Resolving misaligned spatial data with integrated species distribution models

Krishna Pacifici[1,*], Brian J. Reich[2], David A. W. Miller[3], and Brent Pease[1]

[1] *Department of Forestry and Environmental Resources and Program in Fisheries, Wildlife, and Conservation Biology, North Carolina State University*

[2] *Department of Statistics, North Carolina State University*

[3] *Department of Ecosystem Science and Management, Pennsylvania State University*

[*] *corresponding author: jkpacifi@ncsu.edu*

Abstract

Advances in species distribution modeling continue to be driven by a need to predict species responses to environmental change coupled with increasing data availability. Recent work has focused on development of methods that integrate multiple streams of data to model species distributions. Combining sources of information increases spatial coverage and can improve accuracy in estimates of species distributions. However, when fusing multiple streams of data, the temporal and spatial resolutions of data sources may be mismatched. This occurs when data sources have fluctuating geographic coverage, varying spatial scales and resolutions, and differing sources of bias and sparsity. It is well documented in the spatial statistics literature that ignoring the misalignment of different data sources will result in bias in both the point estimates and uncertainty. This will ultimately lead to inaccurate predictions of species distributions. Here, we examine the issue of misaligned data as it relates specifically to integrated species distribution models. We then provide a general solution that builds off work in the statistical literature for the change of support problem. Specifically, we leverage spatial correlation and repeat observations at multiple scales to make statistically valid predictions at the ecologically relevant scale of inference. An added feature of the approach is that addressing differences in spatial resolution between data sets can allow for the evaluation and calibration of lesser quality sources in many instances. Using both simulations and data examples, we highlight the utility of this modeling approach and the consequences of not reconciling misaligned spatial data. We conclude with a brief discussion of the upcoming challenges and obstacles for species distribution modeling via data fusion.

Key words: black-throated blue warbler, change of support, occupancy modeling, integrated species distribution modeling, spatial modeling.

Introduction

Determining how species respond to changing environmental conditions is fundamental to sound management and species conservation (Yoccoz et al., 2001). Accomplishing this requires leveraging empirical evidence to inform and ultimately validate decision making. This need for data-driven decision making has motivated significant advances in the ability to collect and store spatially and temporally referenced data. At the same time there has been an influx in the development and application of methods that integrate multiple streams of data. These new data integration approaches seek to exhaust all available data sources to model species distributions while explicitly accounting for differences among data types (Dorazio, 2014; Fithian et al., 2015; Giraud et al., 2016; Pacifici et al., 2017; Coron et al., 2018). The advantages of combining multiple data sources in integrated species distribution models (ISDMs) include increased spatial coverage, bias reduction and overall improvement in estimator accuracy (Dorazio, 2014; Fithian et al., 2015; Giraud et al., 2016; Pacifici et al., 2017). Several authors have put forth different approaches for integrating different data sources, typically when one source is collected through standardized surveys and the other source is not (Miller et al., 2018+; Fletcher et al., 2018+). As a result we now have a range of methods that leverage information across different data types (Dorazio, 2014; Pacifici et al., 2017; Zipkin et al., 2017), among multiple species (Giraud et al., 2016; Thorson et al., 2016, 2017), and among neighboring locations by incorporating spatial correlation (Thorson et al., 2017). As data becomes more available and easier to access the propensity to combine data will only increase as will the demand to rigorously apply it to inform decision making.

In light of the increased interest in ISDMs, it is essential to explore the implications that come with combining different data sources. As with all species distribution modeling the goal is to correlate observations of individual species with environmental layers that are driving the observed patterns of occurrence. In some cases, the focus will be on large geographic areas or on species that are difficult to sample. Alternative data sources can fill in gaps that might occur in data collection and improve inference (Pacifici et al., 2017; Fletcher et al., 2018+; Miller et al., 2018+). ISDMs can increase precision and reduce bias in certain settings (Pacifici et al., 2017) and are flexible enough to incorporate a wide range of auxiliary data sources (Fletcher et al., 2018+; Miller et al., 2018+). Despite this, two major problems need to be addressed when fusing multiple streams of data. The first problem is to ensure that the ISDM rigorously combines each data source so that relevant and valid statistical inference is possible. The second is to properly reconcile spatial and temporal observations when they are collected at multiple differing spatial and temporal resolutions. The first problem has already received significant attention (Fletcher et al., 2018+; Miller et al., 2018+). The result is a range of flexible approaches that have been developed to rigorously integrate multiple data sources (Pacifici et al., 2017; Fletcher et al., 2018+). The second problem, however, has not been formally addressed for ISDMs in the ecological literature. This stands in contrast to significant coverage given to the topic in the spatial statistics literature where it is often referenced as the general *change of support* (COS) problem (Mugglin et al., 2000; Gelfand et al., 2001; Gotway and Young, 2002; Wikle and Berliner, 2005; Gotway and Young, 2007; Young and Gotway, 2007; Berrocal et al., 2010a,b; Ren and Banerjee, 2013; Reich et al., 2014; Parker et al., 2015; Kim and Berliner, 2016).

Before exploring the challenges of combining data sources and COS we first need to understand COS as it relates to a single data source. Here, we briefly describe three general COS problems. We encourage the reader to explore the topic more thoroughly (Gelfand et al., 2001; Gotway and Young, 2002) and *Journal of the Royal Statistical Society, Series A, Volume 164 Issue 1* dedicated to the topic. Generally, COS arises from three causes: 1) spatial or temporal misalignment, 2) modifiable areal unit problem (MAUP), and 3) the ecological fallacy problem. It is important to recognize that the effect of COS can exist in relation to either the response variable (e.g., counts, occurrences), the covariates driving the response (e.g., landcover, elevation), or both.

First, data may be ”misaligned” either spatially or temporally (Mugglin et al., 2000; Cressie and Wikle, 2015) meaning that data may come from different classifications or partitions of parcels of land or from different years or seasons. Take for example the case where a predictor variable (e.g., elevation) is measured at one spatial scale (e.g., county) and another variable (e.g., human population density) is measured at a different spatial scale (e.g., zip code). Our interest may be in using both covariates to explain variation, in say, abundance. However, the misalignment of the covariate information needs to be reconciled to make proper inference (Mugglin et al., 2000).

The same will hold for temporal mismatch wherein one covariate may be measured at a different temporal resolution (e.g., annually) than a second covariate (e.g., daily) and the differences must be recognized (Cressie and Wikle, 2015). The problem of misalignment also occurs when the response variable (e.g., counts, presence/absence data) is mismatched with the covariate information either spatially (e.g., counts occur at different spatial scale than covariate) or temporally (e.g., covariate information, say landcover, comes from different year than counts were collected).

The second problem classified under COS, is referred to as the modifiable areal unit problem (MAUP) in the geography and statistics literature (Gotway and Young, 2002). MAUP is essentially two separate problems: spatial aggregation and the grouping effect. Spatial aggregation is the process of grouping data into increasingly larger geographic units. This might occur when covariate information is aggregated or grouped to a larger scale to match another covariate or response variable (Latimer et al., 2006) or response variables (e.g., counts) are summarized at increasingly larger geographic scales (e.g, collected at point, summarized to county level). Spatial aggregation will change inferences for estimated parameters. The second problem of MAUP, the grouping effect, occurs when there are differences in the size, shape or formation of the geographic units (Gotway and Young, 2002). Grouping effects have been studied extensively in ecology for some time (Turner, 1989; Levin, 1992).

A third challenge, 'ecological fallacy' is often listed separately from MAUP, but also can be considered a special case of MAUP. Ecological fallacy deals with the case where the underlying individual response to a covariate differs from the response estimated from grouping the individuals (Gotway and Young, 2002; Bradley et al., 2016, 2017). The result is that conclusions based on an analysis using fine resolution data differ from analyses that are conducted using an aggregate or summary of the fine resolution data (Gotway and Young, 2002). 'Downscaling' is often used to address this problem in the environmental and remote sensing fields (Bradley et al., 2017). Often the most difficult piece is identifying which variables are responsible for significantly altering the results when data is scaled up or aggregated from individual level to group level (Gotway and Young, 2002).

In all three cases, notable bias can occur if it is not properly handled and choosing different scales to conduct the analysis results in different magnitudes of error (Bradley et al., 2017). Bias can occur not only in estimating the mean and variance of parameters of interest, but extends to any statistic that is estimated at multiple scales (Waller and Gotway, 2004; Bradley et al., 2017). The consequences of ignoring COS are hard to predict, and they can result in severe biases.

Although the statistical literature is rich with examples of COS (Gotway and Young, 2002) it generally remains unaddressed in species distribution modeling. Several authors note that COS occurs when species presence/absence data which are referenced to point locations and environmental data used to predict occurrence are typically referenced to grid cells (Latimer et al., 2006; Finley et al., 2014). Another example is when location errors arising from georeferenced covariate information that is summarized or aggregated to a grid cell (Hefley et al., 2017). Latimer et al. (2006) describe a solution as either working at the scale of the responses by assigning the environmental data to that level or alternatively working at the grid cell level by scaling up the response data to match the environmental data. However, this leads to a loss of information from rescaling to match either the level of the response or the level of the environmental data. Ideally we want a method to formally account for and circumvent the loss of information due to aggregating data and to recognize the variation within and between the aggregated units. The frequency with which COS will occur and associated issues becomes greater when combining multiple sources of data as now both the covariate information and auxiliary response data can be from mismatched scales.

The nuances of each type of COS warrants careful consideration of the appropriate solution. A wide range of methods exist (Gotway and Young, 2002) with the general goal being to make spatial predictions or estimate variables (covariates or responses) on regions over which they were not measured (Mugglin et al., 2000; Gotway and Young, 2002). As Cressie and Wikle (2015) recommend, the only logical solution is to build models at different scales and evaluate the differences in inference when doing so, ideally first building the model at the finest scale and then aggregating or scaling up to fit additional models. We will apply this general philosophy to COS with multiple data sources as well.

Here we lay out a framework for accommodating COS when combining multiple sources of data in ISDMs. First we describe the theoretical underpinnings of COS in the context of ISDMs, then develop COS extensions to a suite of data fusion models which vary in the level of shared information between data sources (Pacifici et al., 2017). We explore the properties of these models via simulation and apply these methods to our motivating data set on Black-throated Blue Warblers (*Setophaga caerulescens*; BTBW) in Pennsylvania, USA (Figure 1). Our overall objectives are two-fold: 1) introduce the concept of COS and demonstrate its relevancy to ISDMs, and 2) identify specific situations when COS most matters and provide recommendations for how it should be handled.

Change of Support for Integrated Species Distribution Models

We will use a case study of BTBW in Pennsylvania, USA to demonstrate the challenges of accounting for COS in ISDMs (Figure 1). Here two different data sources provide useful, yet different information about the distribution of BTBW. The first data source is collected at finer resolution standardized surveys across the state (Breeding Bird Atlas point counts) while the second data source (eBird) has been summarized at a coarser resolution to account for data that is not collected at a single point location. Combining these data create a conflict in spatial resolution and necessitates a method that addresses the misalignment. This requires being able to reconcile the misalignment between the two data sources and the differing spatial scales to make inference about the underlying distribution of BTBW.

Modeling Framework

We envision two general approaches to handle misalignment to accommodate COS. The first naive method is a two-stage approach (we formally define this approach below as the 'Covariate' model). The first step consists of imputing the second data source in the spatial locations where the response of the first data source is observed. The prediction could be done in any number of ways depending on the characteristics of the second data source (e.g., presence only, presence-absence, counts) and could be accomplished using any number of appropriate species distribution modeling techniques (Guillera-Arroita et al., 2015). The second step uses those predicted values as known constants and linear predictors in an ISDM (Dorazio, 2014; Fithian et al., 2015; Fletcher et al., 2016; Pacifici et al., 2017; Fletcher et al., 2018+; Miller et al., 2018+). However, this approach does not account for uncertainty in the predictions from the second data source during the first step and can result in potentially biased inference. The second general approach, and the one we will focus on here, is a joint-modeling strategy. In this case, both sources of data are modeled simultaneously. As a result, uncertainty is properly accounted for and propagated through to the predictions of the joint response. Below we describe the framework for joint-modeling of ISDMs to account for spatial misalignment.

Species distributions can generally be thought of as a continuous point process that describes the distribution of individuals across a species' range. The local intensity of the process (i.e., the local probability an individual occurs at any point in space) determines the density of animals across space. Building a statistical model for the distribution requires carefully aggregating the intensity function for a point process to the scale of the data. As with any probability density function of a continuous random variable, the probability of an observation at any single spatial location is zero. As a result, non-zero probabilities arise only when considering the number of observations in a spatial region. Therefore, some minimum form of aggregation is required. For example, if a camera trap is placed at location $\mathbf{s}_0$ and animals that pass within distance r from the camera are recorded, then the region of the survey, $\mathcal{B}$, is the circle with center $\mathbf{s}_0$ and radius r and the expected number of observations in $\mathcal{B}$ is $\tilde{\lambda}(\mathcal{B})$, which increases with r. Given that all observations are made with reference to an area, it is generally difficult to estimate the function $\lambda(\mathbf{s})$ for all $\mathbf{s}$ without simplifying assumptions about its smoothness.

Following much of the literature on species distribution models (SDMs), we specify our model for individual spatial locations (i.e., latitude/longitude) $\mathbf{s} \in \mathcal{R}^2$ using an inhomogeneous Poisson process (IPP) (Warton et al., 2010; Dorazio, 2014; Fithian et al., 2015). An IPP is simply a point process where the intensity (i.e., local density of individuals) varies across space. Let $\lambda(\mathbf{s})$ be the intensity of an inhomogeneous Poisson process so that the number of individuals in an arbitrary region $\mathcal{B}$ follows a Poisson distribution with mean

$$\tilde{\lambda}(\mathcal{B}) = \int_{\mathcal{B}} \lambda(\mathbf{s}) d\mathbf{s}. \tag{1}$$

The log-intensity process can be regressed onto covariates via the model $\log[\lambda(\mathbf{s})] = \mathbf{X}(\mathbf{s})^T\boldsymbol{\beta} + \theta(\mathbf{s})$ where $\mathbf{X}(\mathbf{s})$ is a vector of spatial covariates, $\boldsymbol{\beta}$ are the corresponding effects and $\theta(\mathbf{s})$ is the residual spatial process. Several models for the spatial intensity function have been proposed and we discuss three in the AppendixS1.

Data Fusion Models with COS

As we note previously, the focus of our paper is to integrate multiple data sources collected at different spatial resolutions. Assume that data source k is available for m_k regions $\mathcal{G}_{k1}, ..., \mathcal{G}_{km_k}$. Consistent with our motivating data example with BTBW in PA, we address the case where there are two data sources and: (1) the first data source Y_{1j} is the number of the N_j sampling occasions in region $\mathcal{G}_{1j}$ for which the species was observed, so that $Y_{1j} \in \{0, 1, ..., N_j\}$; and (2) that the second data source Y_{2j} is the total number of individuals observed in grid cell $\mathcal{G}_{2l}$, so that $Y_{2l} \in \{0, 1, 2, ...\}$. The approaches below are easily generalized to other cases. In our analyses we will treat the first data source as the "gold standard". The data is collected using a systematic sampling design where effort and location are well defined and offer a benchmark for our data integration model. The second data source contains auxiliary data for which we have less confidence and this is reflected in how we formulate models in some cases (Pacifici et al., 2017). We describe the methods below in the context of the discretized model that assumes the true intensity is constant with each fine-resolution grid cell $\mathcal{B}_1, ..., \mathcal{B}_n$ described above. For our motivating example this model is amenable to implementation in standard software (e.g., `OpenBUGS`, see available code in DataS1). However, we emphasize that other approaches can also be used in the data-fusion models developed in this section.

Here we layout three approaches for data fusion that vary in the degree of influence and reliance on the auxiliary data source (Pacifici et al., 2017) and extend each to allow for COS.

Covariate model

The simplest approach is to use the second data source or a summary of the second data source as a covariate in the model for the first data source (i.e. ad-hoc two-stage approach described above). Note that technically there is no modification for COS, instead information from the response is scaled up or matched to the covariate scale and therefore the misalignment is reconciled. We continue to include this model because it is a simple and effective way to address spatial misalignment. In this case, information from the auxiliary data only informs the species distribution model to the extent it can predict data from the second. The model for the first data source is

$$Y_{1j}|Z_j, p \sim \text{Binomial}(N_j, pZ_j) \tag{2}$$

where Z_j is the binary indicator that cell $\mathcal{G}_{1j}$ is occupied and $p \in [0, 1]$ is the probability of detection given that the cell is occupied. If the number of individuals in region j follows a Poisson distribution with mean $\tilde{\lambda}(\mathcal{G}_{1j})$ defined as in (Eq. S2) then the probability that the number of individuals is zero, i.e., that $Z_j = 0$, is $\exp[-\tilde{\lambda}(\mathcal{G}_{1j})]$. Therefore the probability that $\mathcal{G}_{1j}$ is occupied given $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_n)$ is

$$\text{Prob}(Z_j = 1|\boldsymbol{\lambda}) = 1 - \exp[-\tilde{\lambda}(\mathcal{G}_{1j})]. \tag{3}$$

We use a spatial log-Gaussian model for the intensity function therefore each fine-resolution cell $\mathcal{B}_i$ is

$$\log(\lambda_i) = \mathbf{X}_i^T \boldsymbol{\beta} + \theta_i \tag{4}$$

where $\mathbf{X}_i$ is the vector of covariates; $\boldsymbol{\beta}$ is the corresponding vector of regression coefficients; and θ_i is a spatial random effect. By including a summary of the auxiliary data in the covariate vector we use information in both data sources. For example, if fine resolution cell i falls in coarse resolution cell $\mathcal{G}_{2l}$ then we will use $\log(Y_{2l})$ as an element of $\mathbf{X}_i$. As a result, the coarse-resolution covariate might prove useful for capturing large-scale spatial patterns, but cannot resolve fine-scale variation. Pacifici et al. (2017) show this is a useful model when belief in the second data source is low or uncertain as nothing is assumed about the auxiliary data source.

We chose to estimate the vector of spatial random effects $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)^T$ using a conditional autoregressive prior (CAR) (Banerjee et al., 2014), but note that any model for a continuous spatial process could be used with a few modifications, however, we chose to use the CAR model because of its easy implementation in available software (e.g., BUGS). The CAR model can be motivated using the full conditional specification of θ_i given the value of the process at all other cells,

$$\theta_i|\theta_k \text{ for all } k \neq i \sim \text{Normal}(\rho\bar{\theta}_i, \sigma^2/m_i),$$

where $\bar{\theta}_i$ is the mean of $\boldsymbol{\theta}$ at the m_i cells adjacent to cell i and the two spatial dependence parameters $\rho \in (0,1)$ and $\sigma^2 > 0$ determine the strength of spatial dependence and conditional variance, respectively. These full conditional distributions lead to the joint multivariate normal distribution

$$\boldsymbol{\theta} \sim \text{Normal}\left[\mathbf{0}, \sigma^2\left(\mathbf{M} - \rho\mathbf{A}\right)^{-1}\right], \tag{5}$$

where $\mathbf{M}$ is the diagonal matrix with i^{th} diagonal element equal to m_i and $\mathbf{A}$ is the adjacency matrix with (i,k) element equal one if cells i and k are adjacent and zero otherwise. We denote this as $\boldsymbol{\theta} \sim \text{CAR}(\rho, \sigma, \mathbf{A})$.

Shared model

An alternative to the simple naive approach is to treat both data sources as outcomes of the same underlying distribution and relate them directly to the shared underlying inhomogeneous point process. In this case there is a single underlying species distribution, but each data source is allowed to have its own model that describes the observation process (i.e., the probability of collecting a given observation conditional on the true number of individuals within a given location). All outcomes are taken to be conditionally independent given the intensities. The joint model is

$$\begin{aligned} Y_{1j}|Z_{1j},p &\sim \text{Binomial}(N_j, pZ_j) \quad \text{with} \quad \text{Prob}(Z_j = 1|\boldsymbol{\lambda}) = 1 - \exp[-\tilde{\lambda}(\mathcal{G}_{1j})] \qquad (6)\\ Y_{2l}|\boldsymbol{\lambda}, a, b &\sim \text{Poisson}\left[a + b\tilde{\lambda}(\mathcal{G}_{2l})\right], \end{aligned}$$

where $a > 0$ and $b > 0$ are additive and multiplicative bias terms, respectively and represent the degree of relatedness between the two data sources (e.g., when $b = 0$ the second data source is completely uninformative). The intensity surface $\boldsymbol{\lambda}$ is modeled as in (4) except without Y_{2l} as an element of $\mathbf{X}_i$. This is the typical joint-likelihood approach found in many applications of integrated population models and ISDMs (Dorazio, 2014; Fletcher et al., 2016; Zipkin et al., 2017; Fletcher et al., 2018+). It assumes that the secondary data source is of high quality and/or information is available to model the sources of bias and variability (e.g., false positives, variable sampling effort).

Correlation model

A second alternative to fully modeling the joint-likelihood is to specify separate, but correlated, underlying processes for the two data sources. The idea here is to estimate two separate species distributions, one with each data set, while allowing the two distributions to be correlated. If they are perfectly correlated, information is completely shared across the two distribution models. If the correlation is <1, then information is shared in proportion to the strength of correlation. This allows us to relax the reliance on the auxiliary data source necessary for joint-likelihood approaches while still permitting information to be shared between the sources of data (Pacifici et al., 2017). Let $\lambda_k(\mathbf{s})$ be the Poisson intensity function for data source $k \in \{1, 2\}$ and $\tilde{\lambda}_k(\mathcal{G}) = \int_{\mathcal{G}} \lambda_k(\mathbf{s}) d\mathbf{s}$ be the aggregated intensity for process k. As with the shared model, the data are conditionally independent given the Poisson intensities,

$$\begin{aligned} Y_{1j}|Z_{1j}, p &\sim \text{Binomial}(N_j, pZ_j) \quad \text{with} \quad \text{Prob}(Z_j = 1|\boldsymbol{\lambda}_1) = 1 - \exp[-\tilde{\lambda}_1(\mathcal{G}_{1j})] \\ Y_{2l}|\boldsymbol{\lambda}_2 &\sim \text{Poisson}\left[\tilde{\lambda}_2(\mathcal{G}_{2l})\right]. \end{aligned} \tag{7}$$

Both processes are defined on the same fine grid n grid cells, with

$$\log(\lambda_{ki}) = \mathbf{X}_i^T \boldsymbol{\beta}_k + \theta_{ki}, \tag{8}$$

where $\mathbf{X}_i$ is the vector of covariates; $\boldsymbol{\beta}_k$ is the corresponding vector of regression coefficients for data type k; and θ_{ki} is a spatial random effect. The spatial random effects $\boldsymbol{\theta}_i = (\theta_{1i}, \theta_{2i})^T$ are modeled using a multivariate CAR model (Banerjee et al., 2014), defined by its full conditional distributions

$$\boldsymbol{\theta}_i | \boldsymbol{\theta}_{i'} \text{ for all } i' \neq i \sim \text{Normal}(\rho \bar{\boldsymbol{\theta}}_i, \boldsymbol{\Sigma}/m_i),$$

where $\bar{\boldsymbol{\theta}}_i = (\bar{\theta}_{1i}, \bar{\theta}_{2i})^T$ is the mean of $\boldsymbol{\theta}$ at the m_i cells adjacent to cell i; $\rho \in (0,1)$ controls the strength of spatial dependence; and the 2×2 covariance matrix $\boldsymbol{\Sigma}$ controls the dependence between θ_{1i} and θ_{2i}. This model allows for the processes underlying two data sources to be correlated. Thus each data source informs predictions from the other, but in an indirect manner. As a result, there is a reduced burden for the auxiliary data being of equally reliable to the first data source.

Simulation Study: Aggregating Spatial Covariates with a single data source

Now that we have formally defined COS in ISDMs we want to explore one of the most common challenges that researchers first face when fitting SDMs, how to use spatial covariates that have been collected at different spatial scales. Below we describe a brief simulation study to evaluate the effect of aggregating spatial covariates with a single data source. In this simulation the true intensity surface is generated on a 20×20 fine grid of $n = 400$ grid cells $\mathcal{B}_1, ..., \mathcal{B}_n$. Data are generated on grid cells $\mathcal{G}_1, ..., \mathcal{G}_m$ where each cell contains regular grid of k^2 of the n fine-resolution cells, with S_j denoting the indices of the fine resolution cells in $\mathcal{G}_j$ so that $\mathcal{G}_j = \bigcup_{i \in \mathcal{S}_j} \mathcal{B}_i$ (e.g., Figure S1a for $k = 3$). We first simulate the spatial random effects $\boldsymbol{\theta} = (\theta_1, ..., \theta_n) \sim$ CAR$(0.99, 1, \mathbf{A})$ and covariate $\mathbf{X} = (X_1, ..., X_n)^T \sim$ CAR$(\rho, 1, \mathbf{A})$. The true intensity is then set to $\log(\lambda_i) = \beta_1 + X_i \beta_2 + \theta_i$ with $\beta_1 = 0$ and $\beta_2 = 1$. The data for $\mathcal{G}_j$ is then generated as $Y_j \sim$ Binomial$(5, pZ_j)$ where Prob$(Z_j = 1) = 1 - \exp(-\sum_{i \in \mathcal{S}_j} \lambda_i)$ with detection probability $p = 0.5$. Data are simulated with aggregation level either $k = 2$ or $k = 3$ spatial correlation of the covariate equal either $\rho = 0.50$ or $\rho = 0.99$. For all combinations of these settings we simulate 500 datasets.

For each simulated data set we fit two models. The first model (“naive”) ignores COS and fits a standard spatial occupancy model using m observations where the log intensity in cell $\mathcal{G}_j$ is

$\beta_1 + \tilde{X}_j\beta_2 + \gamma_j$ where $\tilde{X}_j$ is the average of X_i over $\mathcal{G}_j$ and the m spatial effects $\gamma_1, ..., \gamma_m$ follow a CAR prior defined via the adjacency matrix of $\mathcal{G}_1, ..., \mathcal{G}_m$. The second model ("COS") is the COS model used to generate the data wherein we account for COS by modeling the process at the same fine resolution that we generated the data (instead of using the average as in the naive model). Both models assume priors $\beta_1, \beta_2 \sim \text{Normal}(0, 10)$, $\sigma^2 \sim \text{InvGamma}(0.1, 0.1)$, $\rho \sim \text{Beta}(10, 1)$ and $p \sim \text{Uniform}(0, 1)$. Models are fit in `OpenBUGS` using 10,000 MCMC samples after a burn-in period of 2,500 iterations (see DataS1 for code). For each model and each dataset we compute the posterior distribution of the slope β_2, and present the bias and mean square error of the posterior mean and empirical coverage of 90% intervals averaged over the 500 datasets in Table 1.

The naive method that ignores COS is positively biased in all cases. The bias and MSE are the largest in the cases with a highly correlated covariate process ($\rho = 0.99$). Although the COS method does not completely eliminate the bias, it is greatly reduced especially in the cases with spatial correlation therefore highlighting the need to account for COS even with mismatched covariate and response data.

Simulation Study: ISDMs with and without COS

To fully evaluate the newly developed COS ISDMs we conduct a simulation study to determine the conditions in which fusing data sources with different spatial resolutions improves estimates, and compare the efficiency of various ways to account for the COS. We generated data from the shared model on a 20 x 20 rectangular grid. The data are originally generated on the same grid as the true process, i.e., $\mathcal{B}_i = \mathcal{G}_{1i} = \mathcal{G}_{2i}$. The observed data in cell i is a function of the latent spatial process λ_i. The true binary occupancy status is generated as $\text{Prob}(Z_i = 1|\lambda_i) = 1 - \exp(-\lambda_i)$. The fine-scale data are then sampled as

$$Y_{i1}|Z_i \sim \text{Binomial}(N, pZ_i) \quad \text{and} \quad Y_{i2}|\lambda_i, \sim \text{Poisson}(E\lambda_i), \tag{9}$$

where N = 5, the detection probability p is either 0.2 or 0.5 and $E = 10$ is the offset for the second data source. The latent intensities λ_i are simulated as $\log(\lambda_i) = S_i$ where $(S_i, ..., S_n)$ is generated from the CAR model (with rook neighbors) with mean zero, variance parameter $\sigma^2 = 1$, spatial dependence parameter ρ set to either 0.50 or 0.99. The first data source, Y_{1i}, is observed for all $n = 400$ grid cells; the second data source, Y_{2i}, is only observed as aggregated counts over $k \times k$ (k is either 2 or 4) rectangular grids, denoted $\bar{Y}_{2j}$ for coarse-resolution grid cell $j = 1, ..., n/k^2$. Figure S1 plots one realization with $p = 0.2$, $\rho = 0.99$ and $k = 4$.

For each combination of k, ρ, and p we generate 100 datasets and fit the following models:

- **Single**: The second data source is ignored
- **Covariate**: The covariate model with $\log(\bar{Y}_{2j} + 1)$ is used as a covariate
- **Shared**: The joint model for $\bar{Y}_{2j}$ and Y_{1i}
- **Correlation**: The correlation model for $\bar{Y}_{2j}$ and Y_{1i}
- **Shared - no COS**: $\bar{Y}_{2j}$ is assumed to represent one central fine scale grid cell and the data are analyzed using the shared method without COS (Figure S1d)
- **Correlation - no COS**: $\bar{Y}_{2j}$ is assumed to represent one central fine scale grid cell and the data are analyzed using the correlation method without COS (Figure S1d)

Each model is fit using `OpenBUGS` with three chains each with 20,000 iterations and the first 5,000 iterations discarded as burn-in (see DataS1 for code). We used uninformative priors for all parameters and evaluated convergence using the Gelman-Rubin statistic and examining trace plots.

For each method and each dataset we compute the posterior probability that the fine-resolution cells are occupied, denoted $\bar{Z}_i$, and the declaration that the cell is occupied, denoted $\hat{Z}_i = 1$ if $\bar{Z}_i \geq 0.5$ and $\hat{Z}_i = 0$ if $\bar{Z}_i < 0.5$. Methods are evaluated using the Brier Score which is a proper score function to evaluate predictive performance for binary outcomes (Gneiting and Raftery, 2007; Pacifici et al., 2016) (BS: lower is better) and classification accuracy (CA) averaged over cells,

$$BS = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z}_i)^2 \text{ and } CA = \frac{1}{n}\sum_{i=1}^{n} Z_i\hat{Z}_i + (1 - Z_i)(1 - \hat{Z}_i), \tag{10}$$

where Z_i is the true occupancy status generated from the CAR model. Table 2 reports the median Brier score and classification accuracy across the 100 datasets for each method and each simulation scenario.

Including the second data source only shows substantial improvement compared to the single data source model when the grid cells are small ($k = 2$) and detection is low ($p = 0.2$). With large grid cells the aggregated data is too coarse to provide useful spatial information, and with high detection the first data sources provide sufficient information to produce precise maps since we included data from all cells within the area for this data source. Strong spatial correlation improves classification accuracy for all methods, but the second data source provides roughly the same increase in precision regardless of the spatial correlation.

Focusing on the two cases with $k = 2$ and $p = 0.2$ where including the second data source is useful, the results are fairly robust to the COS method. The two simplest COS methods are the covariate model and the naive methods that include the aggregated data as a data point without accounting for COS. These two simple models perform comparably to the more sophisticated shared and correlation models. The average run-times for these methods (Table 2c) are approximately 50% less than the full correlation model. In summary, these two methods provide simple and effective means of accommodating COS in ISDMs.

Case Study: Black-throated Blue Warblers in Pennsylvania

We next apply the data fusion models with and without COS on a data set for BTBW in PA, USA. Our goal is to examine the real world consequences of ignoring COS and to make recommendations for modeling. We have two data sets collected from two different sources. We further subsample these data at different spatial scales (i.e., observations are assigned to cells of increasing sizes) to understand the utility of incorporating COS into ISDMs.

The first data set we use includes point count survey data collected as part of the second Pennsylvania Breeding Bird Atlas (BBA data) (Wilson et al., 2012). During a five year period from 2005 to 2009, 33,846 point count surveys were conducted across the state of Pennsylvania. An even distribution of points was achieved by randomly selecting 8 roadside locations within each standard 1/24 degree latitude by 1/16 degree longitude blocks used for the atlas (Grid 1; Table 3). Point counts occurred during morning hours in the peak breeding season (last week of May through the end of June). Observers recorded singing males of all species during a 6 minute 15 second survey. Observations were divided into 5 75-second intervals and whether the bird was located less than or greater than 150-m from the observer. In our analysis we used all observations of singing male BTBW. We excluded observations > 150 m from the observer.

Our second data set consists of eBird observations (Sullivan et al., 2009). We filtered eBird records to only include observations during the same 5-year period (2005-2009) and only included records during the breeding season (late May to July). Records that did not include measures of survey effort were excluded. A subset of the BBA data was entered into the eBird database. To avoid duplication these records were also removed for analysis. A total of 4937 checklists were included in our analyses. eBird data was summarized at three different resolutions not including the original scale of the BBA data (Grid 1; Table 3).

Preliminary analyses found that percent forest cover has a positive relationship with the occurrence of black-throated blue warbler. We therefore include this covariate in all of the models to understand the consequences of spatial misalignment on the ability to estimate the covariate effects. In addition we summarize the second data source (eBird) in two different ways, first we take the sum of the eBird counts for a particular grid size and average it across all of the BBA cells at grid 1 within the larger grid (denoted by 'Avg' following the model name). Second, we explore the effects of an ad hoc approach wherein we reconcile the misalignment by matching the grids for all of the data (referenced by 'Scaled' following the model name). That is we scale up the BBA data to match the eBird grid. This is to mimic the case where nothing is known about the location of the finer resolution data and instead scale it up to match the second data source.

To fully evaluate the effects of ignoring vs accommodating COS we fit the data fusion models described in the *Data Fusion Models with COS* section with and without COS to 20% of the BBA data and compare the results with a model fit to all of the BBA data. The full BBA data set (33,846 points across PA) has excellent geographic coverage and by subsetting this data set we were able to explore the contrast in performance among the approaches.

Case Study Results

Overall models ignoring COS perform poorly compared to models incorporating COS. Figure 2 shows the estimated occupancy probability across data fusion models and whether COS was incorporated. All of the models incorporating COS had smaller credible intervals and were centered around the value estimated by the full BBA data set. Models ignoring COS and using both data sources equally (Shared) resulted in most estimates that are much higher than the full data set although this is not the case when the covariate is aggregated up to match the eBird grid size (models with 'Scaled' after name). The covariate model using the averaged covariate across all of the finer resolution cells (models with 'Avg' after name) performs well comparatively to more complex models (shared and correlation).

Individual site level estimates of ψ show similar results. Figure S2 plots the estimates when both data sources are at grid level 1. Models that do not account for COS tend to over smooth the estimated occurrence probabilities compared to the full data set. This becomes more pronounced as the degree of spatial misalignment increases (Figure 3). Again the covariate model performs competitively with the more complex shared COS model and outperforms the models ignoring COS.

We can compare the performance of the two models using different approaches to summarizing the second data source in the covariate model. Figure 4 shows the performance at grid level 2 while Figure S4 depicts the performance at grid level 4. Here we can see how the second approach (scaling up the first data source to match the second) clearly averages over the spatial variation at a finer scale and over smooths the predictions.

Figure 5 shows the differences in estimated effects of percent forest cover when ignoring COS vs accommodating it for data fusion models. The full data set (denoted by 'Single) shows a positive relationship with % forest cover and occurrence probabilities. This relationship is not as clear with the data fusion models although this is probably do to the reduction in data (full data set vs 20% of the data being used for all of the data fusion models). Overall the models incorporating COS tend to perform less variably and have reduced uncertainty estimates. It is also important to note as the degree of misalignment increases the amount of uncertainty increases as well. Models using the second data source summarized at grids 3 and 4 have highly variable and uncertain estimates relative to models using the second data source summarized at grids 1 and 2. This pattern is especially pronounced for the models ignoring COS.

Discussion

We present the first comprehensive treatment of spatial misalignment for ISDMs in the ecological literature. Within the spatial statistics literature, it is well known that spatial alignment matters when making predictions (Gelfand et al., 2001; Gotway and Young, 2002). Thus it is not surprising that our results show that COS matters and when unaddressed leads to biased parameter estimates when combining data sources to build ISDMs. Data integration methods have shown both utility and future promise to improve our inferences about species distributions as well as population and community dynamics (Fletcher et al., 2018+; Miller et al., 2018+; Zipkin and Saunders, 2018). While much of the current effort has focused on the development of estimators for different data types, (Dorazio, 2014; Fletcher et al., 2016; Pacifici et al., 2017), a parallel effort is needed to deal with scale and alignment in building models.

Our results highlight cases where not accounting for COS may be especially prone to introduce bias and reduce accuracy in results. We found bias and misclassification errors to be greatest when spatial correlation was high and when detection was low. Error due to COS was also greater when the relationship between distribution and the environment is defined by small-scale processes. For example, greater bias would be expected in our estimated relationships for BTBW if abundance was more correlated to local forest cover within 100 m of a location rather than at the landscape scale measured when values are taken for whole grid cells. In general, summarizing covariate information to match the grid size of observations smooths over important spatial variation, and can result in a loss of power to detect relationships and fine-scale trends. The likely result is that the strength of ecological relationships are underestimated. This is not a result unique to data integration methods, but is the case any time we fit models at coarse scales and ignore the COS issue.

We explored three general approaches to data integration, which we refer to as a shared, correlation, and covariate models for integrating two data sources (Pacifici et al., 2017). The covariate modeling approach provides a simple and efficient method for dealing with COS when it occurs between two data sets. By using data collected at a coarser scale as a covariate, it is possible to estimate the relationship of fine level processes while sufficiently accounting for information loss due to spatial misalignment. The extent of the spatial misalignment will define the extent to which the two data sets are correlated. As demonstrated previously (Pacifici et al., 2017; Miller et al., 2018+), the covariate approach also provides a flexible method to deal with other observational errors such as misidentification and misspecification of locations.

What we refer to as a shared modeling approach or a joint-likelihood approach leads to the greatest preservation of information when COS is accounted for while combining data sources. Using a shared approach requires that both data sets be of high quality and that COS can accurately be modeled between the two data sets. If this is the case, then information from both data sets are placed on equal footing and are used to model a shared (or joint) underlying process. In contrast, when it is difficult to specify the COS, the covariate approach performed relatively well, especially when the primary data set can be specified at a fine scale.

Our results point to some recommendations for SDMs in general, not just when data integration is used. Misalignment between covariate resolution and the size of the grid cell for which responses are modeled is not unique to integrated methods (Latimer et al., 2006). One insight from our specific results is that fine-scale relationships between covariate and species distribution are more affected by ignoring misalignment than coarse-scale relationships. This suggests that covariates such as average climate, which tend to follow smoother gradients, especially in non-mountainous regions, should be relatively robust to spatial misalignment. Alternatively, estimating fine-scale habitat relationships, such as the effect of forest cover in a fragmented landscape will be more sensitive to misalignment. In addition, many of the data sets we use to predict species distributions such as museum records, citizen science data, or even large-scale designed surveys include large uncertainty about spatial location of where records are located (Dickinson et al., 2010). Therefore there is a need to better understand how scale influences inferences made from all SDMs (Steenweg et al., 2018).

COS Model Steps

We are unable to provide general recommendations that are ubiquitous to fitting ISDMs. However, we provide five steps that we believe should be followed when addressing COS in ISDMs.

1. Define the stochastic model for ecological process at the finest scale or resolution.

2. Define support for observed data and determine the desired scale for predictions, i.e. scale that conservation and management decisions will be made.

3. Identify best way to link data sources based on underlying ecological process. Here a second data source may provide a diversity of information including sources of error or effort.

4. Develop joint model for data sources and the underlying ecological process and conduct inference.

5. Conduct model evaluation and check for sensitivity (e.g., significant change in results when adding new data sources) specifically when rescaling the data.

Temporal Mismatch

Here we have purposely excluded a full evaluation of temporal mismatch because we believe it deserves its own treatment in a separate paper. However, we can provide a few insights into handling temporal mismatch based on our experiences with ISDMs. The first question an analyst must address is whether or not there is interest in a static or dynamic model of species distribution. This question dictates the types of data collected and the temporal resolution necessary to assume that distributional patterns are changing through time. If the analyst is interested in modeling distributional changes via dynamic models then the temporal resolution of the data must represent the appropriate time scale to allow changes in the distribution at an ecologically relevant scale. When combining multiple sources of data this can present challenges when opportunistic data potentially arises from historic records (e.g., museum records) creating a gap in time. For example it is common to use presence-only data that may have originated decades earlier than survey data. In this case the appropriate inference depends on the interpretation of ’distribution’ in that a coarse time scale suggests a coarser definition of distribution and is akin to results from redefining the response of interest (Guillera-Arroita et al., 2015). We believe that this definition can be relaxed when interest involves a static distribution of species occurrence, but this is still an important and active area of research to fully understand the implications of temporal mismatch when combining multiple sources of data.

Furthermore, to fully understand the implications of combining different data sources, it is necessary to classify the use of auxiliary data by how it is used to inform SDMs. Similar to integrated population models (IPMs) wherein the goal is to include supplemental data sources which inform specific vital rates that drive populations (Zipkin and Saunders, 2018), we can identify the components of SDMs and how integrating new data improves our understanding of distribution and distributional changes in populations. Specifically, we are interested in how additional data sources improve our understanding of the drivers of distributions and we do this by classifying new sources of data into two categories, spatial and/or temporal, wherein new information can be added. The spatial category can be thought of as including additional observations (presences and/or absences) that modify the geographic footprint of a species, provide information about sampling effort or variation in sampling effort, sources of bias or error (e.g., false positives or false negatives) or that help reduce these sources of error, and uncover or identify relationships with environmental covariates or other species (especially at different spatial scales). Adding temporal information includes observations (presences and/or absences) that modify the geographic range over a temporal scale (e.g., annual or seasonal variation) of interest, or improve our understanding of error and/or sampling effort (similar to spatial) except which occurs over a temporal scale instead of spatial scale. The classification of how additional data will inform SDMs is a critical step in fully understanding whether it is worth using auxiliary information and how it will help.

Future Directions

As we move forward and the number of opportunities to combine data sources increases we believe future directions for research include the need to more fully explore situations where spatial misalignment has the greatest influence on SDMs. In addition, new applications such as dynamic distribution models are also likely to be affected by COS specifically because the ability to estimate changes in distribution are dependent on differentiating when local changes did and did not occur, often at a finer-scale than the resolution of many data sets (Kery et al., 2013; Zurell et al., 2016). Finally, spatial alignment is not a problem unique to data integration for SDMs. Other integrated models, such as IPMs will benefit from better understanding the effects of spatial misalignment and accounting for COS (Schaub and Abadi, 2011; Zipkin et al., 2017).

Acknowledgments

We would like to thanks Steve Beissinger, Brian Inouye, and Elise Zipkin and two anonymous reviewers for helpful comments on earlier drafts of the manuscript.

References

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2014) *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press.

Berrocal, V. J., Gelfand, A. E. and Holland, D. M. (2010a) A bivariate space-time downscaler under space and time misalignment. *The Annals of Applied Statistics*, **4**, 1942.

— (2010b) A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural, Biological, and Environmental Statistics*, **15**, 176–197.

Bradley, J. R., Wikle, C. K. and Holan, S. H. (2016) Bayesian spatial change of support for count-valued survey data with application to the american community survey. *Journal of the American Statistical Association*, **111**, 472–487.

— (2017) Regionalization of multiscale spatial processes by using a criterion for spatial aggregation error. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **79**, 815–832.

Coron, C., Calenge, C., Giraud, C. and Julliard, R. (2018) Bayesian estimation of species relative abundances and habitat preferences using opportunistic data. *Environmental and Ecological Statistics*, **25**, 71–93.

Cressie, N. and Wikle, C. K. (2015) *Statistics for spatio-temporal data*. John Wiley & Sons.

Dickinson, J. L., Zuckerberg, B. and Bonter, D. N. (2010) Citizen science as an ecological research tool: challenges and benefits. *Annual Review of Ecology, Evolution, and Systematics*, **41**, 149–172.

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

Finley, A. O., Banerjee, S. and Cook, B. D. (2014) Bayesian hierarchical models for spatially misaligned data in R. *Methods in Ecology and Evolution*, **5**, 514–523.

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

Fletcher, R., Hefley, T., Robertson, E., Zuckerberg, B., McCleery, R. and Dorazio, R. (2018+) A practical guide for combining data to predict species distributions. *Ecology*. In press.

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

Gelfand, A. E., Zhu, L. and Carlin, B. P. (2001) On the change of support problem for spatio-temporal data. *Biostatistics*, **2**, 31–45.

Giraud, C., Calenge, C., Coron, C. and Julliard, R. (2016) Capitalizing on opportunistic data for monitoring relative abundances of species. *Biometrics*, **72**, 649–658.

Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359–378.

Gotway, C. A. and Young, L. J. (2002) Combining incompatible spatial data. *Journal of the American Statistical Association*, **97**, 632–648.

— (2007) A geostatistical approach to linking geographically aggregated data from different sources. *Journal of Computational and Graphical Statistics*, **16**, 115–135.

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

Hefley, T. J., Brost, B. M. and Hooten, M. B. (2017) Bias correction of bounded location errors in presence-only data. *Methods in Ecology and Evolution*.

Kery, M., Guillera-Arroita, G. and Lahoz-Monfort, J. J. (2013) Analysing and mapping species range dynamics using occupancy models. *Journal of Biogeography*, **40**, 1463–1474.

Kim, Y. and Berliner, L. M. (2016) Change of spatiotemporal scale in dynamic models. *Computational Statistics & Data Analysis*, **101**, 80–92.

Latimer, A. M., Wu, S., Gelfand, A. E. and Silander, J. A. (2006) Building statistical models to analyze species distributions. *Ecological Applications*, **16**, 33–50.

Levin, S. A. (1992) The problem of pattern and scale in ecology: the robert h. macarthur award lecture. *Ecology*, **73**, 1943–1967.

Miller, D. A., Pacifici, K., Sanderlin, J. and Reich, B. J. (2018+) The recent past and promising future for data integration methods to estimate species distributions. *Methods in Ecology and Evolution*. In press.

Mugglin, A. S., Carlin, B. P. and Gelfand, A. E. (2000) Fully model-based approaches for spatially misaligned data. *Journal of the American Statistical Association*, **95**, 877–887.

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

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

Parker, R. J., Reich, B. J. and Sain, S. R. (2015) A multiresolution approach to estimating the value added by regional climate models. *Journal of Climate*, **28**, 8873–8887.

Reich, B. J., Chang, H. H. and Foley, K. M. (2014) A spectral method for spatial downscaling. *Biometrics*, **70**, 932–942.

Ren, Q. and Banerjee, S. (2013) Hierarchical factor models for large spatially misaligned data: A low-rank predictive process approach. *Biometrics*, **69**, 19–30.

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

Steenweg, R., Hebblewhite, M., Whittington, J., Lukacs, P. and McKelvey, K. (2018) Sampling scales define occupancy and underlying occupancy–abundance relationships in animals. *Ecology*, **99**, 172–183.

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D. and Kelling, S. (2009) eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, **142**, 2282–2292.

Thorson, J. T., Ianelli, J. N., Larsen, E. A., Ries, L., Scheuerell, M. D., Szuwalski, C. and Zipkin, E. F. (2016) Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring. *Global Ecology and Biogeography*, **25**, 1144–1158.

Thorson, J. T., Munch, S. B. and Swain, D. P. (2017) Estimating partial regulation in spatiotemporal models of community dynamics. *Ecology*, **98**, 1277–1289.

Turner, M. G. (1989) Landscape ecology: the effect of pattern on process. *Annual review of ecology and systematics*, **20**, 171–197.

Waller, L. A. and Gotway, C. A. (2004) *Applied spatial statistics for public health data*, vol. 368. John Wiley & Sons.

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

Wikle, C. K. and Berliner, L. M. (2005) Combining information across spatial scales. *Technometrics*, **47**, 80–91.

Wilson, A. M., Brauning, D. W. and Mulvihill, R. S. (2012) Second atlas of breeding birds in Pennsylvania.

Yoccoz, N., Nichols, J. D. and Boulinier, T. (2001) Monitoring of biological diversity in space and time. *Trends in Ecology and Evolution*, **16**, 446–453.

Young, L. J. and Gotway, C. A. (2007) Linking spatial data from different sources: the effects of change of support. *Stochastic Environmental Research and Risk Assessment*, **21**, 589–600.

Zipkin, E. F., Rossman, S., Yackulic, C. B., Wiens, J. D., Thorson, J. T., Davis, R. J. and Grant, E. H. C. (2017) Integrating count and detection–nondetection data to model population dynamics. *Ecology*, **98**, 1640–1650.

Zipkin, E. F. and Saunders, S. P. (2018) Synthesizing multiple data types for biological conservation using integrated population models. *Biological Conservation*, **217**, 240–250.

Zurell, D., Thuiller, W., Pagel, J., Cabral, J. S., Münkemüller, T., Gravel, D., Dullinger, S., Normand, S., Schiffers, K. H., Moore, K. A. et al. (2016) Benchmarking novel approaches for modelling species range dynamics. *Global Change Biology*, **22**, 2651–2664.

Table 1**Simulation study results: Single Data Source with Spatially Aggregated Covariate:** Here we are exploring the consequences of rescaling a covariate with a single data source. The data generation depends on the dimension of the aggregate cells (k) and the CAR spatial dependence of the covariate process (ρ); the two methods are the naive method that models the process only at the course resolution and the change of the support ("COS") method that models the process on the fine resolution. Methods are compared using Bias, mean squared error ("MSE") and coverage of 90% intervals for the covariate effect parameter, β_2.

Settings		Bias		MSE		Coverage	
k	ρ	Naive	COS	Naive	COS	Naive	COS
2	0.50	0.69	0.14	1.57	0.58	0.89	0.88
	0.99	1.02	0.25	1.80	0.39	0.79	0.86
3	0.50	0.34	0.09	1.77	1.27	0.94	0.92
	0.99	1.07	0.41	2.11	0.76	0.84	0.85

Table 2**Simulation study results: Accounting for COS in ISDMs:** Here we are interested in evaluating the consequences of ignoring COS in fitting Integrated Species Distribution Models. The data generation depends on the dimension of the aggregate cells (k), the detection probability (p) and the MCAR dependence parameter (ρ); the five methods are the model that uses only one source of data ("Single"), the three change of support methods ("Shared", "Correlation" and "Covariate") and the data fusion methods that ignore change of support ("Shared - no COS" and "Correlation - no COS"). The Brier score and classification accuracy are the median of 100 simulated datasets for each scenario.

(a) Brier score

Settings				Change of support			No COS	
k	p	ρ	Single	Shared	Correlation	Covariate	Shared	Correlation
2	0.2	0.50	0.141	0.126	0.135	0.132	0.133	0.137
2	0.2	0.99	0.117	0.101	0.110	0.102	0.108	0.117
2	0.5	0.50	0.018	0.018	0.018	0.018	0.019	0.018
2	0.5	0.99	0.017	0.016	0.016	0.016	0.017	0.017
4	0.2	0.50	0.138	0.135	0.139	0.138	0.137	0.143
4	0.2	0.99	0.118	0.110	0.116	0.112	0.112	0.119
4	0.5	0.50	0.018	0.018	0.018	0.018	0.018	0.018
4	0.5	0.99	0.017	0.016	0.017	0.016	0.017	0.017

(b) Classification accuracy

Settings				Change of support			no COS	
k	p	ρ	Single	Shared	Correlation	Covariate	Shared	Correlation
2	0.2	0.50	0.775	0.802	0.794	0.789	0.791	0.786
2	0.2	0.99	0.820	0.852	0.833	0.848	0.840	0.823
2	0.5	0.50	0.981	0.981	0.981	0.981	0.980	0.981
2	0.5	0.99	0.982	0.982	0.982	0.981	0.980	0.981
4	0.2	0.50	0.777	0.784	0.786	0.780	0.781	0.780
4	0.2	0.99	0.821	0.836	0.824	0.831	0.834	0.820
4	0.5	0.50	0.981	0.981	0.981	0.981	0.981	0.981
4	0.5	0.99	0.982	0.982	0.982	0.981	0.981	0.981

(c) CPU times (minutes)

Settings				Change of support			no COS	
k	p	ρ	Single	Shared	Correlation	Covariate	Shared	Correlation
2	0.2	0.50	2.57	2.75	4.54	2.56	2.54	4.27

Table 3**Spatial Resolutions for COS:**. We evaluated 4 different spatial resolutions to explore the consequences of spatial misalignment in ISDMs. Two different data sources were used which came from different spatial resolutions. BBA data came from the finest resolution (Grid 1) and eBird data was summarized for each of the other resolutions. We used these mismatches in scale to highlight the utility of accommodating COS in ISDMs.

Spatial Resolution	Grid Size (degrees)	Grid Size (km^2)
Grid 1	1/24 x 1/16	24.3
Grid 2	1/12 x 1/8	97.5
Grid 3	1/3 x 1/2	1553.6
Grid 4	2/3 x 1	6230.5

List of Figures

1 **Motivating data sets: Black-throated Blue Warbler in Pennsylvania, USA**:Example of combining two data sources collected at different spatial scales used in integrated species distribution modeling. The first data source is collected at fine resolution surveys (Breeding Bird Atlas) while the second data source (eBird) is summarized at a coarser resolution. 33

2 **Estimates of total occupancy probabilities from ISDMs incorporating COS vs ignoring COS at 4 different spatial resolutions**. The second data source is aggregated at different spatial scales (Grids 1 - 4) to mimic different degrees of spatial misalignment. The solid red line represents the point estimate of occupancy probability from the full, primary data source (BBA) with uncertainty (95% CI: dotted red line). ”Avg” and ”Scaled” in model names (y axis) refer to method used to summarize the second data source (see case study description for complete details). COS models are separated from No COS models by a horizontal dashed line. 34

3 **Occupancy estimates from the single, covariate, shared no COS, and shared COS models with eBird summarized at coarsest resolution (grid size = 6231 sq km)**: The plot in the upper left panel shows the distribution of occupancy for the full, primary data source (BBA). Clockwise, the plot in the upper right panel shows the results from the covariate model, shared COS, and shared no COS in bottom left. 35

4 **Occupancy estimates from the covariate model with second data source summarized in two different ways at grid level 2 (98 sq km)**: The plot in the upper left panel shows the results from the covariate model (’Cov-Avg’) with the second data source averaged across all smaller sites, with uncertainty on upper right panel. The lower left panel shows results from the covariate model (’Cov-Scaled’) with the covariates summarized by scaling up the covariate values from the finer resolution to match the larger grid size. 36

5 **Estimated effect of % forest cover under data fusion models which incorporate or ignore spatial misalignment in ISDMs.** The second data source is aggregated at different spatial scales (Grids 1 - 4) to mimic different degrees of spatial misalignment. Single model is provided as a reference to the estimated effect using full BBA dataset. ”Avg” and ”Scaled” in model names (y axis) refer to method used to summarize the second data source (see case study description for complete details). Solid red vertical line placed at 0 provided to visualize significance of the effect of % forest cover on occupancy probability. Dashed black horizontal lines used to separate models using one data source and COS and No COS models. . . . 37

Figure 1

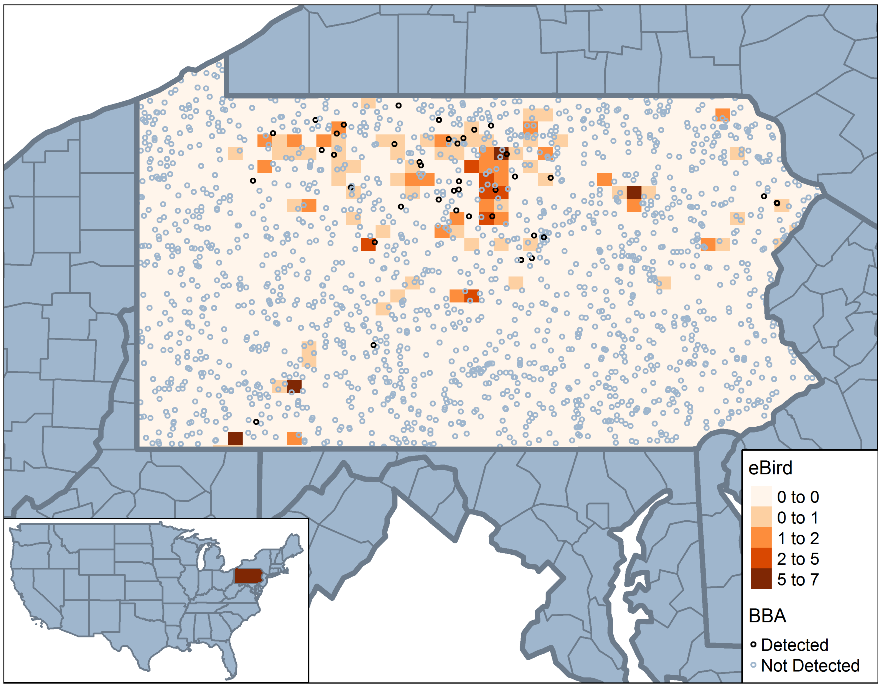


Figure 2

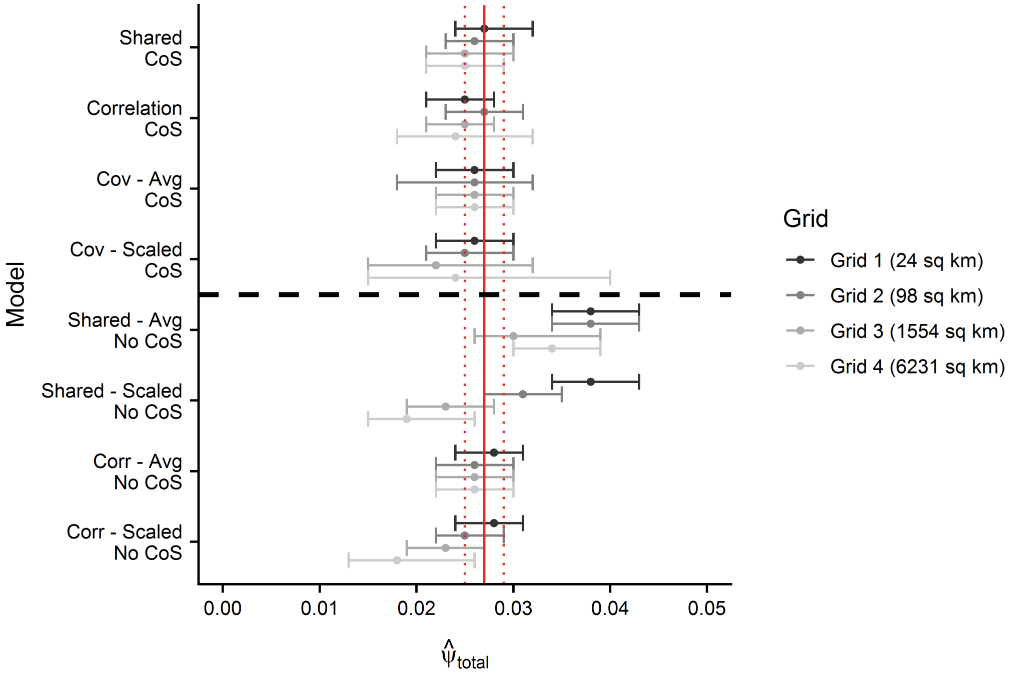

Shared
CoS
Correlation
CoS
Cov - Avg
CoS
Cov - Scaled
CoS
Shared - Avg
No CoS
Shared - Scaled
No CoS
Corr - Avg
No CoS
Corr - Scaled
No CoS
Model
0.00
0.01
0.02
0.03
0.04
0.05
$\hat{\psi}_{total}$
Grid
Grid 1 (24 sq km)
Grid 2 (98 sq km)
Grid 3 (1554 sq km)
Grid 4 (6231 sq km)

Figure 3

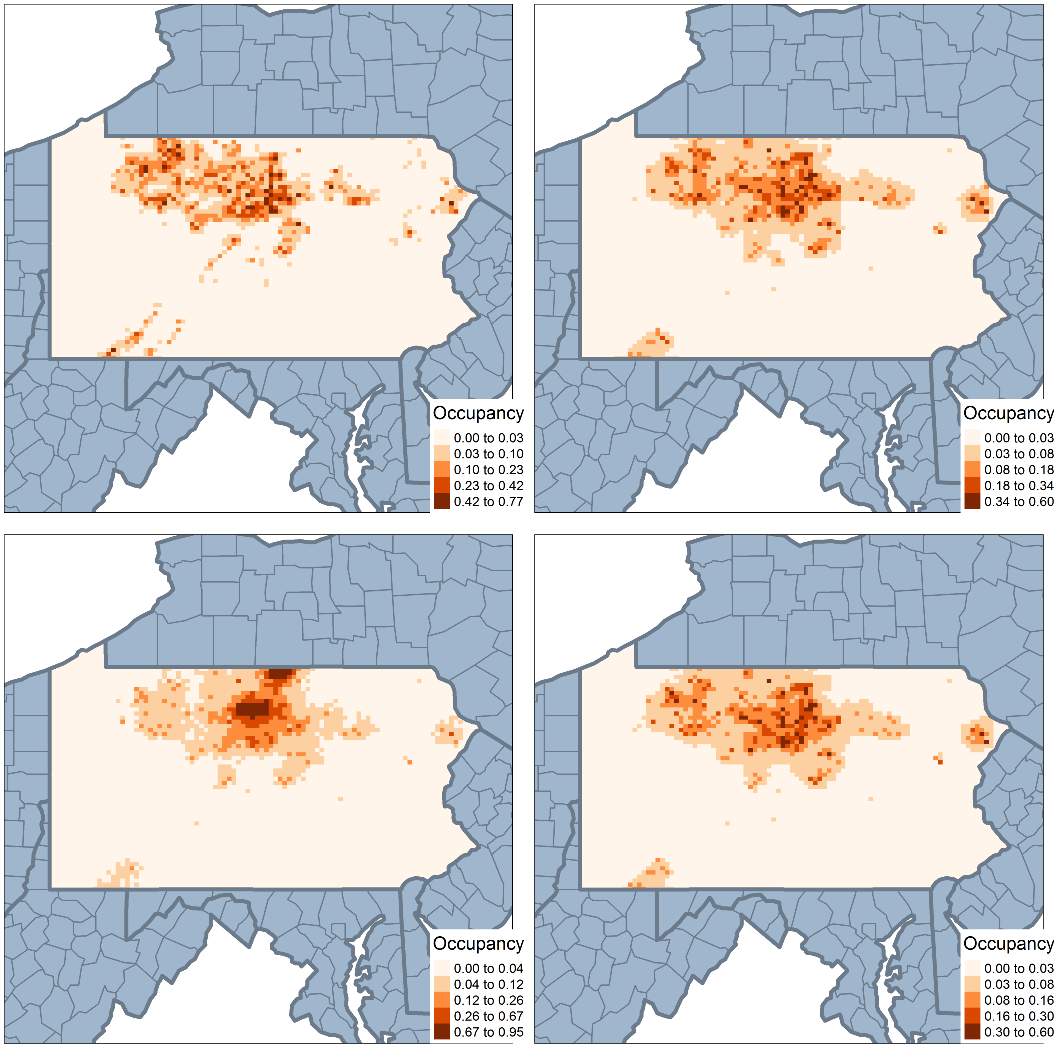

Occupancy
0.00 to 0.03
0.03 to 0.10
0.10 to 0.23
0.23 to 0.42
0.42 to 0.77
Occupancy
0.00 to 0.03
0.03 to 0.08
0.08 to 0.18
0.18 to 0.34
0.34 to 0.60
Occupancy
0.00 to 0.04
0.04 to 0.12
0.12 to 0.26
0.26 to 0.67
0.67 to 0.95
Occupancy
0.00 to 0.03
0.03 to 0.08
0.08 to 0.16
0.16 to 0.30
0.30 to 0.60

Figure 4

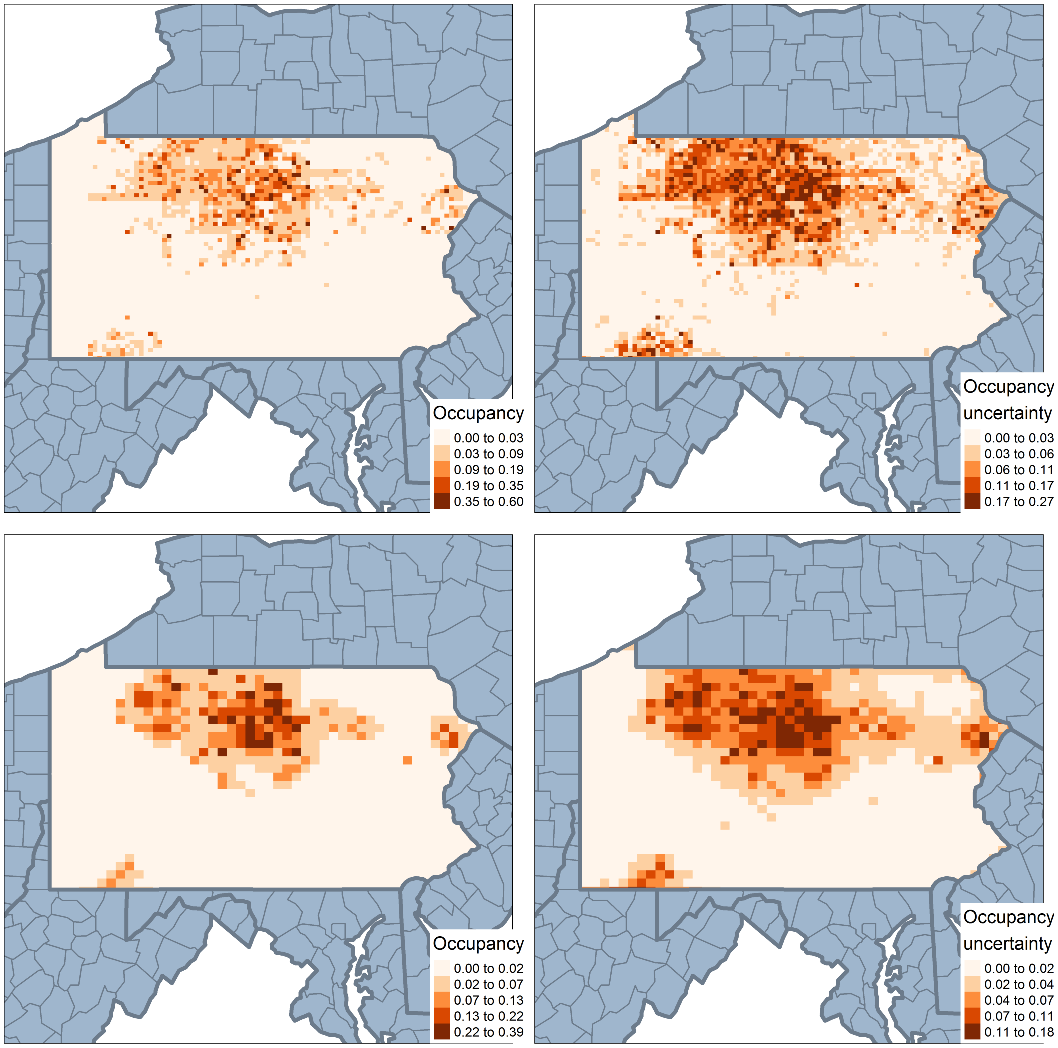
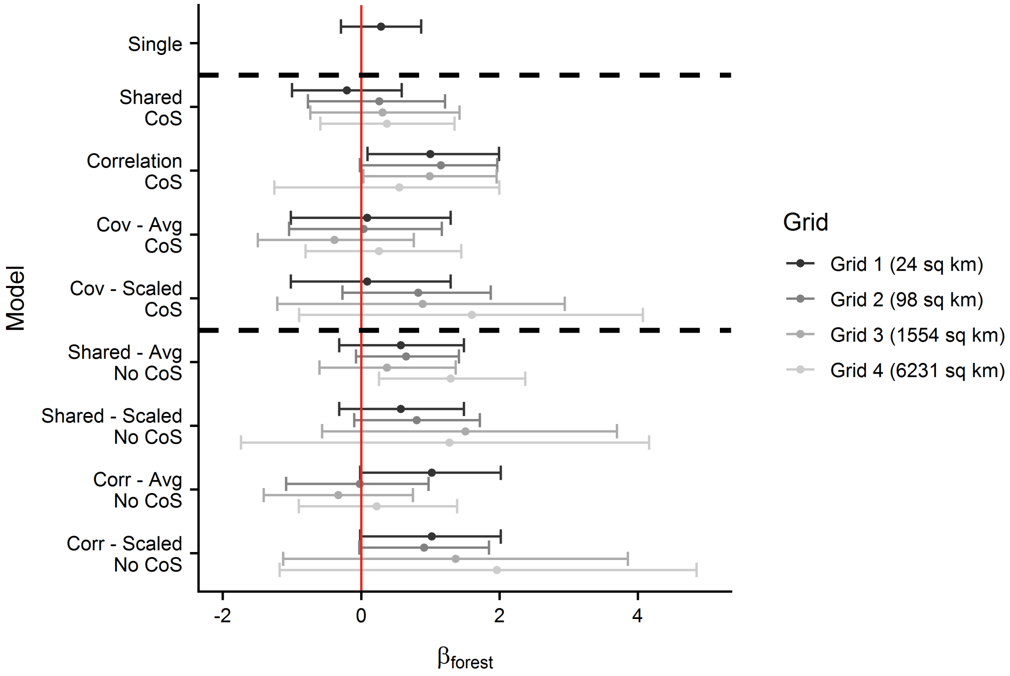

Occupancy
0.00 to 0.03
0.03 to 0.09
0.09 to 0.19
0.19 to 0.35
0.35 to 0.60
Occupancy uncertainty
0.00 to 0.03
0.03 to 0.06
0.06 to 0.11
0.11 to 0.17
0.17 to 0.27
Occupancy
0.00 to 0.02
0.02 to 0.07
0.07 to 0.13
0.13 to 0.22
0.22 to 0.39
Occupancy uncertainty
0.00 to 0.02
0.02 to 0.04
0.04 to 0.07
0.07 to 0.11
0.11 to 0.18

Figure 5


Single
Shared CoS
Correlation CoS
Cov - Avg CoS
Cov - Scaled CoS
Shared - Avg No CoS
Shared - Scaled No CoS
Corr - Avg No CoS
Corr - Scaled No CoS
Model
-2
0
2
4
β_{forest}
Grid
Grid 1 (24 sq km)
Grid 2 (98 sq km)
Grid 3 (1554 sq km)
Grid 4 (6231 sq km)