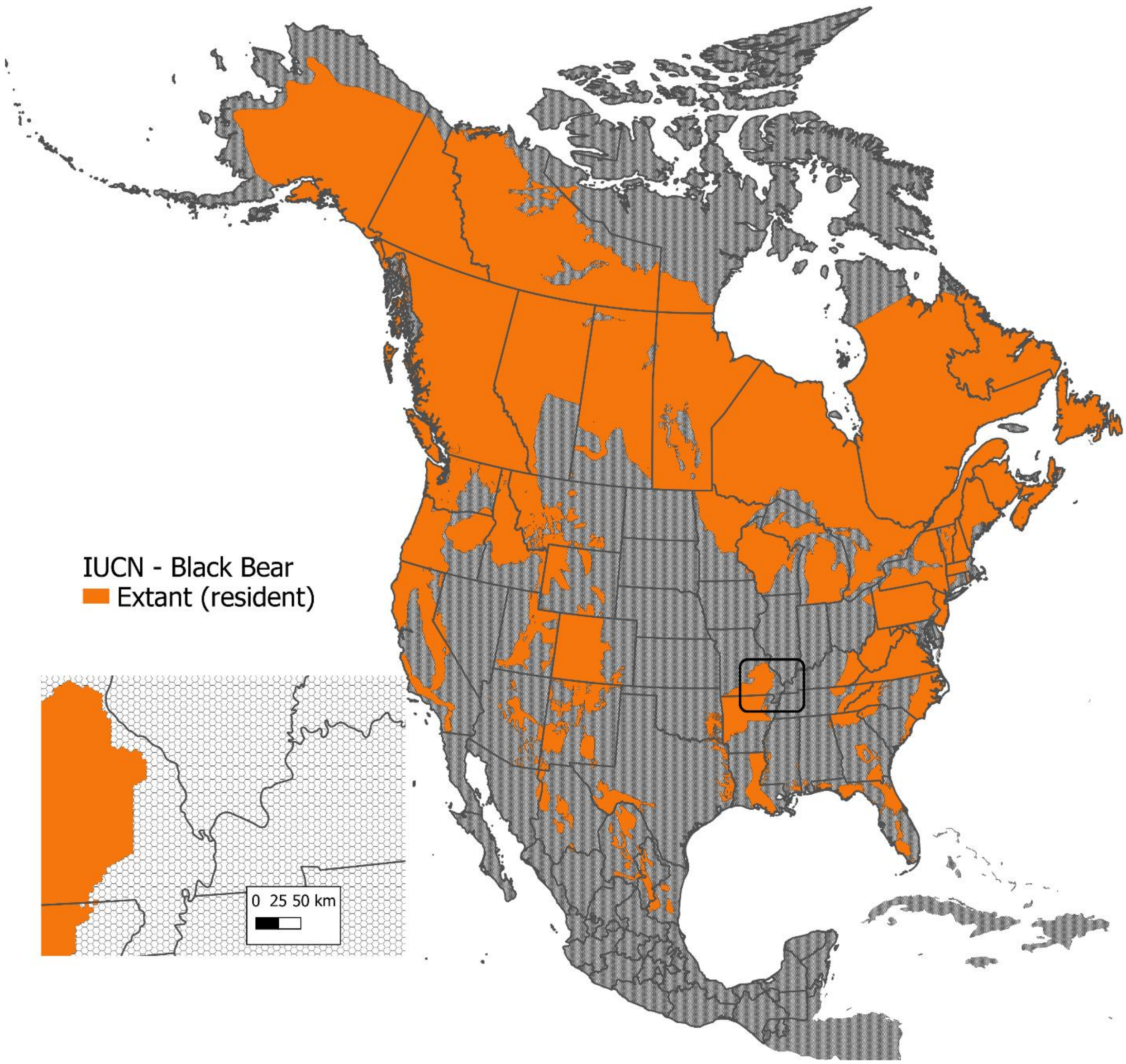


Figure C1: Current range map of American Black Bear (Ursus americanus) across North America (citation) with overlapping grid cells (n = 525,913,~50 sq km) used to summarize all available animal and environmental information.

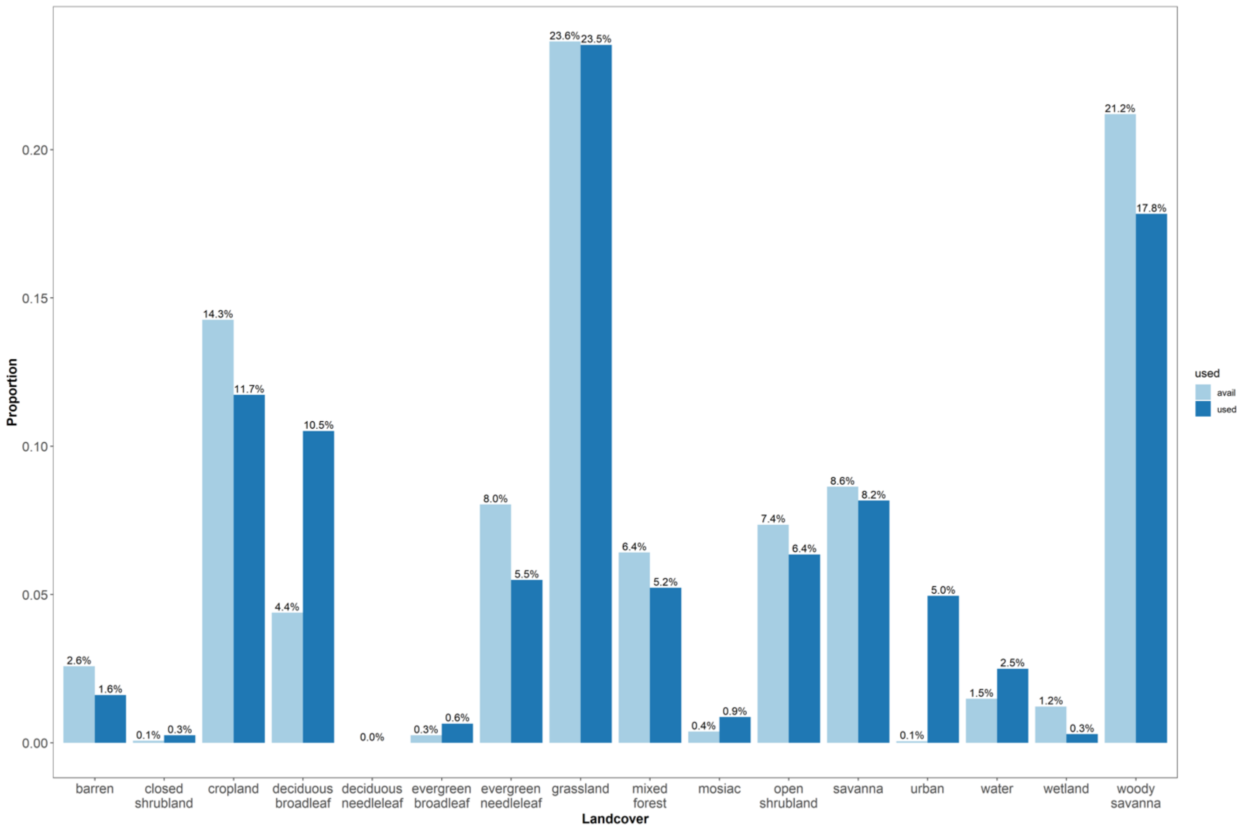


Figure C2: Use (grid cells where iNaturalist Mammalia observations occurred) vs Availability (grid cells without iNaturalist observations) across 16 land cover classes described by the Moderate Resolution Imaging Spectroradiometer (MODIS) Land Cover Type (MCD12Q1) product from 2019. Lighter shades of blue indicate availability while darker indicate conditions where iNaturalist observations occurred.