Received: 19 March 2018 | Accepted: 2 October 2018

DOI: 10.1111/2041-210X.13110

ADVANCES IN MODELLING DEMOGRAPHIC PROCESSES

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

The recent past and promising future for data integration methods to estimate species' distributions

David A. W. Miller[1] | Krishna Pacifici[2] | Jamie S. Sanderlin[3] | Brian J. Reich[4]

[1]Department of Ecosystem Science and Management, Penn State University, University Park, Pennsylvania

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

[3]Rocky Mountain Research Station, USDA Forest Service, Flagstaff, Arizona

[4]Department of Statistics, North Carolina State University, Raleigh, North Carolina

Correspondence
David A. W. Miller
Email: dxm84@psu.edu

Handling Editor: Beth Gardner

Abstract

1. With the advance of methods for estimating species distribution models has come an interest in how to best combine datasets to improve estimates of species distributions. This has spurred the development of data integration methods that simultaneously harness information from multiple datasets while dealing with the specific strengths and weaknesses of each dataset.
2. We outline the general principles that have guided data integration methods and review recent developments in the field. We then outline key areas that allow for a more general framework for integrating data and provide suggestions for improving sampling design and validation for integrated models.
3. Key to recent advances has been using point-process thinking to combine estimators developed for different data types. Extending this framework to new data types will further improve our inferences, as well as relaxing assumptions about how parameters are jointly estimated. These along with the better use of information regarding sampling effort and spatial autocorrelation will further improve our inferences.
4. Recent developments form a strong foundation for implementation of data integration models. Wider adoption can improve our inferences about species distributions and the dynamic processes that lead to distributional shifts.

KEYWORDS
data fusion, integrated distribution model, joint likelihood, spatial point process, species distribution modelling

1 | INTRODUCTION

> Perhaps out of the two approaches, we thought, there might emerge a picture more complete and even more accurate than either alone could produce. And so we went.
>
> Sea of Cortez—Steinbeck and Ricketts (1941)

Species distributional modelling has its roots dating back to niche models first developed more than a century ago (Grinnell, 1917). However, the field has come of age as an ecological and management tool in the past two decades (Elith & Leathwick, 2009; Guillera-Arroita et al., 2015; Guisan & Thuiller, 2005). Species distribution models (SDMs) are used to infer range extent and niche breadth, as well as to predict the effects of changing climate (Pearson & Dawson, 2003; Thomas et al., 2004) and spread after invasive species introductions (Ficetola, Thuiller, & Miaud, 2007), to quantify community interactions (Clark, Gelfand, Woodall, & Zhu, 2013; Pollock et al., 2014), and to inform management of threatened species (Dicko et al., 2014; Guisan et al., 2013). Many of the most heavily referenced ecological papers in recent years define methods used

for modelling species distributions (e.g. MacKenzie et al., 2002; Guisan & Thuiller, 2005; Elith et al., 2006; Phillips, Anderson, & Schapire, 2006, among others). The result is that SDMs now incorporate many types of data, functional forms for covariate relationships, and methods to account for uncertainty.

In parallel to method development are extensive ongoing efforts to generate, catalog and disseminate data streams that provide these raw data needed to build species distribution models (Edwards, Lane, & Nielsen, 2000; Sullivan et al., 2009). Ideally, we would collect these data using systematic study designs, where

TABLE 1 Definitions of terms used to describe integrated species distribution models

Measures of distribution	
Abundance	Number of individuals within an area (*density* is abundance scaled by area)
Occurrence	Whether the focal species occurs within an area (also *occupancy* or *presence*)
Index	A metric positively correlated with abundance or probability of occurrence
Static distribution	The distribution of individuals at a single point in time
Dynamic distribution	The distribution of individuals as it changes across time (measures of change, including *local extinction* and *local colonization*, intrinsic rate of increase, or expansion and contraction of species' range). Change in distribution is a function of the individual processes with demographic parameters (survival, reproduction, and movement)
Data	
Standardized data	Data collected using a standardized sampling design and fixed protocol at known sampling locations
Nonstandardized data	Data not collected under standardized protocol, where sampling locations and sampling effort are often unknown and sampling protocol varies
Detection/Nondetection data	Recorded observations of whether or not a species was observed during a given sampling occasion (note this is not the same as true presence)
Count data	Observations of number of individuals observed during a given sampling occasion
Presence-only data	Observations only include locations where the species was observed
Cell	The spatial extent of sample units for which observations are gathered and predictions made (also *grain* or *site*)
Domain	The extent of the study area for which estimates are generated
Z/N/B	Denotes the true state of a location. *Z* for presence–absence, *N* for abundance, and *B* for biomass
Y/C/b	Denotes the observed data for a sampling occasion. *Y* for detection nondetection, *C* for count data, and b for biomass collections
Observation uncertainty	
False-negative	Species is present in a given cell, but no individuals are detected. (also *sensitivity*)
False-positives	Species is not present in a given cell, but it is recorded as being present. (also *mis-identification* or 1-*specificity*)
Effort	A direct or indirect measure of sampling intensity at a site. (e.g. survey hours, no. observers, or total individuals collected for all species). Costs associated with sampling intensity are an important component of effort
Location error	Mis-specification of the spatial location of observations
Integrated estimator types	
Joint Likelihood	Estimator where parameters are constrained to be equal across data types and parameter estimates maximize fit across all data types
Data-weighting	A joint likelihood estimator where likelihoods for individual datasets are differentially weighted with weights assigned a priori or based on model validation of fit
Covariance	Estimator where parameter values are specified to as correlated across data types
Covariate	Estimator where values from one data type are used to generate a set of predictors used in the model for another data type
Offsets	Additional information from a second data type is included as an offset in a log-linear model
Priors	Additional information for a second data type is included as Bayesian priors in a species distribution model
Ensemble	Multiple model estimates are averaged post hoc to generate a consensus estimate
Spatial Autocorrelation	Accounting for additional spatial structure and cross-correlation between parameters using models of spatial autocorrelation
Model checks	
Goodness-of-fit	Measures of how well a model meets assumptions about the data. Are data a realistic reflection of parameter values estimated by the distribution model?
Validation	Measures of how well a model predicts independent data collected for the system

the sample location, effort, and sampling methods follow a standardized protocol (hereafter, standardized data; see Table 1 for terminology and definitions used in this paper). Some examples of large-scale standardized datasets include the North American Breeding Bird Survey (Sauer et al., 2017), North American Forest Inventory and Analysis (Smith, 2002), and camera trap data collected by the Tropical Ecology Assessment and Monitoring Network (Ahumada et al., 2011). Standardizing data collection facilitates direct comparisons among locations, times, and conditions and more accurate accounting for observational uncertainty. However, for most species, nonstandardized data will be more abundant and widely distributed. For nonstandardized data collection methods, effort, collection protocol and even exact location may not be specified prior to collection and may not be included in collection records. Examples of widely available nonstandardized datasets include museum samples (Newbold, 2010), citizen science observations (Bird et al., 2014; Hochachka et al., 2012), and historical records (Tingley & Beissinger, 2009), as well as other opportunistic data collection methods. These data sources provide an incredible wealth of knowledge about where species occur now and in the past. However, inference from these datasets is challenged by our ability to account for observational uncertainty and to estimate true occurrence rates (Yackulic et al., 2013a).

Methods that combine and integrate these data are crucial to take advantage of the wealth of data now available (Dorazio, 2014; Pacifici et al., 2017; Zipkin & Saunders, 2018). Our goal is to facilitate efforts to integrate multiple data sources to estimate species' distributions. We review recent work on data integration, highlight key concepts underlying integrated estimators, and outline the process for combining data and testing results. In addition, we outline key innovations that we believe will usher a more general framework for data integration. Finally, we discuss approaches to study design and model validation in the context of data integration. Throughout, we place a strong emphasis on methods that combine standardized and nonstandardized data, leveraging the strengths that come with each of these data types.

2 | WHAT HAS BEEN DONE SO FAR?

> We wanted to see everything our eyes would accommodate, to think what we could, and, out of our seeing and thinking, to build some kind of structure in modeled imitation of the observed reality.
>
> Steinbeck and Ricketts (1941)

Integrated methods combine multiple datasets to improve predictions made about species distributions. Recent developments have been predicated on the following premises: (a) Species' distributions are the aggregated spatial locations of all individuals of the same species across a geographic domain; (b) The distribution can be described as a spatial point process in which the local intensity (i.e. density) of individuals varies. (c) Species distribution models are a direct or indirect model of this underlying point process. (d) Data integration requires linking each data source to the common underlying point process while accounting for differences among data types.

Before proceeding, we need to discuss two key concepts that underlie most integrated models: spatial point processes and joint likelihood methods. Each has a long history of use in statistics and ecological applications. However, their combined application to distribution modelling is relatively recent and has spurred much of the development in this area.

2.1 | A Unifying Framework: Spatial Point Processes

In the following we will highlight features of point processes that can be used to draw a link between occurrence and abundance, among different data types, and across different spatial extents (Dorazio, 2014; Hefley & Hooten, 2016; Renner et al., 2015). Detailed and thorough descriptions of spatial point processes can be found elsewhere (Cressie, 2015; Diggle, 2013; Dorazio, 2014; Hefley & Hooten, 2015; Renner et al., 2015).

A *spatial point process* describes the distribution of event locations across some spatial domain (e.g. locations of individuals of the same species). Points arise from a random process, described by the local intensity λ_s, which measures the expected density of points at a given location, *s*, in space. If points arise independently and at random, this local density can be described by a homogenous Poisson distribution and is referred to as a *Poisson point process*. If event locations are independent but the intensity varies spatially, the distribution arises from a *inhomogeneous point process* (λ_s varies). If all events are not observed, such as is the case when sampling and detection are incomplete, the observed points come from a *thinned point process*. Point processes can be described as continuous distributions or as discretized measures across a gridded set of cells. Modelling distribution data can directly employ a set of statistical methods developed to model continuous point processes (Renner et al., 2015); Or, by understanding the underlying distributions that govern the point process, one can specify relationships to other methods used to model species distributions.

We plot random realizations with different intensities for a homogenous Poisson point process in Figure 1. As is common for many species distribution modelling methods, we can discretize observations into a set of smaller cells. A couple of patterns emerge when this is done. Local abundance and occurrence depend on the intensity of the process and the cell size. As intensity increases and cell size increases, the proportion of occupied cells increases. Alternatively, when the expected number of individuals per cell is small, occupancy and density are nearly equivalent. This occurs when density is low enough so that cells rarely contain more than one individual.

If we know the underlying point generating process, we can also specify a mathematical relationship between density and occupancy and how the relationship changes based on the cell area. Table 2 shows the expected occupancy given local intensity and some data generating distribution. A Poisson distribution is usually the default

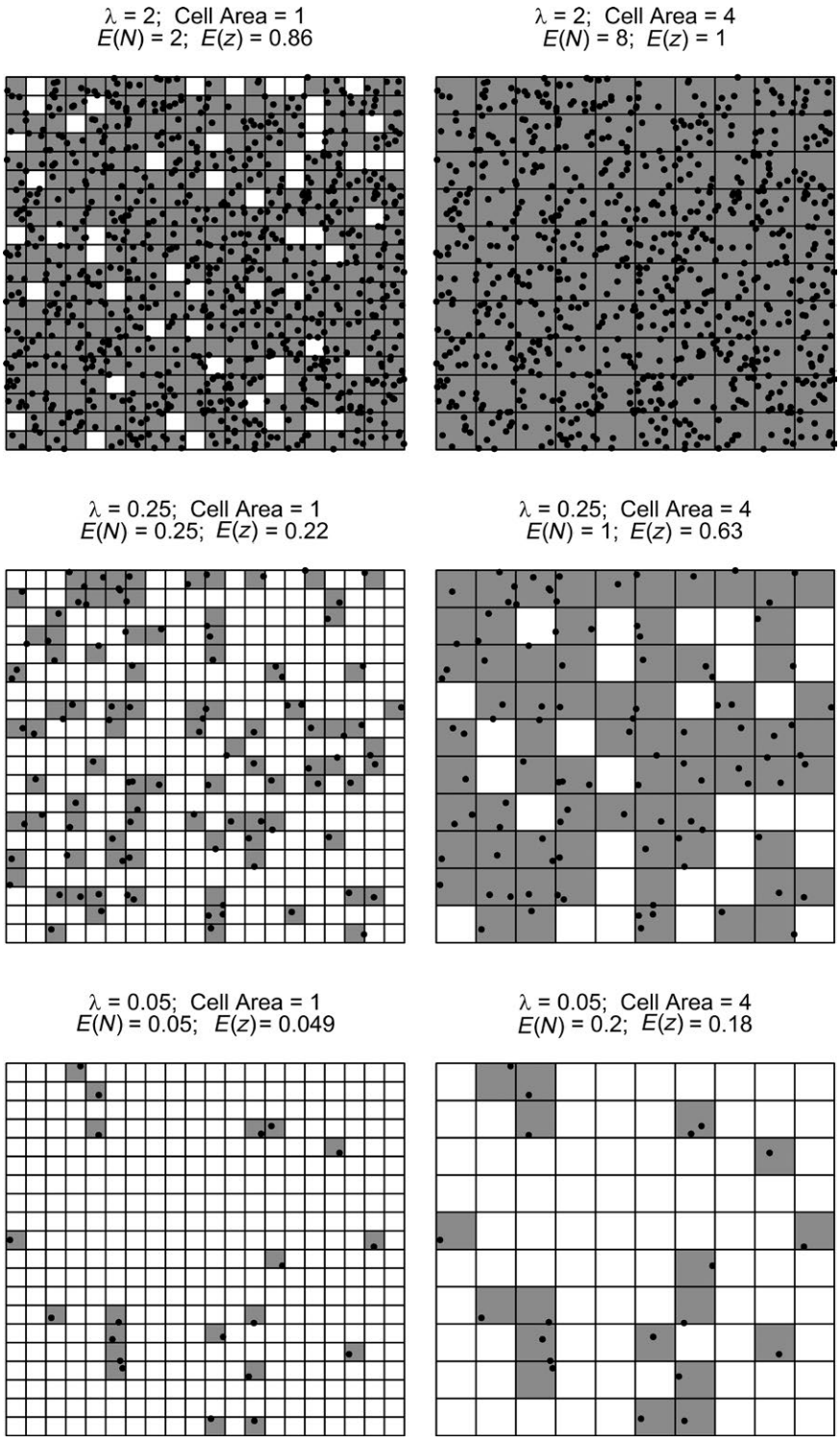

FIGURE 1 We simulated homogenous Poisson point processes for three levels of intensity, λ, and for two cell sizes. The expected abundance per cell, *E*(*N*), and expected proportion of occupied cells, *E*(*z*) depends on the combination of intensity and cell size

for integrated models. Counts will be distributed as a Poisson random variable regardless of cell size when animals occur randomly across space at a constant intensity (i.e. a homogenous point process). However, when over-dispersion occurs (zero-inflation or local clumping) it may be useful to consider other data generating distributions (e.g. Hostetler, 2015; Knapp et al., 2016), which in many cases are straight forward to fit if space is discretized. An extreme example of over-dispersion occurs when counts are not actual individuals but signals (e.g. scat samples or bird calls) clumped around the locations of individuals (Chambert, Waddle, Miller, Walls, & Nichols, 2017).

It is possible to link observations across different spatial scales. A useful result is that if occurrence is modelled as a function of local intensity, occupancy is no longer scale dependent, but instead can be determined based on cell area (Dorazio, 2014; Koshkina, Wang, Gordon, Dorazio, & White, 2017). Consider the following example for a Poisson point process. Let A be the area of the cell, λ the expected

TABLE 2 It is possible to link count-based approaches with occurrence-based approaches by determining the probability the count model is equal to 0 (i.e. the species does not occur at a site)

Distribution	Probability $z_i = 0/N_i = 0$
Poisson	$e^{-\lambda}$
Zero-inflated Poisson	$\sigma + (1 - \sigma)e^{-\lambda}$
Negative Binomial	$(1-p)^r$
Zero-inflated Negative Binomial	$\sigma+(1-\sigma)(1-p)^r$

Note. parameters are as follows: z_i is the true occurrence state of site i where $z_i = 1$ if the site is occupied; N_i is the true abundance of a site; λ is the intensity parameter for a Poisson distribution; p and r are parameters defining a negative binomial distribution; σ is zero inflation correction for a zero-inflated model.

density per unit of area, ψ expected occupancy in a cell of area 1. The expected number of individuals per cell is $E(N) = \lambda \times A$ and the expected probability a cell is occupied is $E(Z) = 1 - (1 - \psi)^A$. We can now combine this information to scale across distributions and areas. For example, expected occupancy for area A given λ is $E(Z) = 1 - (e^{-\lambda})^A$. And the expected density given ψ and a Poisson point process is $E(N) = -\ln(1 - \psi) \times A$. This ability to generate expectations across scales and data types creates a flexible framework for linking different data types and species distribution models. Similarly, it is possible to account for scale dependency in relationships to predictor variables when fitting integrated distribution models (Pacifici et al., 2017).

Recent studies demonstrate it is possible to use the toolbox developed for modelling spatial point processes to estimating distributions for many data types (Dorazio, 2014; Fithian, Elith, Hastie, & Keith, 2014; Hefley & Hooten, 2015, 2016; Renner & Warton, 2013; Renner et al., 2015; Warton, Renner, & Ramp, 2013). The result is that if it is possible to link multiple data types to a common point process, it is possible to combine the methods in an integrated estimator.

2.2 | Joint-likelihood methods

Joint-likelihoods have been the preferred methods for which multiple data types are integrated in a single estimator. The joint likelihood approach uses multiple data types to simultaneously estimate a shared set of parameters (Figure 2). For each data type, a likelihood is specified that is to be maximized or fit using Bayesian methods. In joint likelihood methods the shared set of parameters within each of the individual likelihoods is constrained to be equal across likelihoods. Parameter estimates are those that maximize the combined fit of all data. It is possible to develop shared parameterizations across a wide set of data types and spatial resolutions. This may involve different measured response variables (e.g. probability of occurrence vs. local density) and responses measured at different spatial resolutions.

Joint-likelihood methods are not new to ecological parameter estimation (see Table 3 for examples of joint-likelihood methods to estimate ecological parameters). The most analogous example might be integrated population models (not to be confused with integral projection models [Ellner & Rees, 2006]), which combine data types (often

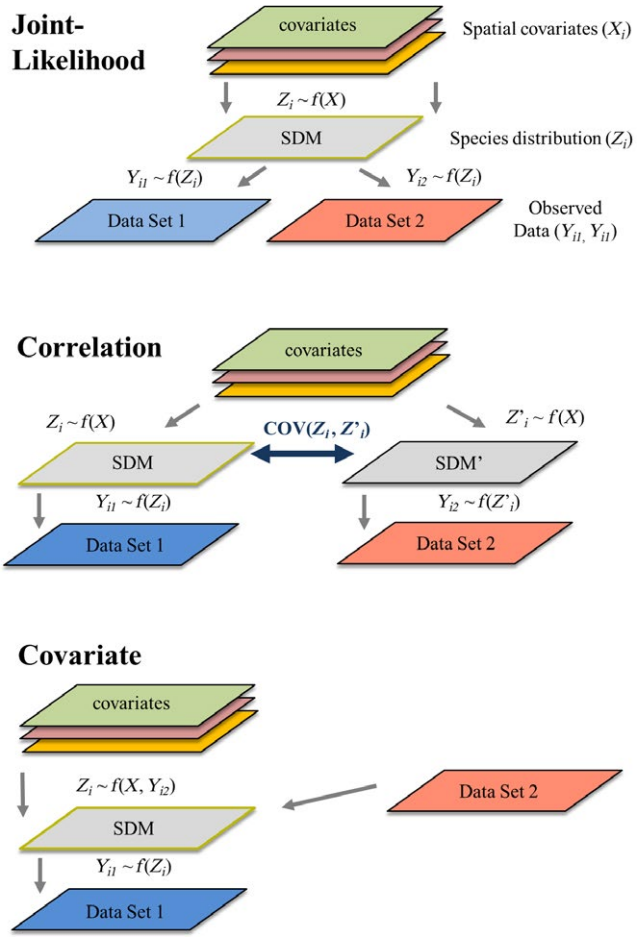

FIGURE 2 Data integration methods have typically relied on joint-likelihood methods, where two or more data types are related to a common underlying species distribution. Efforts based on specifying correlation or covariate structure are possible in cases where joint likelihoods cannot be specified or where both data types should not be put on an equal footing. In the case of the correlation method, a separate species distribution is estimated for each data type in the analysis, with a correlation structure specified to share information across the datasets. In the case of the covariate method, the first dataset is used to generate covariates that are then used to predict the distribution of the second dataset. In the case of both, the correlation and covariate models, the information content of the second dataset is estimated

mark-recapture and count or abundance data) to improve precision of estimates of population growth rate and underlying demographic parameters (Besbeas, Freeman, Morgan, & Catchpole, 2002; Besbeas, Lebreton, & Morgan, 2003; Schaub, Gimenez, Sierro, & Arlettaz, 2007).

2.3 | Recent efforts at data integration

Recent efforts to construct species distribution models using data integration methods have primarily been motivated by two goals: (a) build a more accurate SDM and (b) relax assumptions, increase identifiability, and overcome limitations that might come with models incorporating only nonstandardized data. We recognize that species distribution models often lump multiple data types into the set of

TABLE 3 Examples of data types and demographic parameters for studies combining multiple data sources. Although not exhaustive, this list provides an overview of different types of studies that combine multiple data sources with joint-likelihood methods

Data types	Demographic parameters	Example(s) from the literature
Mark-recapture-recovery or capture–recapture, count data, and/or fecundity data	Abundance, survival, reproduction	Besbeas et al. (2002, 2003), Brooks, King, and Morgan (2004), Schaub et al. (2007a), King, Brooks, and Coulson (2008), Schaub and Abadi (2010), Abadi et al. (2012), Wilson, Gil-Weir, Clark, Robertson, and Bidwell (2016), Ahrestani et al. (2017), and Eacker, Lukacs, Proffitt, and Hebblewhite (2017)
Mark-recapture-recovery or capture–recapture, count data	Survival	Catchpole, Freeman, Morgan, and Harris (1998)
Radiotelemetry, mark-recapture data	Survival, movement rates	Powell, Conroy, Hines, Nichols, and Krementz (2000)
Secondary sources of data	Detection probability in occupancy models	Ruiz-Gutierrez et al. (2016)
Count, presence–absence data	Abundance, survival, reproduction	Zipkin et al. (2017)
Capture–recapture data, presence–absence data	Abundance, survival, reproduction	Sanderlin et al. (2018) and Freeman and Besbeas (2012)
Spatial capture–recapture	Density	Gopalaswamy et al. (2012), Sollmann et al. (2013a,b), and Chandler and Clark (2014)
Capture–recapture, presence–absence data	Abundance	Blanc, Marboutin, Gatti, Zimmermann, and Gimenez (2014)

observations used for estimation (e.g. combining multiple sources of nonstandardized data into a presence-only estimator). However, for our purposes we will only consider estimators that treat each data type on an individual basis, dealing with the intricacies that went into collecting data from that source.

The ability to combine data types has been recognized since the early implementation of occupancy estimation methods. The ability to treat data collected from different methods as independent visits, with separate detection probabilities but a shared underlying distribution of occurrence has been recognized as a valid way to estimate occupancy (MacKenzie et al., 2006). The ad hoc integration of data types has been expanded to more fully deal with specific nuances of multiple data types by Nichols et al. (2008) and to improve estimates when both false negatives and false positives occur in a data type (Hanks, Hooten, & Baker, 2011; Miller et al., 2011, 2013, 2015). Although these can be thought of as joint-likelihood methods, the use of a common estimation framework (i.e. occupancy estimation) makes data integration relatively straight-forward.

More recently, efforts have focused on combining data from designed surveys and presence-only data (Dorazio, 2014; Fithian et al., 2014; Fletcher, Robert, Greene, & Tye, 2016; Koshkina et al., 2017; Pacifici et al., 2017). A common approach has been to treat presence-only data as coming from a thinned point-Poisson process and jointly estimating the observation intensity with the response variable for the standardized dataset. This approach was predicated on the development of methods for analysing presence-only data using a point-process framework. Point process models have been adapted to estimate distributions and model observation error from presence-only data (Warton & Shepherd, 2010; Warton et al., 2013) as an alternative to MaxEnt (Phillips, Dudík, & Schapire, 2004; Phillips et al., 2006). Other recent papers build on these methods (e.g. Fithian & Hastie, 2013; Renner et al., 2015) including work by Renner and Warton (2013) that demonstrate the functional equivalence of point process and MaxEnt methods.

This flexibility to model presence-only data using a likelihood framework underlies the more recent advances that implement data integration for multiple data types. These models focus on combining presence-only with standardized datasets where observation error is explicitly addressed. Examples include using occurrence data where detection is assumed to be 1 (Fithian et al., 2014), an *N*-mixture model to estimate abundance from repeat count data (Dorazio, 2014) and spatial capture recapture data (Tenan, Pedrini, Bragalanti, Groff, & Sutherland, 2017), and occupancy data from temporal and spatial replicated observations (Fletcher et al., 2016; Koshkina et al., 2017; Pacifici et al., 2017).

Other recent examples exploit the relationship between individual abundance and probability of detecting a species that occurs at a site (i.e. Zipkin et al., 2017). This relationship has previously been recognized in the occupancy literature (Royle & Nichols, 2003). Pacifici et al. (2017) also exploit this relationship, although indirectly through correlation structure rather than a joint parameterization structure. Clare, McKinney, Depue, and Loftin (2017) address another key issue for data integration, which is when observations from two datasets are not independent.

Consistent among these studies is that when data integration has been employed, the integrated model represented an improvement over using a single data source (increased accuracy, increased precision). In addition, linking presence only observations to a standardized dataset allows for local density or occurrence probability to be directly estimated rather than just an estimate of relative occurrence probability. This feature is especially appealing, as methods that attempt to derive true occurrence probabilities based on presence-only data and single surveys with incomplete detection (e.g. Lele & Keim, 2006; Lele, Moreno, & Bayne, 2012; Royle, Chandler, Yackulic, & Nichols, 2012)

have come under criticism for strong assumptions (Hastie & Fithian, 2013; Phillips & Elith, 2013; Ward, Hastie, Barry, Elith, & Leathwick, 2009). Likewise correcting for relative sampling bias with presence only data requires that predictors be known and uncorrelated with the true distribution (Dorazio, 2012). Data integration can make detection and sampling effort identifiable, as well as true occurrence probabilities, by conditioning on the standardized dataset. Fithian et al. (2014) employ a unique approach that also incorporates information from other species to identify where sampling effort has been greatest, while Pacifici et al. (2017) model effort using covariates derived from the data source they employ. We explore this in more detail in the next section.

3 | CREATING A MORE FLEXIBLE AND GENERAL FRAMEWORK FOR DATA INTEGRATION

> Let's go wide open. Let's see what we see, record what we find, and not fool ourselves with conventional scientific strictures.
>
> Steinbeck and Ricketts (1941)

Recent developments lay a strong foundation for data integration methods. For the most part, this effort has focused on simple detection nondetection or count data and joint-likelihood approaches to link other data types to presence-only data. Our goal now is to draw attention to other concepts that we believe can underlie an even more general framework for future developments of data integration methods.

3.1 | Rethinking what are distribution data

Many of the methods used to estimate abundance or occurrence are prohibitively costly to implement when the interest is in understanding patterns at the scale of a species' range. This precludes their use when the goal is to estimate large-scale patterns of distribution. As a result most large-scale distribution models rely solely on presence only and presence absence data collected using nonstandardized designs. Data integration opens the possibility to combine spatially rich data sources with a more focused set of standardized data. We believe this will improve our inferences about species distribution and will provide added resolution regarding spatial variation in density as well as patterns of occurrence.

There are a range of data types and estimators that can be used to estimate species abundance. However, they are rarely used for distribution models because limitations prohibit collecting data across a species' range. In general, abundance estimation methods rely on either mark-recapture or count-based methodology. Mark-recapture uses three general classes of models, depending on if the population is closed to births, deaths, immigration and/or emigration. These are closed abundance estimators (Otis, Burnham, White, & Anderson, 1978), open abundance estimators (Jolly, 1965; Schwarz & Arnason, 1996; Seber, 1965) or both (Kendall, Pollock, & Brownie, 1995; Pollock, 1982). Mark-recapture data also allow one to obtain estimates of survival and reproduction of individuals (Jolly, 1965; Schwarz & Arnason, 1996; Seber, 1965), key demographic parameters influencing the spatial distribution of individuals through time. More recently spatially explicit capture recapture methods have been used to estimate density across a defined spatial domain (Efford, Borchers, & Byrom, 2009; Royle, Chandler, Sollman, & Gardner, 2014; Royle, Chandler, Sun, & Fuller, 2013). Alternatively, distance sampling has emerged as a standardized approach for estimating density based on locations of observations relative to the observer (Buckland et al., 2001). Other count-based methods include N-mixture models (Royle, 2004), double-observer methods (Nichols, Hines, Sauer, Fallon, & Heglund, 2000), and removal designs (Farnsworth et al., 2002). All these methods estimate abundance or density at some given spatial scale and can readily be linked to other data types usually collected at much larger spatial scales, including those used for presence only data, by relating these patterns to the underlying point process. Alternatively, fisheries and plant applications often focus on biomass, which similarly might be combined with data across scales or be linked to density or occurrence data.

Opportunities for data integration are not limited to traditional static species distribution models, but also could be utilized when distribution dynamics are of interest (Zipkin et al., 2017). Dynamic models range from simple abundance models that estimate local changes in occurrence or abundance across space and time (Amburgey et al., 2018; Miller et al., 2018) to more complex models that incorporate demographic parameters and life-stage-specific abundances (Davis, Hooten, Phillips, & Doherty, 2014; Zipkin et al., 2014). Changes in the distribution of species through time (expansion, contraction) are influenced by individual population processes governed by survival, reproduction, and movement. Presence-only data have been under-utilized for estimation of dynamics because it is difficult to disentangle changes in effort and detection from actual changes. However, single-species occupancy models (MacKenzie et al., 2002) can be used to evaluate changes in occupancy from local extinction and colonization, while accounting for imperfect detection. Recent efforts with single-species occupancy models (Dail & Madsen, 2011; Rossman et al., 2016; Zipkin et al., 2014) allow one to estimate occupancy/abundance, survival, and reproduction with occupancy only data. Integrated approaches with large-scale efforts to collect multiple types of data (i.e. BBS count data and capture–recapture data from the Monitoring Avian Productivity and Survivorship [MAPS] program illustrated in Ahrestani, Saracco, Sauer, Pardieck, and Royle (2017) could be used to evaluate changes in species distribution.

Data integration methods do not need to be limited to single-species distribution models. Community approaches to estimating distributions have been especially promising for sharing information among species about both where species occur and how they are observed given the data collection method (Fithian et al., 2014; Iknayan, Tingley, Furnas, & Beissinger, 2013). Multi-species occupancy models (Dorazio, Royle, Söderström, & Glimskär, 2006; MacKenzie et al., 2006) extend single-species occupancy models (MacKenzie, Nichols, & Hines, 2003; MacKenzie et al., 2002) and allow estimation of unknown species richness and species turnover (local extinction and

colonization), while accounting for imperfect detection probabilities (Royle & Dorazio, 2008). Another set of community modelling methods focus on estimating species interactions (Clark et al., 2013; Lapointe, Giroux, Bélanger, & Filion, 2000; Mackenzie, Bailey, & Nichols, 2004; Ovaskainen, Roy, Fox, & Anderson, 2016; Rota et al., 2016; Warton et al., 2015). Information about the occurrence or abundance of one species informs the presence of another. Similarly, these models have been extended to incorporate community dynamics (Davis et al., 2017; Miller, Brehme, Hines, Nichols, & Fisher, 2012; Yackulic et al., 2013b). These methods could naturally be extended to incorporate data integration methods.

3.2 | Flexible data integration methods

Most of the effort devoted to integrated models has focused on joint-likelihood approaches. However, Pacifici et al. (2017) and Merow, Wilson, and Jetz (2016) introduce data integration approaches that relax the need to specify parameter relationships directly. Why is this useful? First, it is not uncommon to collect nonstandardized data that is informative but for which it may not be possible to develop an unbiased estimator of species distribution. The goal in this case is to harness the information from nonstandardized data while letting these data tell us the degree to which that information informs the distribution. Second, it is not always possible to specify a shared set of parameters. It may instead be possible to specify a set of parameters that will be correlated across data types based on the shared underlying species distribution. Finally, in some cases it may be computationally infeasible to combine data types into a single joint-likelihood estimator. In this case, alternative methods that incorporate additional information may reduce computation time while still improving accuracy and precision of distribution estimates.

The first alternative approach requires the specification of a likelihood for each dataset. However, instead of constraining parameters to be equal as is done in joint-likelihood methods, one can specify a correlation structure between parameters in the two models (Pacifici et al., 2017; Figure 2). Specifically, the spatial variation in the state variable for the two datasets is assumed to be correlated (i.e. if correlation is high, then where abundance or occurrence is greatest in one dataset will match where it is greatest in the other). Consider data types used in the estimator of Dorazio (2014). The goal was to build an integrated model based on density estimates from an N-mixture and presence-only data types. Now rather than constraining local densities estimated for each cell across the landscape to be the same for each data type, they could instead be specified as correlated, where the degree of correlation is estimated. If the estimated correlation is high (i.e. $\rho \sim 1$), this would suggest the two models are estimating the same process and information sharing will be high. In cases where unexplained variation in effort or other sources of heterogeneity occur in the nonstandardized data, the estimated correlation should be <1 and the degree of information transfer to estimates from the standardized dataset reduced. Specifying a correlation structure between datasets may also be useful when spatial grain of observations varies greatly among datasets. For example, spatial location from broad-scale surveys such as eBird data may be imprecise, suggesting that distribution be estimated for large-grain sizes while avian point count surveys, say from BBS, may have very precise locations where observations occur. It may be tenuous to assume both datasets can describe common parameters in this case (see the worked example in Supporting Information Appendix A for an example of this).

An alternative approach also suggested by Pacifici et al. (2017) is to incorporate the second dataset as a set of predictors for the distribution model for the first dataset (Figure 2). The covariate or covariates are used as explanatory variables in the distribution model the same way that other environmental predictors are included. The idea is to generate a set of covariates from one dataset and determine the ability to predict the distribution for the second. The covariate used may be these data directly incorporated, a summary of these data, or even a distribution model built using these data. This approach can be implemented using standard analytical packages, negating the need to develop new code, reducing both time and effort needed to generate integrated estimates.

Other approaches also fall under this category. The first is the offset approach employed by Merow et al. (2016) to integrate expert opinion data into distribution estimates. By specifying an offset within a log-regression model, they were able to incorporate expert maps into predictions, greatly improving the accuracy of estimated distributions. Another obvious approach to incorporate additional information is using priors in a Bayesian hierarchical model (e.g. Gopalaswamy et al., 2012). Predictions from one dataset can inform predictions from the second through the specification of informed prior estimates for relevant parameters. Finally, methods that employing data-weighting still rely on a joint-likelihood framework to combine data, but allow for the weight given to each dataset in fitting estimates to be adjusted (Francis, 2011; Maunder, Crone, Punt, Valero, & Semmens, 2017).

In combination, all these methods provide a much more flexible tool kit for data integration than the typical joint likelihood approach seen in most applications to date.

3.3 | Improved accounting for observational uncertainty

> We knew that what seemed to us true could be only relatively true anyway. There is no other kind of observation.
>
> Steinbeck and Ricketts (1941)

The relationship between the true state of a location and our observed data depends on multiple sources of uncertainty. These include whether sampling occurred at a site, how much effort was expended sampling the site, the rate of species detection, the false positive detection rate, and whether location error occurs in observations (Hefley, Brost, & Hooten, 2017; MacKenzie et al., 2002; Miller et al.,

2011; Royle & Link, 2006; Tyre, Tenhumberg, & Field, 2003; Yackulic et al., 2013a). A robust set of methods are available for accounting for multiple sources of uncertainty when standardized protocols are used to collect data, especially in the case where repeated observations occur at the same sites. Less attention has been given to explicitly dealing with uncertainty for nonstandardized data types. Data integration methods provide an opportunity to improve our observation models for presence only datasets. Here, we focus on observational processes often ignored in traditional models, variation in sampling effort and false positive errors due to mis-identification.

3.3.1 | Effort

> We welcomed this help, for in general work … the more hands and eyes involved, the better.
>
> Steinbeck and Ricketts (1941)

Survey effort is generally held constant or directly measured when collecting standardized datasets. Even for nonstandardized data collection, some measure of effort is often recorded or can be extracted from the collected data. For example, the eBird citizen science database of bird location records includes information regarding the length of time, distance travelled, number of observers, and whether a complete list of species observed was entered as part of each data record (Sullivan et al., 2009). In other cases, one can frequently derive locations and timing of most visits from multi-species databases based on whether at least one species' record occurs in a cell for a given date. Similarly, the number of species observed or the observation of common species is a function of overall effort for that cell (Fithian et al., 2014). In other cases, sampling effort is a function of spatial features such as distance to road or city, or whether a location is on publicly accessible land (Fithian et al., 2014).

Pacifici et al. (2017) and Stauffer, Miller, Williams, and Brown (2017) provide examples where effort is accounted for in nonstandardized datasets. The relationship between effort and detection can be exploited when estimating occurrence to make true occurrence probabilities identifiable (Stauffer et al., 2017). This is because expected probability of observing a species in a cell, given a local occurrence probability ψ, detection rate per unit effort p, and effort E is given by

$$P(Y=1|p,\psi,E)=\psi\times(1-(1-p)^E).$$

For abundance, in the case where the count is a measure of observation intensity (e.g. total number of bird calls heard) and individuals are not identified, effort can be used to correct for variation in observations. The relationship between the observed count C, true abundance N, intensity of observed signals per unit effort S, and effort E is as follows:

$$E(C)=N\times S\times E.$$

In the case where individuals can be identified and each individual has an equal and constant probability of being detected per unit of effort, N is identifiable when effort varies the observed count will take the form of the following:

$$E(C)=N\times(1-(1-p)^E),$$

where p is the probability of detecting an individual per unit effort and E is effort. These represent versions of thinned point processes and can be modelled as such (Warton et al., 2013).

While in many cases, effort is a known variable (e.g. hours of sampling), it is also possible to treat E as an unknown variable to be estimated. Estimating effort is useful when actual effort is ill defined and when there is a need to control for multiple measures of effort. E will be a function of some set of covariates X, which positively index effort: $E \sim f(X)$. Such a function should be constrained so that E is always ≥0, E is equal to 0 when all X's are 0, and the relationship between E and X is positive.

3.3.2 | Mis-identification

> They so wanted it to be a sea-serpent. Even we hoped it would be.
>
> Steinbeck and Ricketts (1941)

Another source of uncertainty, which may be more prevalent in nonstandardized data is the presence of false positive errors. These may occur due to mis-identification, as well as mis-specification of when and where the observation occurred. False positive errors can lead to significant bias when estimating distributions and dynamics (Miller et al., 2015). However, once mis-identification is accounted for, including data sources with ambiguous detections can still improve inferences about species distributions and factors that predict those distributions (Louvrier et al., 2018b; Miller et al., 2011, 2015). The covariate and correlation methods described previously relax the need to account for false positive errors in the secondary dataset for estimates to be unbiased. Instead, presence of false positive errors should reduce covariance between results for the second dataset and reduce the extent to which the second informs results for the first.

In the case of a joint-likelihood, or where the primary dataset also includes false positive errors, accounting for this source of error is important when the goal is an unbiased estimate of occurrence. Methods to account for false positives in occurrence modelling are well developed (Chambert, Miller, & Nichols, 2015; Louvrier, Chambert, Marboutin, & Gimenez, 2018a; Miller et al., 2011, 2013; Ruiz-Gutierrez, Hooten, & Grant, 2016) and all build on the simple hierarchical mixture model allowing for false positive detections first developed by Royle and Link (2006). A key feature of much of this effort has been integrating multiple data sources to better deal with mis-identification.

3.4 | Leveraging spatial dependence

Recent developments have focused on the explicit incorporation of spatial structure in SDMs (Johnson, Conn, Hooten, Ray, & Pond, 2012) and autocorrelation in the occupancy process (Royle & Dorazio, 2008). The advantages touted for spatial modelling are the ability to borrow information across nearby sites to reduce uncertainty, improved

ability to make predictions for locations without data, and accounting for dependence between nearby observations necessary to obtain valid estimates of uncertainty. Leveraging information from an additional source(s) of data extends the potential of incorporating spatial structure in SDMs. For example, it is often rare to record observations at the same locations for all data sources, making direct comparisons of the data sources difficult. However, given that nearby locations are likely to have similar densities of individuals, it is possible to leverage spatial dependence to provide an indirect comparison of data sources.

The SDM literature provides many different examples of spatial models, however, the general form is to assume $g(\psi)$ varies by location with a mean that depends on covariates such as land-use category or regional climate, and covariance between the occupancy probabilities at two locations is modelled in terms of the spatial configuration of the data locations and $g()$ refers to the appropriate link function (Johnson et al., 2012). The choice of covariance model depends on the nature of these data. If the spatial locations are areal, that is, defined only at a fixed number of spatial regions (such as counties), then modelling spatial dependence using the adjacency matrix of the regions in the study using a conditionally autoregressive model (CAR) is appropriate (Banerjee, Carlin, & Gelfand, 2014). If the spatial locations are single points in space, such as trap locations, then a geostatistical model (Banerjee et al., 2014) for the covariance as a function of the distance between points is preferred. The CAR model has the advantage of computational simplicity because dependence is defined locally which can be exploited using sparse matrix operations. On the other hand, geostatistical models are arguably more flexible and realistic, even when applied to areal data defining distance using regional centroids.

An important complication likely to arise in spatial data integration is the so-called "change of support problem" (Finley, Banerjee, & Cook, 2014; Gelfand, Diggle, Guttorp, & Fuentes, 2010; Hefley et al., 2017). That is, data sources may produce data that measure occupancy or abundance at different spatial resolutions, such as measurements of abundance at a single point location vs. a summary of an entire transect. One approach is to use the courser data as a covariate in the model for the fine-scale data, as in the covariate model (Pacifici et al., 2017). This has the advantage of simplicity but may suppress uncertainty in the courser data source. Another approach is to assume that the latent process of interest (occupancy or abundance) is shared by all data sources, and then average the latent response to the data's resolution (Gelfand et al., 2010) in the likelihood. As discussed previously in the context of the joint-likelihood model, this is ideal when both data sources are of high quality, but may be misleading if one is poor.

3.5 | Flexible integration methods—A worked example

To help illustrate concepts in this section, we include an example analysis which illustrates combining data without a joint likelihood and including spatial information and explicit modelling of effort as an unknown random variable in nonstandardized data collection. A full description of the data, statistical models, and code can be found in Supporting Information Appendix A. To summarize, we consider two datasets, both collected for black-throated blue warblers for a subset of the state of Pennsylvania, USA. The first dataset follows a formal study design with standardized fixed radius point counts conducted at random locations throughout the study area. The second dataset comprises nonstandardized data collected by eBird. Location and effort come with a large degree of uncertainty. For this dataset, we summarize counts at a larger scale, 1/24th by 1/16th degree latitude and longitude blocks along with three measures of relative effort. Combining data using a joint likelihood presents difficulties in specifying a shared parameter and instead we link data via correlation and by using the secondary dataset as a covariate. For the correlation model, we explicitly estimate effort as an unknown variable and estimate the response variable for the eBird dataset as either a Poisson count process and occurrence. In the covariate model, we scale eBird counts by assumed effort. For all models we also explicitly estimate spatial autocorrelation using a multivariate conditional autoregressive model and account for forest cover and elevation when predicting responses. Data is analysed using the freely available WinBUGS software with model likelihoods written in the standard bugs language.

4 | DESIGNING AND VALIDATING INTEGRATED SPECIES DISTRIBUTION MODELS

> 'If you believe this,' he says in effect, 'perhaps you are not right, but at least you are not a fool'.
>
> Steinbeck and Ricketts (1941)

4.1 | Model validation

Model validation is an essential step in any statistical analysis. In data integration analyses, it will often be important to validate that combining data source improves fit as compared to a single-source analysis, or to select among different data integration strategies (e.g. a joint-likelihood vs. covarying models). If both data sources are considered to be equal, measures of fit and validation based on the joint or full likelihood of both datasets may be reasonable. When one data source instead comes from a standardized data collection protocol and the others are nonstandardized, it seems reasonable to validate the model using the standardized data source to compare methods. For example, one might perform cross-validation and select the model with the smallest prediction error on the withheld observations from standardized data source. In other cases, such as when the quality of all data sources is unknown then model selection is not straightforward. If all data are treated equally, then it is likely that fit to the largest data sources will overwhelm the others, and if this data source is noisy or biased then the selection results are dubious.

Multi-objective optimization (Branke, Deb, & Miettinen, 2008) may provide a path forward. One possible solution is to combine measures of fit across the different sources of data using a weighted average, either post hoc as is done in ensemble methods or within a joint-likelihood weighting framework (Francis, 2011). The weights control our faith in each data source. On one extreme, if one data source is of high quality, it would be given weight one and all other weights would be set to zero. On the other extreme, setting all the weights to be equal puts all data sources on an equal footing. Of course, selecting the weights remains a subjective decision and so this approach does not provide a definitive solution. However, a Pareto diagram (Branke et al., 2008) that plots each model's measure of fit in a scatterplot can help visualize the results, and rule out models that are not Pareto efficient, that is, models that are outperformed by another model for all data sources.

4.2 | Optimal sampling design

> We planned to collect marine animals in a remote place on certain days and at certain hours indicated on the tide charts.
>
> Steinbeck and Ricketts (1941)

Carefully considering the sampling design, that is, the times and locations to make observations, is paramount when survey resources are limited (Legg & Nagy, 2006). A general framework for identifying the optimal design for ecological studies includes setting the objective function (e.g. classification accuracy) and the constraints (e.g. sample size) on the design (Laber et al., 2018; Reich, Pacifici, & Stallings, 2018; Taha, 2011). The most important component is having clear objectives (Boulinier, Yoccoz, Nichols, & Boulinier, 2001) as this can have a major impact on the optimal design. For example, Sanderlin et al. (2014) determined a different optimal sampling design with single-season, multi-species occupancy models when the objective was maximum accuracy of species richness or detection probability (fewer sites, more visits) than rare species occupancy probability (more sites, fewer visits), concurrent with the results of single species occupancy sampling designs (Field, Tyre, & Possingham, 2005; Mackenzie & Royle, 2005). Guillera-Arroita, Ridout, and Morgan (2010, 2014) similarly found that when detection was of interest, more effort should be used conducting repeat surveys to obtain more accurate detection estimates. Optimization of designs could be solved with linear programming or with graphical solutions (i.e. Sanderlin, Lazar, Conroy, & Reeves, 2012; Sanderlin et al., 2014). Part of the optimization framework should also include exploring both parametric uncertainty (Guillera-Arroita et al., 2010, 2014; Sanderlin et al., 2014) and decision uncertainty.

There have been many model developments combining multiple data sources. In contrast much less attention has been given to sampling design of integrated models. While we often see improved precision with joint models (Besbeas et al., 2002; Schaub & Abadi, 2010) these will come with some cost associated with collecting additional data. Collecting some additional data sources may be more efficient (improved precision per unit effort) than others (Clement, 2016; Sanderlin, Block, Strohmeyer, Saab, & Ganey, 2018). These considerations are important during initial study design or when there is an opportunity to collect additional data. By using an optimal design framework, you can explore sampling design and cost trade-offs associated with additional sources of data and how to spatially allocate samples to optimize parameter accuracy. Determining the optimal locations of additional samples is particularly important with SDMs, especially with citizen scientists. Programmes could be designed to target specific habitats or regions of interest, or areas with the most uncertainty, to meet study objectives. With rare or elusive species characterized by low occurrence or detection probabilities, additional sampling may be beneficial. Often, rare species are of most conservation concern. Multi-phase/stage designs, or when pilot data or first stage/phase of a project, can be used in an occupancy framework to optimally allocate samples with rare species (Guillera-Arroita et al., 2014; Pacifici, Dorazio, & Conroy, 2012; Pacifici, Reich, Dorazio, & Conroy, 2016).

With spatial occupancy designs, the goal is to select locations from the set of all possible locations to optimize some criteria (efficient estimation of state variable or covariate, prediction of response across design space). Although the spatial design literature is rich in examples and applications of these types of designs (Mateu & Müller, 2012), ecological data pose unique design challenges. Data are often non-Gaussian, subject to many different sources of error, and often severely correlated spatially and temporally. Typical spatial designs attempt to maximize coverage for spatial interpolation by placing sampling locations uniformly over the study domain (e.g. space-filling designs) or to improve covariance estimation by clustering sample locations. Reich et al. (2018) apply these principles of spatial optimal design to ecological data while exploring the advantages of integrating auxiliary data to inform the design of species distribution modelling. Reich et al. (2018) use a novel objective function that minimizes misclassification rate of occurrence probabilities and show that using auxiliary data refines the appropriate sampling locations and ultimately reduces uncertainty across a species' distribution. We believe this avenue of explicitly incorporating error rates in the optimality criteria is a fruitful endeavour and holds promise to improve the overall ability of mapping species distributions. It is natural to think of auxiliary data as not only informing the design, but potentially as a calibration sample if the quality is high or with genotyping error (i.e. Sanderlin et al., 2012).

5 | CONCLUSIONS

Data integration can improve species distribution estimates while properly accounting for data collected under different sampling designs. Recent developments in integrated distribution models, many of which are highlighted here, provide a framework for future work. The key point that we believe are most relevant are:

1. Point process thinking provides a common framework that facilitates the unification of a wide range of different estimators and data types.
2. When fitting integrated models, consideration should be given to the wide range of flexible methods available for integrating data and not just joint-likelihood approaches.
3. While much of the current development in data integration has focused on static models of species distributions, applications of data integration methods to estimate range dynamics and community responses to change are natural extensions and will be at the frontier of new developments.
4. Predictions can be improved by careful thought about how data is collected and utilized. Opportunities for improvement include greater incorporation of spatial information, measures of effort, and explicitly modelling sources of observation error.
5. It is important to be cognizant of the new challenges that come with integrated estimators including how to validate data, dealing with nonindependence among datasets, and design of surveys.

The rapid development of methods in recent years demonstrates the potential for data integration to provide "a picture more complete." Our hope is that this review points to a flexible and robust toolbox that is key to future developments.

ACKNOWLEDGEMENTS

We acknowledge SAMSI for organizing a session on Mathematical/Statistical Ecology, which supported this work and the EURING Technical Conference for organizing a session on the topic that stimulated this review.

AUTHORS' CONTRIBUTIONS

All authors made substantial contributions to all stages including the conception, research, writing and revision of this review. Data analysis for the worked example was completed by D.A.W.M. and K.P.

DATA ACCESSIBILITY

Black-throated blue warbler detection data were downloaded from publicly available datasets. The eBird data was accessed on 01 May 2013 from the public data portal (http://ebird.org/data/download/). Pennsylvania Breeding Bird Atlas point count data was accessed from the PA Game Commission (http://www.pabirdatlas.psu.edu/) All data used in the worked example is provided in the supplementary files, which allows the complete analysis to be replicated and to explore other formulations of the data integration model.

ORCID

David A. W. Miller http://orcid.org/0000-0002-3011-3677
Jamie S. Sanderlin http://orcid.org/0000-0001-8651-9804

REFERENCES

Abadi, F., Gimenez, O., Jakober, H., Stauber, W., Arlettaz, R. R., & Schaub, M. (2012). Estimating the strength of density dependence in the presence of observation errors using integrated population models. *Ecological Modelling*, *242*, 1–9. https://doi.org/10.1016/j.ecolmodel.2012.05.007

Ahrestani, F. S., Saracco, J. F., Sauer, J. R., Pardieck, K. L., & Royle, J. A. (2017). An integrated population model for bird monitoring in North America. *Ecological Applications*, *27*, 916–924. https://doi.org/10.1002/eap.1493

Ahumada, J. A., Silva, C. E., Gajapersad, K., Hallam, C., Hurtado, J., Martin, E., … Andelman, S. J. (2011). Community structure and diversity of tropical forest mammals: Data from a global camera trap network. *Philosophical transactions of the Royal Society of London Series B, Biological Sciences*, *366*, 2703–2711. https://doi.org/10.1098/rstb.2011.0115

Amburgey, S. M., Miller, D. A. W., Campbell Grant, E. H., Rittenhouse, T. A. G., Benard, M. F., Richardson, J. L., … Werner, E. E. (2018). Range position and climate sensitivity: The structure of among-population demographic responses to climatic variation. *Global Change Biology*, *24*, 439–454. https://doi.org/10.1111/gcb.13817

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. Boca Raton, FL: CRC Press.

Besbeas, P., Freeman, S. N., Morgan, B. J. T., & Catchpole, E. a. (2002). Integrating mark-recapture–recovery and census data to estimate animal abundance and demographic parameters. *Biometrics*, *58*, 540–547. https://doi.org/10.1111/j.0006-341X.2002.00540.x

Besbeas, P., Lebreton, J.-D., & Morgan, B. J. T. (2003). The efficient integration of abundance and demographic data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *52*, 95–102. https://doi.org/10.1111/1467-9876.00391

Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., … Frusher, S. (2014). Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, *173*, 144–154. https://doi.org/10.1016/j.biocon.2013.07.037

Blanc, L., Marboutin, E., Gatti, S., Zimmermann, F., & Gimenez, O. (2014). Improving abundance estimation by combining capture–recapture and occupancy data: Example with a large carnivore. *Journal of Applied Ecology*, *51*, 1733–1739. https://doi.org/10.1111/1365-2664.12319

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

Branke, J., Deb, K., & Miettinen, K. (2008). *Multiobjective optimization: Interactive and evolutionary approaches*. Berlin: Springer Science & Business Media.

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

Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J., Borchers, D., & Thomas, L. (2001). *Introduction to distance sampling: estimating abundance of biological parameters*. Oxford, UK: Oxford University Press.

Catchpole, E. A., Freeman, S. N., Morgan, B. J. T., & Harris, M. P. (1998). Integrated recovery/recapture data analysis. *Biometrics*, *54*, 33–46. https://doi.org/10.2307/2533993

Chambert, T., Miller, D. A. W. D., & Nichols, J. D. J. D. (2015). Modeling false positive detections in species occurrence data under different study designs. *Ecology*, *96*, 332–339. https://doi.org/10.1890/14-1507.1

Chambert, T., Waddle, J. H., Miller, D. A. W., Walls, S. C., & Nichols, J. D. (2017). A new framework for analysing automated acoustic species detection data: Occupancy estimation and optimization of recordings postprocessing. *Methods in Ecology and Evolution*, *9*, 560–570. https://doi.org/10.1650/CONDOR-17-83.1

Chandler, R. B., & Clark, J. D. (2014). Spatially explicit integrated population models. *Methods in Ecology and Evolution*, *5*, 1351–1360. https://doi.org/10.1111/2041-210X.12153

Clare, J., McKinney, S. T., Depue, J. E., & Loftin, C. S. (2017). Pairing field methods to improve inference in wildlife surveys while accommodating detection covariance. *Ecological Applications*, *27*, 2031–2047. https://doi.org/10.1002/eap.1587

Clark, J., Gelfand, A., Woodall, C., & Zhu, K. (2013). More than the sum of the parts: Forest climate response from joint species distribution models. *Ecological Applications*, *24*, 990–999.

Clement, M. J. (2016). Designing occupancy studies when false-positive detections occur. *Methods in Ecology and Evolution*, *7*, 1538–1547. https://doi.org/10.1111/2041-210X.12617

Cressie, N. (2015). *Statistics for spatial data* (2nd ed.). New York, NY: John Wiley and Sons.

Dail, D., & Madsen, L. (2011). Models for estimating abundance from repeated counts of an open metapopulation. *Biometrics*, *67*, 577–587. https://doi.org/10.1111/j.1541-0420.2010.01465.x

Davis, A. J., Hooten, M. B., Phillips, M. L., & Doherty, P. F. (2014). An integrated modeling approach to estimating Gunnison sage-grouse population dynamics: Combining index and demographic data. *Ecology and Evolution*, *4*, 4247–4257.

Davis, C. L., Miller, D. A. W. D. A., Walls, S. C. S. C., Barichivich, W. J. W. J., Riley, J. W., & Brown, M. E. M. E. (2017). Species interactions and the effects of climate variability on a wetland amphibian metacommunity. *Ecological Applications*, *27*, 285–296. https://doi.org/10.1002/eap.1442

Dicko, A. H., Lancelot, R., Seck, M. T., Guerrini, L., Sall, B., Lo, M., & Vreysen, M. J. B. (2014) Using species distribution models to optimize vector control in the framework of the tsetse eradication campaign in Senegal. *Proceedings of the National Academy of Sciences*, *111*, 10149-10154. https://doi.org/10.1073/pnas.1407773111

Diggle, P. (2013). *Spatial analysis of spatial and spatio-temporal point patterns* (3rd ed.). Boca Raton, FL: CRC Press. https://doi.org/10.1201/b15326

Dorazio, R. M. (2012). Predicting the geographic distribution of a species from presence-only data subject to detection errors. *Biometrics*, *68*, 1303–1312. https://doi.org/10.1111/j.1541-0420.2012.01779.x

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. https://doi.org/10.1111/geb.12216

Dorazio, R. M., Royle, J. A., Söderström, B., & Glimskär, A. (2006). Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology*, *87*, 842–854. https://doi.org/10.1890/0012-9658(2006)87[842:ESRAAB]2.0.CO;2

Eacker, D. R., Lukacs, P. M., Proffitt, K. M., & Hebblewhite, M. (2017). Assessing the importance of demographic parameters for population dynamics using Bayesian integrated population modeling. *Ecological Applications*, *27*, 1280–1293. https://doi.org/10.1002/eap.1521

Edwards, J. L., Lane, M. A., & Nielsen, E. S. (2000). Interoperability of biodiversity databases: Biodiversity information on every desktop. *Science*, *289*, 2312–2315. https://doi.org/10.1126/science.289.5488.2312

Efford, M., Borchers, D., & Byrom, A. E. (2009). Density estimation by spatially explicit capture–recapture: Likelihood-based methods. In E. G. Cooch, D. L. Thomson, & M. J. Conroy (Eds.), *Modeling demographic processes in marked populations* (pp. 255–269). New York, NY: Springer.

Elith, J., Graham, C. H., Anderson, R., Dudík, M., Ferrier, S., Guisan, A., … Zimmermann, N. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*, 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, *40*, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Ellner, S. P., & Rees, M. (2006). Integral projection models for species with complex demography. *The American Naturalist*, *167*, 410–428. https://doi.org/10.1086/499438

Farnsworth, G. L., Pollock, K. H., Nichols, J. D., Simons, T. R., Hines, J. E., & Sauer, J. R. (2002). A removal model for estimating detection probabilities from point-count surveys. *Auk*, *119*, 414–425. https://doi.org/10.1642/0004-8038(2002)119[0414:ARMFED]2.0.CO;2

Ficetola, G. F., Thuiller, W., & Miaud, C. (2007). Prediction and validation of the potential global distribution of a problematic alien invasive species—The American bullfrog. *Diversity and Distributions*, *13*, 476–485. https://doi.org/10.1111/j.1472-4642.2007.00377.x

Field, S. A., Tyre, A. J., & Possingham, H. P. (2005). Optimizing allocation of monitoring effort under economic and observational constraints. *Journal of Wildlife Management*, *69*, 473–482. https://doi.org/10.2193/0022-541X(2005)069[0473:OAOMEU]2.0.CO;2

Finley, A. O., Banerjee, S., & Cook, B. D. (2014). Bayesian hierarchical models for spatially misaligned data in R. *Methods in Ecology and Evolution*, *5*, 514–523. https://doi.org/10.1111/2041-210X.12189

Fithian, W., Elith, J., Hastie, T., & Keith, D. A. (2014). Bias correction in species distribution models: Pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, *6*, 424–438.

Fithian, W., & Hastie, T. (2013). Finite-sample equivalence in statistical models for presence-only data. *Annals of Applied Statistics*, *7*, 1917–1939. https://doi.org/10.1214/13-AOAS667

Fletcher, R. J., Robert, J., Greene, D. U., & Tye, C. A. (2016). Integrated models that unite local and regional data reveal larger-scale environmental relationships and improve predictions of species distributions. *Landscape Ecology*, *31*, 1369–1382. https://doi.org/10.1007/s10980-015-0327-9

Francis, R. I. C. C. (2011). Data weighting in statistical fisheries stock assessment models. *Canadian Journal of Fisheries and Aquatic Sciences*, *68*, 1124–1138. https://doi.org/10.1139/f2011-025

Freeman, S. N., & Besbeas, P. (2012). Quantifying changes in abundance without counting animals: Extensions to a method of fitting integrated population models. *Journal of Ornithology*, *152*, 409–418. https://doi.org/10.1007/s10336-011-0667-4

Gelfand, A. E., Diggle, P., Guttorp, P., & Fuentes, M. (Eds.) (2010). *Handbook of spatial statistics*. Boca Raton, FL: CRC Press.

Gopalaswamy, A. M., Royle, J. A., Delampady, M., Nichols, J. D., Karanth, K. U., & Macdonald, D. W. (2012). Density estimation in tiger populations: Combining information for strong inference. *Ecology*, *93*, 1741–1751. https://doi.org/10.1890/11-2110.1

Grinnell, J. (1917). The niche-relationshipos of the California Thrasher. *Auk*, *34*, 427–433. https://doi.org/10.2307/4072271

Guillera-Arroita, G., Lahoz-Monfort, J. J., Elith, J., Gordon, A., Kujala, H., Lentini, P. E., … 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. https://doi.org/10.1111/geb.12268

Guillera-Arroita, G., Ridout, M. S., & Morgan, B. J. T. (2010). Design of occupancy studies with imperfect detection. *Methods in Ecology and Evolution*, *1*, 131–139. https://doi.org/10.1111/j.2041-210X.2010.00017.x

Guillera-Arroita, G., Ridout, M. S., & Morgan, B. J. T. (2014). Two-stage Bayesian study design for species occupancy estimation. *Journal of Agricultural, Biological, and Environmental Statistics*, *19*, 278–291. https://doi.org/10.1007/s13253-014-0171-4

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, *8*, 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I. T., … Buckley, Y. M. (2013). Predicting species distributions for conservation decisions. *Ecology Letters*, *16*, 1424–1435. https://doi.org/10.1111/ele.12189

Hanks, E. M., Hooten, M. B., & Baker, F. A. (2011). Reconciling multiple data sources to improve accuracy of large-scale prediction of forest

disease incidence. *Ecological Applications*, *21*, 1173–1188. https://doi.org/10.1890/09-1549.1

Hastie, T., & Fithian, W. (2013). Inference from presence-only data; the ongoing controversy. *Ecography*, *36*, 864–867. https://doi.org/10.1111/j.1600-0587.2013.00321.x

Hefley, T. J., Brost, B. M., & Hooten, M. B. (2017). Bias correction of bounded location errors in presence-only data. *Methods in Ecology and Evolution*, *8*, 1566–1573. https://doi.org/10.1111/2041-210X.12793

Hefley, T. J., & Hooten, M. B. (2015). On the existence of maximum likelihood estimates for presence-only data (ed D Warton). *Methods in Ecology and Evolution*, *6*, 648–655. https://doi.org/10.1111/2041-210X.12340

Hefley, T. J., & Hooten, M. B. (2016). Hierarchical species distribution models. *Current Landscape Ecology Reports*, *1*, 87–97. https://doi.org/10.1007/s40823-016-0008-7

Hochachka, W. M., Fink, D., Hutchinson, R. A., Sheldon, D., Wong, W.-K., & Kelling, S. (2012). Data-intensive science applied to broad-scale citizen science. *Trends in Ecology and Evolution*, *27*, 130–137. https://doi.org/10.1016/j.tree.2011.11.006

Hostetler, J. A. (2015). Improved state-space models for inference about spatial and temporal variation in abundance from count data. *Ecology*, *96*, 1713–1723. https://doi.org/10.1890/14-1487.1

Iknayan, K. J., Tingley, M. W., Furnas, B. J., & Beissinger, S. R. (2013). Detecting diversity: Emerging methods to estimate species diversity. *Trends in Ecology & Evolution*, *29*, 97–106.

Johnson, D. S., Conn, P. B., Hooten, M. B., Ray, J. C., & Pond, B. A. (2012). Spatial occupancy models for large data sets. *Ecology*, *94*, 801–808.

Jolly, G. (1965). Explicit estimates from capture–recapture data with both death and immigration-stochastic model. *Biometrika*, *52*, 225–247. https://doi.org/10.1093/biomet/52.1-2.225

Kendall, W. L., Pollock, K. H., & Brownie, C. (1995). A likelihood-based approach to capture–recapture estimation of demographic parameters under the robust design. *Biometrics*, *51*, 293–308. https://doi.org/10.2307/2533335

King, R., Brooks, S. P., & Coulson, T. (2008). Analyzing complex capture–recapture data in the presence of individual and temporal covariates and model uncertainty. *Biometrics*, *64*, 1187–1195. https://doi.org/10.1111/j.1541-0420.2008.00991.x

Knapp, R. A. R. A., Fellers, G. M. G. M., Kleeman, P. M. P. M., Miller, D. A. W. D. A. W., Vredenburg, V. T. V. T., Bree, E., … Briggs, C. J. C. J. (2016). Large-scale recovery of an endangered amphibian despite ongoing exposure to multiple stressors. *Proceeding of the National Academy of Sciences*, *113*, 11889–11894. https://doi.org/10.1073/pnas.1600983113

Koshkina, V., Wang, Y., Gordon, A., Dorazio, R. M., & White, M. (2017). Integrated species distribution models: Combining presence-background data and site-occupancy data with imperfect detection. *Methods in Ecology and Evolution*, *8*, 420–430. https://doi.org/10.1111/2041-210X.12738

Laber, E., Meyer, N. J., Reich, B. J., Pacifici, K., Collazo, J. A., & Drake, J. (2018). Optimal treatment allocations in space and time for online control of an emerging infectious disease. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *67*, 743–789. https://doi.org/10.1111/rssc.12266

Lapointe, S., Giroux, J.-F., Bélanger, L., & Filion, B. (2000). Benefits of rotational grazing and dense nesting cover for island-nesting waterfowl in southern Quebec. *Agriculture, Ecosystems & Environment*, *78*, 261–272. https://doi.org/10.1016/S0167-8809(99)00132-2

Legg, C. J., & Nagy, L. (2006). Why most conservation monitoring is, but need not be, a waste of time. *Journal of Environmental Management*, *78*, 194–199. https://doi.org/10.1016/j.jenvman.2005.04.016

Lele, S. R., & Keim, J. L. (2006). Weighted distributions and estimation of resource selection probability functions. *Ecology*, *87*, 3021–3028.

Lele, S. R., Moreno, M., & Bayne, E. (2012). Dealing with detection error in site occupancy surveys: What can we do with a single survey? *Journal of Plant Ecology*, *5*, 22–31. https://doi.org/10.1093/jpe/rtr042

Louvrier, J., Chambert, T., Marboutin, E., & Gimenez, O. (2018a). Accounting for misidentification and heterogeneity in occupancy studies using hidden Markov models. *Ecological Modelling*, *387*, 61–69. https://doi.org/10.1016/j.ecolmodel.2018.09.002

Louvrier, J., Molinari-jobin, A., Marc, K., Chambert, T., Miller, D., Zimmermann, F., … Gimenez, O. (2018b). Use of ambiguous detections to improve estimates from species distribution models. *Conservation Biology*. https://doi.org/10.1111/cobi.13191

Mackenzie, D. I., Bailey, L. L., & Nichols, J. D. (2004). Investigating species co-occurrence patterns when species. *Journal of Animal Ecology*, *73*, 546–555. https://doi.org/10.1111/j.0021-8790.2004.00828.x

MacKenzie, D., Nichols, J., & Hines, J. (2003). Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology*, *84*, 2200–2207. https://doi.org/10.1890/02-3090

MacKenzie, D., Nichols, J., Lachman, G., Droege, S., Royle, J. A., & Langtimm, C. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, *83*, 2248–2255. https://doi.org/10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. L., & Hines, J. E. (2006). *Occupancy estimation and modeling: Inferring patterns and dynamics of species occurrence*. New York, NY: Academic Press.

Mackenzie, D. I., & Royle, J. A. (2005). Designing occupancy studies: General advice and allocating survey effort. *Journal of Applied Ecology*, *42*, 1105–1114. https://doi.org/10.1111/j.1365-2664.2005.01098.x

Mateu, J., & Müller, W. G. (2012). *Spatio-temporal design: Advances in efficient data acquisition*. West Sussex, UK: John Wiley and Sons.

Maunder, M., Crone, P. R., Punt, A. E., Valero, J. L., & Semmens, B. X. (2017). Data conflict and weighting, likelihood functions and process error. *Fihseries Research*, *192*, 1–4.

Merow, C., Wilson, A. M., & Jetz, W. (2016). Integrating occurrence data and expert maps for improved species range predictions. *Global Ecology and Biogeography*, *26*, 243–258.

Miller, D. A. W., Bailey, L. L., Campbell Grant, E. H., McClintock, B. T., Weir, L. A., & Simons, T. R. (2015). Performance of species occurrence estimators when basic assumptions are not met: A test using field data where true occupancy status is known. *Methods in Ecology and Evolution*, *6*, 557–565. https://doi.org/10.1111/2041-210X.12342

Miller, D. A. W., Brehme, C. S., Hines, J. E., Nichols, J. D., & Fisher, R. N. (2012). Joint estimation of habitat dynamics and species interactions: Disturbance reduces co-occurrence of non-native predators with an endangered toad. *Journal of Animal Ecology*, *81*, 1288–1297. https://doi.org/10.1111/j.1365-2656.2012.02001.x

Miller, D. A. W., Grant, E. H. C., Muths, E., Amburgey, S. M., Adams, M. J., Joseph, M. B., … Sigafus, B. H. (2018). Quantifying climate sensitivity and climate driven change in North American amphibian communities. *Nature Communications*, *9*, 3926. https://doi.org/10.1038/s41467-018-06157-6

Miller, D. A. W., Nichols, J. D., Gude, J. A., Rich, L. N., Podruzny, K. M., Hines, J. E., & Mitchell, M. S. (2013). Determining occurrence dynamics when false positives occur: Estimating the range dynamics of wolves from public survey data. *PLoS ONE*, *8*, e65808. https://doi.org/10.1371/journal.pone.0065808

Miller, D. A., Nichols, J. D., McClintock, B. T., Grant, E. H. C., Bailey, L. L., & Weir, L. A. (2011). Improving occupancy estimation when two types of observational error occur: Non-detection and species misidentification. *Ecology*, *92*, 1422–1428. https://doi.org/10.1890/10-1396.1

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. https://doi.org/10.1177/0309133309355630

Nichols, J., Bailey, L., O'Connell, A. F., Talancy, N. W., Grant, E. H. C., Gilbert, A. T., … Hines, J. E. (2008). Multi-scale occupancy estimation and modelling using multiple detection methods. *Journal of Applied Ecology*, *45*, 1321–1329. https://doi.org/10.1111/j.1365-2664.2008.01509.x

Nichols, J. D., Hines, J. E., Sauer, J. R., Fallon, J. E., & Heglund, P. J. (2000). A double-observer approach for estimating detection probability and abundance from point counts. *The Auk*, *117*, 393–408. https://doi.org/10.1642/0004-8038(2000)117[0393:ADOAFE]2.0.CO;2

Otis, D., Burnham, K. P., White, G. C., & Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs*, *62*, 3–135.

Ovaskainen, O., Roy, D. B., Fox, R., & Anderson, B. J. (2016). Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. *Methods in Ecology and Evolution*, *7*, 428–436. https://doi.org/10.1111/2041-210X.12502

Pacifici, K., Dorazio, R. M., & Conroy, M. J. (2012). A two-phase sampling design for increasing detections of rare species in occupancy surveys. *Methods in Ecology and Evolution*, *3*, 721–730. https://doi.org/10.1111/j.2041-210X.2012.00201.x

Pacifici, K., Reich, B. J., Dorazio, R. M., & 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. P., Reich, B. J., Miller, D. A. W. D., Gardner, B., Stauffer, G., Singh, S., … Collazo, J. A. (2017). Integrating multiple data sources in species distribution modeling: A framework for data fusion. *Ecology*, *98*, 840–850. https://doi.org/10.1002/ecy.1710

Pearson, R. G., & Dawson, T. P. (2003). Predicting the impacts of climate change on the distribution of species: Are bioclimate envelope models useful? *Global Ecology and Biogeography*, *12*, 361–371. https://doi.org/10.1046/j.1466-822X.2003.00042.x

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*, 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Phillips, S. J., Dudík, M., & Schapire, R. E. (2004). *A maximum entropy approach to species distribution modeling* In *Proceedings of the 21st International Conference on Machine Learning* (pp. 655–662). New York, NY: ACM Press.

Phillips, S. J., & Elith, J. (2013). On estimating probability of presence from use-availability or presence-background data. *Ecology*, *94*, 1409–1419. https://doi.org/10.1890/12-1520.1

Pollock, K. H. (1982). A capture–recapture design robust to unequal probability of capture. *Journal of Wildlife Management*, *46*, 752–757. https://doi.org/10.2307/3808568

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., … McCarthy, M. a. (2014). Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM) (ed J McPherson). *Methods in Ecology and Evolution*, *5*, 397–406. https://doi.org/10.1111/2041-210X.12180

Powell, L. A., Conroy, M. J., Hines, J. E., Nichols, J. D., & Krementz, D. G. (2000). Simultaneous use of mark-recapture and radiotelemetry to estimate survival, movement, and capture rates. *Journal of Wildlife Management*, *64*, 302–313. https://doi.org/10.2307/3803003

Royle, J. A., & Nichols, J. D. (2003). Estimating abundance from repeated presence-absence data or point counts. *Ecology*, *84*, 777–790.

Reich, B. J., Pacifici, K., & Stallings, J. W. (2018). Integrating auxiliary data in optimal spatial design for species distribution modeling. *Methods in Ecology and Evolution*, *9*, 1626–1637. https://doi.org/10.1111/2041-210X.13002

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., … Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, *6*, 366–379. https://doi.org/10.1111/2041-210X.12352

Renner, I. W., & Warton, D. I. (2013). Equivalence of MAXENT and poisson point process models for species distribution modeling in ecology. *Biometrics*, *69*, 274–281. https://doi.org/10.1111/j.1541-0420.2012.01824.x

Rossman, S., Yackulic, C., Saunders, S., Reid, J., Davis, R., & Zipkin, E. (2016). Dynamic N-occupancy models: Estimating demographic rates and local abundance from detection-nondetection data. *Ecology*, *97*, 3300–3307. https://doi.org/10.1002/ecy.1598

Rota, C. T., Wikle, C. K., Kays, R., Forrester, T. D., McShea, W. J., Parsons, A. W., & Millspaugh, J. J. (2016). A two-species occupancy model accommodating simultaneous spatial and interspecific dependence. *Ecology*, *97*, 48–53. https://doi.org/10.1890/15-1193.1

Royle, J. A. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, *60*, 108–115. https://doi.org/10.1111/j.0006-341X.2004.00142.x

Royle, J. A., Chandler, R. B., Sollman, R., & Gardner, B. (2014). *Spatial capture–recapture*. Waltham, MA: Academic Press.

Royle, J. A., Chandler, R. B., Sun, C. C., & Fuller, A. K. (2013). Integrating resource selection information with spatial capture–recapture. *Methods in Ecology and Evolution*, *4*, 520–530. https://doi.org/10.1111/2041-210X.12039

Royle, J. A., Chandler, R. B., Yackulic, C., & Nichols, J. D. (2012). Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution*, *3*, 545–554. https://doi.org/10.1111/j.2041-210X.2011.00182.x

Royle, J. A., & Dorazio, R. M. (2008). *Hierarchical modeling and inference in ecology: The analysis of data from populations, metapopulations and communities*. New York, NY: Academic Press.

Royle, J. A., & Link, W. A. (2006). Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, *87*, 835–841. https://doi.org/10.1890/0012-9658(2006)87[835:GSOMAF]2.0.CO;2

Ruiz-Gutierrez, V., Hooten, M. B., & Grant, E. H. C. (2016). Uncertainty in biological monitoring: A framework for data collection and analysis to account for multiple sources of sampling bias. *Methods in Ecology and Evolution*, *7*, 900–909. https://doi.org/10.1111/2041-210X.12542

Sanderlin, J. S., Block, W. M., Ganey, J. L., Mountain, R., Service, U. S. F., & Knoll, S. P. (2014). Optimizing study design for multi-species avian monitoring programmes. *Journal of Applied Ecology*, *51*, 860–870. https://doi.org/10.1111/1365-2664.12252

Sanderlin, J. S., Block, W. M., Strohmeyer, B. E., Saab, V. A., & Ganey, J. L. (2018). Precision gain versus effort with joint models using detection/non-detection and banding data. *Ecology and Evolution*. https://doi.org/10.1002/ece3.4825

Sanderlin, J. S., Lazar, N., Conroy, M. J., & Reeves, J. (2012). Cost-efficient selection of a marker panel in genetic studies. *Journal of Wildlife Management*, *76*, 88–94. https://doi.org/10.1002/jwmg.241

Sauer, J. R., Pardieck, K. L., Ziolkowski, D. J., Smith, A. C., Hudson, M.-A. R., Rodriguez, V., … Link, W. A. (2017). The first 50 years of the North American Breeding Bird Survey. *The Condor*, *119*, 576–593.

Schaub, M., & Abadi, F. (2010). Integrated population models: A novel analysis framework for deeper insights into population dynamics. *Journal of Ornithology*, *152*, 1–11.

Schaub, M., Gimenez, O., Sierro, A., & Arlettaz, R. (2007). Use of integrated modeling to enhance estimates of population dynamics obtained from limited data. *Conservation Biology: The Journal of the Society for Conservation Biology*, *21*, 945–955. https://doi.org/10.1111/j.1523-1739.2007.00743.x

Schwarz, C. J., & Arnason, A. (1996). A general methodology for the analysis of capture–recapture experiments in open populations. *Biometrics*, *52*, 860–873. https://doi.org/10.2307/2533048

Seber, G. (1965). A note on the multiple-recapture census. *Biometrika*, *52*, 249–259. https://doi.org/10.1093/biomet/52.1-2.249

Smith, W. B. (2002). Forest inventory and analysis: A national inventory and monitoring program. *Environmental Pollution*, *116*, S233–S242. https://doi.org/10.1016/S0269-7491(01)00255-X

Sollmann, R., Gardner, B., Parsons, A. W., Stocking, J. J., McClintock, B. T., Simons, T. R., … O'Connell, A. F. (2013a). A spatial mark-resight model augmented with telemetry data. *Ecology*, *94*, 553–559. https://doi.org/10.1890/12-1256.1

Sollmann, R., Tôrres, N. M., Furtado, M. M., De Almeida Jácomo, A. T., Palomares, F., Roques, S., & Silveira, L. (2013b). Combining camera-trapping ªnd noninvasive genetic data in a spatial

capture–recapture framework improves density estimates for the Jaguar. *Biological Conservation*, *167*, 242–247. https://doi.org/10.1016/j.biocon.2013.08.003

Stauffer, G., Miller, D. A. W., Williams, L., & Brown, J. (2017). Ruffed grouse population declines after introduction of West Nile Virus. *Journal of Wildlife Management*, *82*, 165–172.

Steinbeck, J., & Ricketts, E. F. (1941). *Sea of Cortez*. New York, NY: Viking Press.

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, *142*, 2282–2292. https://doi.org/10.1016/j.biocon.2009.05.006

Taha, H. (2011). *Operations research an introduction* (9th ed.). Upper Saddle River, NJ: Prentice Hall.

Tenan, S., Pedrini, P., Bragalanti, N., Groff, C., & Sutherland, C. (2017). Data integration for inference about spatial processes: A model-based approach to test and account for data inconsistency. *PLoS ONE*, *12*, 1–18.

Thomas, C. D., Cameron, A., Green, R. E., Bakkenes, M., Beaumont, L. J., Collingham, Y. C., … Williams, S. E. (2004). Extinction risk from climate change. *Nature*, *427*, 145–148.

Tingley, M. W., & Beissinger, S. R. (2009). Detecting range shifts from historical species occurrences: New perspectives on old data. *Trends in Ecology and Evolution*, *24*, 625–633. https://doi.org/10.1016/j.tree.2009.05.009

Tyre, A., Tenhumberg, B., & Field, S. (2003). Improving precision and reducing bias in biological surveys: Estimating false-negative error rates. *Ecological Applications*, *13*, 1790–1801. https://doi.org/10.1890/02-5078

Ward, G., Hastie, T., Barry, S., Elith, J., & Leathwick, J. R. (2009). Presence-only data and the em algorithm. *Biometrics*, *65*, 554–563. https://doi.org/10.1111/j.1541-0420.2008.01116.x

Warton, D. I., Blanchet, F. G., O'Hara, R., Ovaskainen, O., Taskinen, S., Walker, S. C., & Hui, F. (2015). So many variables: Joint modeling in community ecology. *Trends in Ecology & Evolution*, *30*, 766–779. https://doi.org/10.1016/j.tree.2015.09.007

Warton, D. I., Renner, I. W., & Ramp, D. (2013). Model-based control of observer bias for the analysis of presence-only data in ecology. *PLoS ONE*, *8*, e79168.

Warton, D. I., & Shepherd, L. C. (2010). Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *Annals of Applied Statistics*, *4*, 1383–1402. https://doi.org/10.1214/10-AOAS331

Wilson, S., Gil-Weir, K. C., Clark, R. G., Robertson, G. J., & Bidwell, M. T. (2016). Integrated population modeling to assess demographic variation and contributions to population growth for endangered whooping cranes. *Biological Conservation*, *197*, 1–7. https://doi.org/10.1016/j.biocon.2016.02.022

Yackulic, C. B., Chandler, R., Zipkin, E. F., Royle, J. A., Nichols, J. D., Campbell Grant, E. H., & Veran, S. (2013a). Presence-only modelling using MAXENT: When can we trust the inferences? (ed. R.B. O'Hara). *Methods in Ecology and Evolution*, *4*, 236–243. https://doi.org/10.1111/2041-210x.12004

Yackulic, C., Reid, J., Nichols, J., Hines, J., Davis, R., & Forsman, E. (2013b). The roles of competition and habitat in the dynamics of populations and species distributions. *Ecology*, *95*, 265–279.

Zipkin, E. F., Rossman, S., Yackulic, C. B., Wiens, J. D., Thorson, J. T., Davis, R. J., & Grant, E. H. C. (2017). Integrating count and detection–nondetection data to model population dynamics. *Ecology*, *98*, 1640–1650. https://doi.org/10.1002/ecy.1831

Zipkin, E. F., & Saunders, S. P. (2018). Synthesizing multiple data types for biological conservation using integrated population models. *Biological Conservation*, *217*, 240–250. https://doi.org/10.1016/j.biocon.2017.10.017

Zipkin, E. F., Thorson, J. T., See, K., Lynch, H. J., Grant, E. H. C., Kanno, Y., … Royle, J. A. (2014). Modeling structured population dynamics using data from unmarked individuals. *Ecology*, *95*, 22–29. https://doi.org/10.1890/13-1131.1

SUPPORTING INFORMATION

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

How to cite this article: Miller DAW, Pacifici K, Sanderlin JS, Reich BJ. The recent past and promising future for data integration methods to estimate species' distributions. *Methods Ecol Evol*. 2019;10:22–37. https://doi.org/10.1111/2041-210X.13110