Special Feature: Data Integration for Population Models

Ecology, 100(6), 2019, e02710

A practical guide for combining data to model species distributions

Robert J. Fletcher, Jr.,[1,5] Trevor J. Hefley,[2] Ellen P. Robertson,[1] Benjamin Zuckerberg,[3]
Robert A. McCleery,[1] and Robert M. Dorazio[4]

[1] *Department of Wildlife Ecology and Conservation, University of Florida, P.O. Box 110430, 110 Newins-Ziegler Hall, Gainesville, Florida 32611-0430 USA*
[2] *Department of Statistics, Kansas State University, 205 Dickens Hall, Manhattan, Kansas 66506-0802 USA*
[3] *Department of Forest and Wildlife Ecology, University of Wisconsin, 226 Russell Labs, 1630 Linden Drive, Madison, Wisconsin 53706-1598 USA*
[4] *Department of Biology, San Francisco State University, 1600 Holloway Avenue, San Francisco, California 94132 USA*

Citation: Fletcher, Jr., R. J., T. J. Hefley, E. P. Robertson, B. Zuckerberg, R. A. McCleery, and R. M. Dorazio. 2019. A practical guide for combining data to model species distributions. Ecology 100(6):e02710. 10.1002/ecy.2710

Abstract. Understanding and accurately modeling species distributions lies at the heart of many problems in ecology, evolution, and conservation. Multiple sources of data are increasingly available for modeling species distributions, such as data from citizen science programs, atlases, museums, and planned surveys. Yet reliably combining data sources can be challenging because data sources can vary considerably in their design, gradients covered, and potential sampling biases. We review, synthesize, and illustrate recent developments in combining multiple sources of data for species distribution modeling. We identify five ways in which multiple sources of data are typically combined for modeling species distributions. These approaches vary in their ability to accommodate sampling design, bias, and uncertainty when quantifying environmental relationships in species distribution models. Many of the challenges for combining data are solved through the prudent use of integrated species distribution models: models that simultaneously combine different data sources on species locations to quantify environmental relationships for explaining species distribution. We illustrate these approaches using planned survey data on 24 species of birds coupled with opportunistically collected eBird data in the southeastern United States. This example illustrates some of the benefits of data integration, such as increased precision in environmental relationships, greater predictive accuracy, and accounting for sample bias. Yet it also illustrates challenges of combining data sources with vastly different sampling methodologies and amounts of data. We provide one solution to this challenge through the use of weighted joint likelihoods. Weighted joint likelihoods provide a means to emphasize data sources based on different criteria (e.g., sample size), and we find that weighting improves predictions for all species considered. We conclude by providing practical guidance on combining multiple sources of data for modeling species distributions.

Key words: citizen science; data fusion; ecological niche model; habitat suitability model; integrated model; spatial point process; Special Feature: Data Integration for Population Models; species distribution model.

Introduction

Understanding species distributions is essential to ecology, evolution, and conservation biology. In this context, species distribution models (SDMs) are often used to understand environmental relationships and predict species distributions in both environmental and geographic space (Elith and Leathwick 2009). These models have been used to project potential effects of climate change (Case and Lawler 2017), identify areas with high risk of invasion by exotic species (Palaoro et al. 2013) or land clearing for energy production (Evans et al. 2010), improve biological inventories (Raxworthy et al. 2003), understand niche conservatism (Wiens et al. 2010), and guide a variety of conservation decisions (Guisan et al. 2013).

The usefulness of species distribution models can nonetheless be limited by the data used in model building, which often contain relatively limited information and several sources of bias (Norris 2004, McCarthy et al. 2012). Common sources of species data include presence–only, presence-absence, and abundance data, which can arise from either planned surveys or opportunistic sampling. Each of these data sources has potential strengths but each can also have observation errors (e.g., presence–absence data may be better described as “detection–nondetection” data in some cases; MacKenzie et al.

Manuscript received 14 June 2018; revised 9 November 2018; accepted 2 January 2019. Corresponding Editor: Brian D. Inouye.

Editors’ Note: Papers in this Special Feature are linked online in a virtual table of contents at: www.wiley.com/go/ecologyjournal

[5] E-mail: robert.fletcher@ufl.edu

2002), which we address below. Presence-only data, which often come from opportunistic sampling based on museum specimens, biological inventories or some citizen science programs, are frequently used for modeling species distributions (Graham et al. 2004, Dickinson et al. 2010). Such data are useful because they typically capture a large spatial extent, are readily available, and sometimes include a large amount of data; however, they often suffer from three key issues. First, sample selection bias is common (Phillips et al. 2009, Fourcade et al. 2014) where opportunistic sampling occurs in areas that are easily accessible (e.g., near roads). Second, imperfect detection in areas that are visited can arise, and there is often limited information to account for this issue (Hefley et al. 2013). Third, because absences are not available, only measures of relative suitability/probability can typically be modeled (but see Dorazio 2012, Royle et al. 2012). In contrast to presence-only data, presence–absence and abundance data are often the target of more rigorous planned surveys. Planned surveys may be less affected by sample selection bias and are frequently designed to estimate detection probabilities (Rota et al. 2011, Lawson et al. 2014), such that absolute probabilities of occurrence or estimates of abundance are often possible. However, data from planned surveys frequently suffer from small amounts of data and limited geographic extent in comparison to opportunistic, presence-only data.

To improve our understanding of species distributions, it is tempting to combine multiple data sources when modeling species distributions. Indeed, combining multiple data sources for modeling species distributions has increased substantially in recent years (Fig. 1). In these situations, data are combined for a variety of reasons. For instance, the amount of data on species locations may be limited in planned surveys, particularly for rare species, such that data sources are sometimes pooled to increase sample size (Fletcher et al. 2016). Increasingly available “big data” sets in ecology may provide opportunities in this way (e.g., La Sorte et al. 2018). In addition, some data sources may have biases that can be alleviated by combining a second data source that does not contain the same limitations (Dorazio 2014). Despite this increasing interest, appropriately combining multiple data sources can be challenging because each data source typically has different assumptions and biases, and each may provide different information on species distribution. Recent advances in species distribution modeling have focused on how to better integrate these different sources of data to make more reliable predictions and statistical inference (Dorazio 2014, Fithian et al. 2015, Talluto et al. 2016, Pacifici et al. 2017).

Here, we review and synthesize approaches for combining multiple sources of data to identify environmental relationships and predict species distributions. We then illustrate the implementation of combining data using an example that couples planned survey data on 24 bird species with eBird data collected by citizen scientists (Sullivan et al. 2014) in the southeastern United States. This example highlights ways in which integrated species distribution models may be useful, but also emphasizes some potential challenges to combining data. We conclude by providing practical guidance for combining data in species distribution modeling.

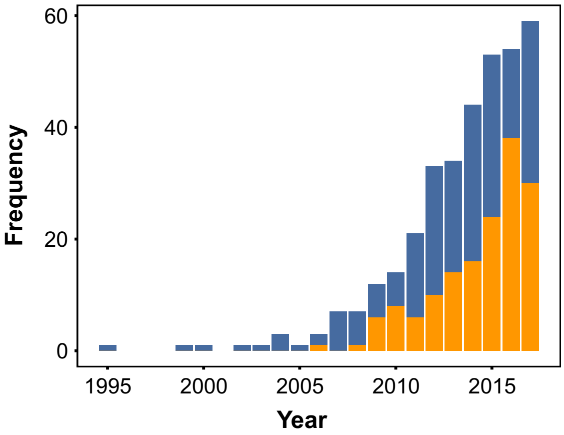

Fig. 1. Combining data is an increasingly common topic in species distribution modeling. Shown is the number of articles that were captured using the search phrase for combining and/or integrating data (blue; see text) and the number of articles that actually combined spatial data on species distribution (presence-only, presence–absence, abundance data; orange).

A diversity of approaches to combining data in species distribution models

We reviewed articles to determine how multiple types of data have been combined within species distribution models. We focused on models that use disparate sources of information collected at different localities across a region of interest that are then combined to model species distributions. Other approaches have combined data at the same locations to estimate occupancy and abundance better by using different sources of information on detectability (e.g., Miller et al. 2011, Hefley et al. 2015, Zipkin et al. 2017). These approaches have either not focused on mapping species distributions or do not link different types of distribution data collected across space, so we do not focus on them here. However, much of what we discuss is relevant to these and other models that combine data sources.

Methods.—We compiled articles within the ISI Web of Science on January 19, 2018 by using a search phrase that captures several terms for combining data: (“informative prior*” or “pool* data” or “data pool*” or “integrat* data” or “data integrat*” or “combin* data” or “data combin*” or “multiple data” or “data fusion” or “integrated” or “integrates”) and (“species” or “animal” or “plant”) and (“distribution model*” or “occupancy model*” or “abundance model*” or “integrated data model*” or “hierarchical Bayes*”). This search resulted in 353 articles. We consider these articles to be a representative sample of data integration within species distribution modeling, but acknowledge that this search likely did not capture all articles on this topic.

We retained all articles within this search that used 2+ data sources on species distribution for the response variable within a species distribution model ($N = 155$). Species distribution data included presence-only, presence–absence, abundance, and density data, with and without information on observation errors (e.g., imperfect detection and/or misidentification). For example, we use the term "presence–absence data" as a general term to reflect situations where observations errors may or may not have occurred.

With these articles, we then compiled information on how data were combined to model species distributions. Information included whether data sources were treated differently in modeling or were simply pooled, the steps of modeling (e.g., sequential or simultaneous modeling of data sources), whether data sources were considered as response variables only or response and predictor variables, whether sample design and/or potential bias of each data source was considered, and the overarching modeling frameworks used (e.g., hierarchical modeling).

Results.—From the sample of articles, there were five common approaches to combining data (Fig. 2). First, the most common approach was simply to pool data without regard to data source and/or sampling issues (Palaoro et al. 2013), or through post hoc criteria for pooling data (Underwood et al. 2010). This approach accounted for 73% of all articles that combined data. Second, separate, independent models were sometimes built from different data sources, and then predictions were combined or compared (e.g., ensembles; 11% of articles; Douma et al. 2012, Case and Lawler 2017). Third, in some cases (8% of articles), models were developed for one data source with a second auxiliary data source (e.g., ancillary information about the species that was used as a predictor for modeling the focal data set) being included in model building, commonly through the use of a covariate or offset (Merow et al. 2016, Regos et al. 2016). Fourth, in a Bayesian context, one data source may be used to derive informed priors for modeling a second data source (1% of articles; Marcantonio et al. 2016, Talluto et al. 2016). Finally, formal data integration was implemented (9% of articles), where explicit data models were developed for each data source and combined to estimate model parameters through the use of joint likelihoods (i.e., the product of individual likelihoods for each data source). We describe these models as "integrated species distribution models" (ISDMs) (Fletcher et al. 2016, Koshkina et al. 2017).

Contrasting approaches for combining data

Approaches for combining data vary in their ability to account for sampling issues, to connect different response variables between data sources, and to account for uncertainty (Table 1). To highlight these differences, we first describe inhomogeneous point process (IPP) models for modeling species distributions (Warton and Shepherd 2010). Several algorithms used for modeling species distributions (e.g., logistic regression, Maxent, boosted regression trees) can be viewed as estimating, with varying degrees of accuracy, parameters of an IPP (Warton and Shepherd 2010, Aarts et al. 2012, Renner and Warton 2013, Renner et al. 2015). In addition, different sources of data, including presence-only, presence–absence, and abundance data, can be formally connected in the context of IPP models (Miller et al. 2019). These connections facilitate understanding the benefits and limitations of different ways to combine multiple sources of data.

The IPP species distribution framework.—As a data-generating mechanism, a realization (i.e., random draw) from an IPP generates random points in geographic space. The IPP is a natural choice to model a species distribution using idealized data (e.g., exact coordinates of individuals, perfect detection, known sampling area) and it provides a well-grounded foundation for extensions to less-ideal data. The probability density function of a realization from an IPP is

$$f(n,\mathbf{s}_1,\ldots,\mathbf{s}_n)=\frac{e^{-\int_S \lambda(\mathbf{s})d\mathbf{s}}\left(\int_S \lambda(\mathbf{s})d\mathbf{s}\right)^n}{n!}\,n!\prod_{i=1}^{n}\frac{\lambda(\mathbf{s}_i)}{\int_S \lambda(\mathbf{s})d\mathbf{s}}, \tag{1}$$

which simplifies to $e^{-\int_S \lambda(\mathbf{s})d\mathbf{s}}\prod_{i=1}^{n}\lambda(\mathbf{s}_i)$, where $\mathbf{s}_i$ is a vector that contains the coordinates of the i^{th} individual location. The spatially varying intensity function $\lambda(\mathbf{s})$ controls the expected number and location of the random points within the study area S. An important quantity in Eq. 1 is the integrated intensity function, $\int \lambda(\mathbf{s})d\mathbf{s}$, which can be used to show the connection between exact point locations (presence-only data) and grid-based locations (presence–absence and abundance data).

Regression-type models (e.g., Poisson regression) can then be constructed by assuming (1) a probability distribution for the response variable and (2) a deterministic relationship between predictor variables (covariates) and the expected value of the response variable. The IPP can be formulated as a regression model by assuming the intensity function depends on location-specific covariates $\mathbf{x}(\mathbf{s})$ (i.e., the covariates at location $\mathbf{s}$) as

$$\log(\lambda(\mathbf{s}))=\beta_0+\mathbf{x}(\mathbf{s})'\boldsymbol{\beta}, \tag{2}$$

where β_0 is the intercept and $\boldsymbol{\beta}\equiv(\beta_1,\ldots,\beta_p)'$ is a vector of regression coefficients associated with the p covariates. We note that these covariates can reflect conditions

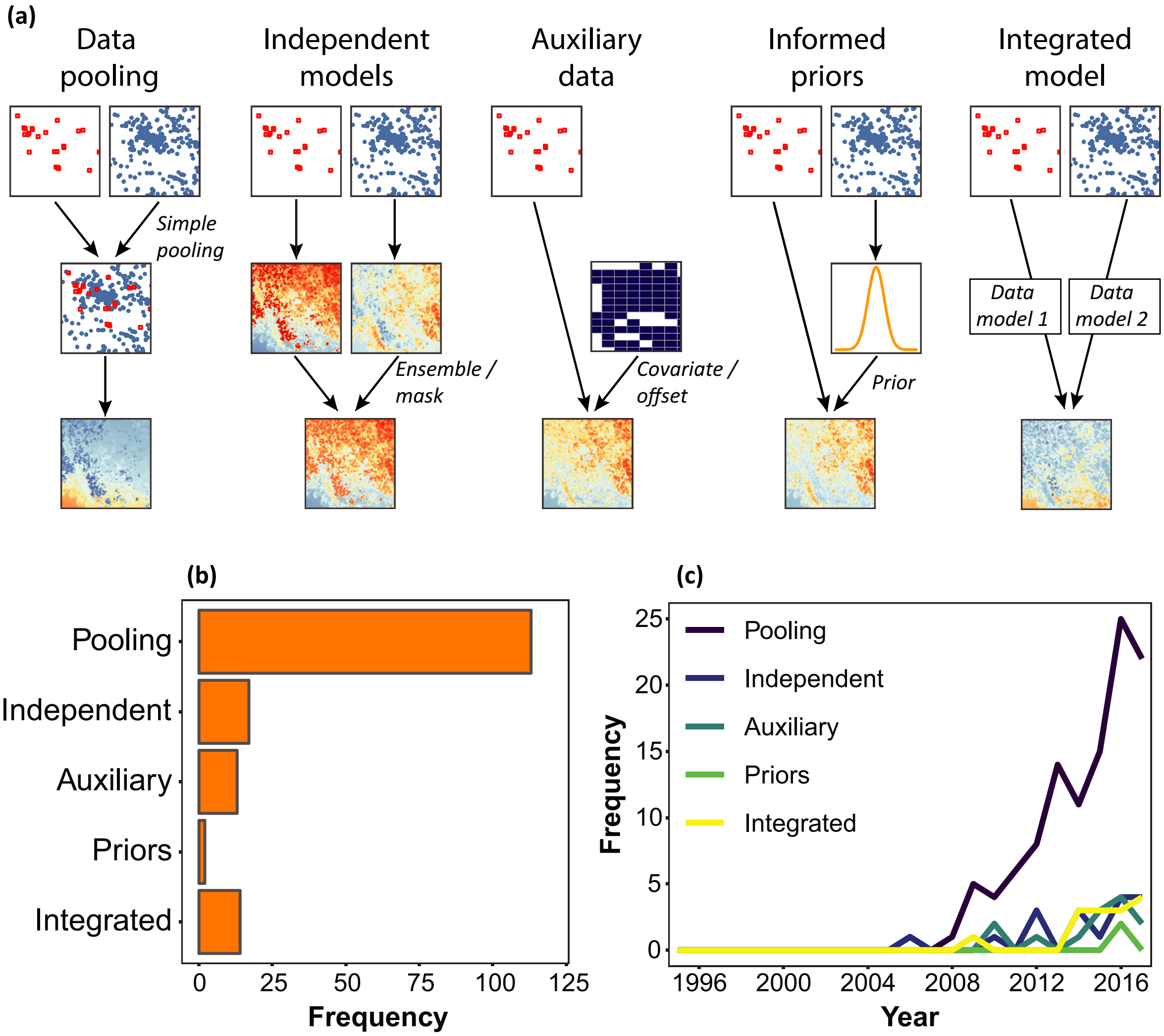

FIG. 2. Combining data for predicting species distributions. Across 353 articles reviewed, (a) there were five general approaches to combining data. First, simple pooling occurred, where different data sources were combined and a single model was fit. Second, independent distribution models were fit to different data sources and results were combined, either through ensemble techniques or through masking/clipping. Third, auxiliary data (not directly species occurrence or abundance) were used in modeling building, typically through the use of covariates or offsets. Fourth, one data source was used to create an informative prior for modeling the primary data source in a Bayesian modeling framework. Finally, multiple data sources were formally integrated by developing separate data models for each source that could then be combined, typically through the use of joint likelihoods. (b) The overall frequency of each approach and (c) the frequency of each approach over time.

at **s** or within the neighborhood of **s** using a buffer or kernel to summarize covariates (e.g,. McCarthy et al. 2012, Chandler and Hepinstall-Cymerman 2016).

The connection between presence-only data, which tend to be the exact coordinates of where individuals occurred, and presence–absence and count data, which tend to be assigned to grid cells, can be understood by using a statistical technique known as the change of support (Cressie and Wikle 2011, Pacifici et al. 2019). The support of a probability distribution is the set of all random values that can be generated. For example, the support of the Poisson distribution is the set of all non-negative integers: $\{0,1,2, \ldots, \infty\}$. The support of the IPP in Eq. 1 is the infinite set of all possible locations within the study area (S). In contrast, the support of count data is limited to a finite number of grid cells where the random variable is the number of points within each cell. The IPP can be formally connected to count data by changing the support from a continuous spatial support to a discrete spatial support. By applying a change of support, the number of points (y_j) contained within the j^{th} grid cell (A_j) becomes a Poisson random variable

Table 1. Some characteristics of different approaches for combining data.

Characteristic	Simple pooling	Independent models	Auxiliary data	Informed priors	Integrated models
Can account for different sampling issues	No	Yes	Yes	Yes	Yes
Can account for variation in spatial or temporal support among data	No	Yes	No	Yes	Yes
Can account for uncertainty from both data sources	No	No	No	Yes	Yes
Can allow for different predictors for each data source	No	Yes	Yes	Yes	Yes
Sequential vs. simultaneous modeling of data sources	Simultaneous	Sequential	Sequential	Sequential	Simultaneous

$$y_j \sim \text{Poisson}\left(\int_{A_j} \lambda(\mathbf{s})d\mathbf{s}\right), \qquad (3)$$

where the expected number of points is $\int_{A_j} \lambda(\mathbf{s})d\mathbf{s}$. Presence–absence data are related to count data by recording the grid cells where $y_j = 0$ or $y_j > 0$. This can be accomplished using a well-known relationship between Poisson and Bernoulli random variables

$$I(y_j > 0) \sim \text{Bernoulli}\left(1 - e^{-\int_{A_j} \lambda(\mathbf{s})d\mathbf{s}}\right), \qquad (4)$$

where $I(y_i > 0)$ is an indicator function that takes on a value of 1 if the species is present and 0 if the species is absent.

Because the IPP can be used to describe several distribution modeling frameworks and provides a means to relate different response variables (Eqs. 3 and 4), the IPP and its extensions provide a formal means to compare the benefits and limitations of different approaches for combining data to species distributions, as we shall see below.

Simple pooling.—Pooling of data is common when using opportunistic presence-only data that are collated from different sources (e.g., museum specimens, biological inventories; Domisch et al. 2016). In these situations, different data sources are combined to model species distribution without explicit acknowledgement of the data sources in the modeling process. Although simple pooling can be helpful to increase the number of point occurrences used for modeling, data sources often vary in their sampling designs, types of bias, and response variables with different attributes (e.g., variation in spatial or temporal support, or the variation in spatial grain and time frame of samples), such that simple pooling may be of limited value.

When presence-only, presence–absence or count data are pooled, it is not clear which distribution should be used to model the combined data. For example, if presence-only data and count data are indiscriminately pooled, then either the Poisson distribution or the IPP could be chosen; however, both require ad hoc manipulation of the data. To use the IPP for count data, one must deaggregate counts and assign the points within each grid cell to a spatial location, thus generating location error (Hefley et al. 2017). Conversely, to use the Poisson distribution one must assign the exact locations contained within the presence-only data to a grid cell (of potentially arbitrary size), which can lead to unwarranted conclusions when the inference at the location of individuals differs from the inference at the grid cells (Hefley et al. 2017).

Combining independent models.—Species distribution modeling has a long history in combining or contrasting predictions from different models. For instance, ensemble modeling techniques are frequently applied to combine predictions from different modeling algorithms (Araújo and New 2007). With respect to multiple data sources, the focus is on developing independent models for each data source and then combining them in some way. This two-step process can be useful for understanding how different data sources can vary in terms of statistical inference and predictions. However, independent models do not allow for "sharing" of information across data sources to estimate parameters more accurately. In addition, it is difficult to combine parameter estimates formally. For example, count data may be modeled using Poisson regression, whereas the IPP may be used to model presence-only data. Both Poisson regression and the IPP model can be used to estimate coefficients, but unless the models have a common spatial support (i.e., the same spatial resolution) the coefficients are not directly comparable. Unless the change of support is applied, Poisson regression explains how the expected abundance among grid cells changes because of spatial covariates, but coefficient estimates are scale dependent and inference is only valid for the spatial resolution of the grid cells. IPP regression coefficients, however, explain how the intensity function (a continuous function with infinite resolution) varies over continuous space, which can be linked to abundance at any spatial resolution. Because the spatial scale of inference differs, comparisons of parameters and statistical inference between the two models should be avoided.

Using auxiliary data.—The use of auxiliary data has been commonly applied in two ways. First, some models have included one species as a covariate to explain the distribution of another species (Araújo and Luoto 2007, Trainor and Schmitz 2014). Second, some modeling approaches have included secondary data to inform modeling the species of interest (e.g,. range maps; Merow et al. 2017). This latter approach has varied considerably. For example, Petitpierre et al. (2016) used information from a global SDM to inform background (pseudoabsence) point selection for a local SDM. Merow et al. (2016) linked prior sources on the relative rate of observing a species occurrence and dispersal-related information data by including it as an offset term in the IPP (Eq. 2). This relationship provides a straightforward means to incorporate auxiliary data. But, similar to combining independent models, it is often unclear how variation in the spatial or temporal support of the data can be accounted for when modeling the primary data source.

Incorporating informed priors.—The use of informed priors in ecological models has been suggested for many years (Ellison 1996), although in practice it has rarely been implemented when combining data for species distribution modeling, where only 1% of the articles we reviewed used this approach (Fig. 1). In this case, one source of data is used to provide a prior distribution for one or more parameters when modeling the second data source. Using informed priors requires that the model for the first set of data (used to obtain the prior) has the same parameters as the model for the second source of data. For example, a prior derived from Poisson regression of counts should not be used as a prior for IPP regression with presence-only data (unless the change of support was used for the Poisson regression model). Using informed priors is similar to the use of joint likelihoods in integrated models (Pawitan 2001; see Integrated distribution models). For instance, when using informed priors, the posterior distribution is proportional to the product of the likelihood for the second source of data and the prior, which is technically equivalent (in a Bayesian paradigm) to the use of joint likelihoods as the product of two component likelihoods based on each data source (see Integrated distribution models). Practical differences include that in this case the approach is sequential and requires priors for all parameters, whereas with the use of joint likelihoods estimation is simultaneous and need not use priors (Table 1).

Integrated distribution models.—Integrated species distribution models (ISDMs) have focused on two ways in which different types of distribution data can be formally combined. First, models have integrated coarse-grain data, such as atlas data, with fine-grain data to downscale predictions of models (Keil et al. 2014). Second, models have been developed to link data of differing quality of information and amount, with a focus on linking presence-only data with other data sources to address known limitations of this data type (Dorazio 2014, Fithian et al. 2015). Overall, this approach has several desirable characteristics (Table 1) and has been shown to improve estimates of environmental relationships and predictions of species distributions (Appendix S1: Table S1), so we provide more detail on this approach.

Models that integrate coarse-grain and fine-grain data aim to link presence–absence or count data being collected at different scales. For example, Keil et al. (2014) developed a model that links coarse-scale gridded atlas data with fine-scale points. The IPP framework can be used to link data of differing resolutions in this way. Let a_i be presence–absence data collected at $i = 1, 2, \ldots, n_\mathrm{a}$ coarse-grain grid cells and let b_j be presence–absence data collected independently of a_i at $j = 1, 2, \ldots, n_\mathrm{b}$ fine-grain grid cells. Separate (independent) binary regression models for each scale could be proposed as follows:

$$a_i \sim \mathrm{Bernoulli}\left(1 - e^{-\int_{A_i} \lambda(\mathbf{s})d\mathbf{s}}\right), \tag{5}$$

$$b_j \sim \mathrm{Bernoulli}\left(1 - e^{-\int_{B_j} \lambda(\mathbf{s})d\mathbf{s}}\right), \tag{6}$$

where A_i and B_j are the grid cells associated with a_i and b_j, respectively. Parameters associated with the intensity functions, such as the regression coefficients $\boldsymbol{\beta}$ in Eq. 2, can be estimated using standard techniques, such as maximum-likelihood estimation.

Because the models for both course-grain and fine-grain data (Eqs. 5 and 6) share a continuous spatial support, parameters associated with the intensity function, such as the regression coefficients $\boldsymbol{\beta}$ in Eq. 2, are directly comparable. Independent estimation, however, does not enable "sharing" of information across data sources to estimate regression coefficients more accurately. Simultaneous estimation remedies this problem, and the likelihood function provides a means for integrating two or more statistical models that share common parameters (Pawitan 2001).

Successful use of multiple sources of data requires carefully considering not only the spatial support, but also the differences in data collection methods. For example, when combining presence–absence data from planned surveys with presence-only locations obtained from opportunistic surveys, it is important to consider bias that may result from spatially heterogeneous search effort. To appropriately combine presence–absence data with presence-only data, Fithian et al. (2015) proposed the following model:

$$a_i \sim \mathrm{Bernoulli}\left(1 - e^{-\int_{A_i} \lambda(\mathbf{s})d\mathbf{s}}\right), \tag{7}$$

$$\mathbf{y} \sim \mathrm{IPP}(\lambda(\mathbf{s})b(\mathbf{s})), \tag{8}$$

where a_i is the presence–absence data collected at grid cell A_i and $\mathbf{y}$ is a matrix that contains the coordinates of the exact locations of the *j*th individuals. The number of locations of presence-only data may depend on location-specific factors that influence search effort (e.g., distance to nearest road). To account for this issue, Fithian et al. (2015) included a spatially varying thinning function $b(\mathbf{s})$ that allows for the possibility of some individuals being missed (when $b(\mathbf{s}) < 1$) or counted more than once (when $b(\mathbf{s}) > 1$; see also Dorazio 2014). Similar to the intensity function, $b(\mathbf{s})$ may depend on covariates. The Fithian et al. (2015) model can also accommodate count data by replacing Eq. 7 with Eq. 3.

In many situations, both data sources are observed with error. For example, it is common that presence–absence data from planned surveys are contaminated with false-negative errors. There is a large literature on models that account for the data collection process (termed "observation models" or "data models") when modeling presence–absence or count data, which include the well-known occupancy and N-mixture models (MacKenzie et al. 2002, Royle 2004). Models that account for the data collection process are often expressed hierarchically and include a level for the unobserved true presence–absence or counts. The true counts are typically modeled with a Poisson distribution; thus Eq. 3 can be used. Similarly, the true presence or absence can be modeled with Eq. 4. This approach combines presence-only data with abundance and/or presence–absence data, but it also explicitly accounts for imperfect detection (Dorazio 2014, Koshkina et al. 2017).

Once models are formulated for each data source, the joint likelihood for the integrated model is the product of the of the component likelihoods:

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{m=1}^{M} L_m(\boldsymbol{\beta}, \boldsymbol{\theta}_m), \tag{9}$$

where $\boldsymbol{\beta}$ are shared parameters (among component likelihoods) explaining $\lambda(\mathbf{s})$ of the IPP and $\boldsymbol{\theta}_m$ are parameters associated with the observation process of the m^{th} data set (e.g., probability of detection for presence–absence data) and are typically not shared. Combining likelihoods using the product requires the assumption of independence among the data sets; however, this assumption could be relaxed (Pawitan 2001).

This framework is quite flexible and many extensions have been considered (Appendix S1: Table S1). For instance, Fithian et al. (2015) applied regularization to the parameter estimation of their model and noted that inclusion of basis functions, such as splines, would be straightforward to implement. Regularization is a method to reduce overfitting and perform model selection, which often increases predictive performance of species distribution models (Gaston and Garcia-Vinas 2011). Pacifici et al. (2017) developed a similar framework that also accounted for spatial dependence through the use of a conditional autoregressive structure, which is another common issue that arises in species distribution models that may affect inferences on a wide variety of problems (Fletcher and Fortin 2018). Giraud et al. (2016) applied a similar set of ideas to model entire communities of species.

An illustration with breeding birds across the Southeastern United States

To illustrate the use of combining data in distribution models, we link data from planned surveys on bird communities in managed forests across the southeastern United States with eBird data. The motivation for these planned surveys was to understand how variation in forest management impacts bird communities, with a focus on forest management practices being considered for bioenergy production (Gottlieb et al. 2017). Reliable estimates of bird occurrence were needed to understand effects of land-use change from bioenergy, which were investigated in detail in Gottlieb et al. (2017). However, there was also interest in making projections of potential effects across the region, given the potential for large-scale land-use changes arising from an increase in bioenergy in the southeastern United States (Galik and Abt 2016). Here, we focus on using these data to project species distributions across the region; see Gottlieb et al. (2017) for inferences on bioenergy.

Methods.—We sampled birds at 78 sites across three breeding seasons, April–July, 2013–2015 (Fig. 3a). Sampling occurred in slash (*Pinus elliottii*) and loblolly (*Pinus taeda*) pine plantations and mature, naturally regenerated longleaf (*Pinus palustris*) pine savannas among three geographic strata within the Southeastern Plains and Southern Coastal Plains ecoregions in Florida, Georgia, and Alabama. At each site, we surveyed birds along two, 200 × 100 m transects between April 1 and June 30 of each year. Each transect was surveyed on three sampling visits separated by approximately 4 weeks, with each transect being surveyed twice on each visit to account for variation in imperfect detection. Here, we pooled information from the two surveys within each visit, such that our detection history included three samples per site. These data can be modeled with a scale-invariant occupancy model of Koshkina et al. (2017), which adjusts the general occupancy model (MacKenzie et al. 2002) with Eq. 4. See Appendix S1 for more details.

Our second data source came from eBird (Sullivan et al. 2014). eBird records include the time, date, and location of bird observations. eBird data can potentially be modeled in two ways. First, data can be modeled assuming it provides presence-only information. Second, data can be modeled as presence–absence data (or more appropriately, "detection-nondetection" data), because much of the eBird data includes a complete "checklist" of all species on the survey. Here, we treat eBird

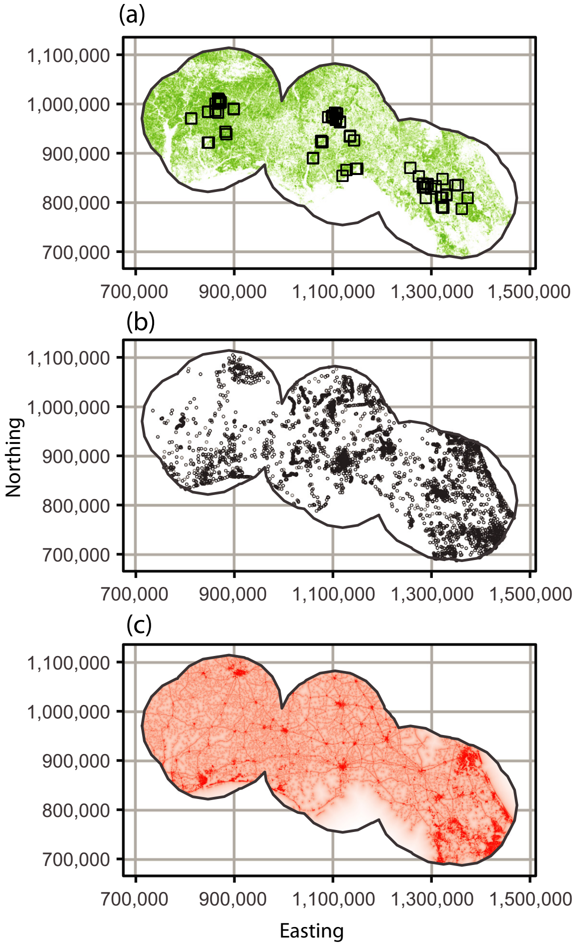

FIG. 3. An example for integrating data in distribution models for 24 species of breeding birds in the southeastern United States. (a) Study area for planned surveys, where line transects at 78 sites stratified across three regions were surveyed over multiple visits. Green shading illustrates forest cover across the region. (b) Presence-only data from eBird across the region (n = 120,484 observations across 24 species). (c) Distance from urban areas (log-transformed; dark red = closer to urban areas). Note the similarity of c to detections from eBird in b.

observations as presence-only data. We downloaded eBird data for species locations from April to June, 2013–2015, the time period of planned surveys (Fig. 3b). We only considered eBird locations within 100 km of the planned surveys occurring during the sampling year. We selected 100 km because biomass for bioenergy is expected to be extracted within approximately 75–100 km from biomass plants (Evans et al. 2013). Using this extent resulted in 120,484 observations across the species considered, with sample size varying from 1,400 to 11,900 per species. These data were then considered using the IPP model with a spatially varying thinning function (Eq. 8).

We fit independent SDMs and contrast these models to an ISDM. We fit models to 24 species of birds (all species that were detected in ≥25% of planned surveys; Appendix S1: Table S2). We chose this cutoff to allow for adequate estimation of the influence of covariates that may affect both occupancy and detection probability in point transect data, which suffered from relatively small sample sizes in comparison to the eBird data. For our presence-only model, we generated quadrature (i.e., background) points by overlaying a 2-km grid across the study area (n = 46,573 points; Renner et al. 2015). Our integrated model used a joint likelihood to combine the presence-only and site-occupancy models (Eq. 9).

Different data sets may vary substantially in sample size as ours do here. This may lead to the larger-sized data set dominating the results because the integrated joint log-likelihood function is additive (Eq. 9), such that the larger data set contributes more to the joint likelihood. Data weighting is a potential solution for possible bias arising from variation in the quality and amount of different data sources used in integrated models (Francis 2011). To address this potential issue, after fitting the ISDM we altered it to include a weight, w, on the joint likelihood function as:

$$\begin{aligned}\log(L(\boldsymbol{\beta},\boldsymbol{\theta})) =& w\log\left(L_{\mathrm{po}}\left(\boldsymbol{\beta},\boldsymbol{\theta}_{\mathrm{po}}\right)\right)\\ &+(1-w)\log\left(L_{\mathrm{pa}}\left(\boldsymbol{\beta},\boldsymbol{\theta}_{\mathrm{pa}}\right)\right),\end{aligned} \tag{10}$$

where $0 < w < 1$, such that $w = 0.5$ would be equivalent to no weighting. We profiled across values of w (ranging from 0.005 to 0.995) for each species and assessed the utility of different weights using three-fold block validation (see below; Wang and Zidek 2005).

For ISDMs, we considered several covariates. We used the 2011 National Landcover Dataset (Homer et al. 2015) to calculate the proportion of forest (pooling deciduous, mixed, and conifer forest categories) and agriculture in the landscape, calculated at a 1-km buffer scale (we initially considered 0.5–5 km scales, but scales were highly correlated; Fletcher and Fortin 2018), as well as Easting and Northing coordinates (**s**) calculated from the center of transects for detection data. In general, we expected that the proportion of forest would have positive effects on most-species, and the proportion of agriculture would have negative effects. We included Easting and Northing coordinates to accommodate broad-scale spatial variation in species distributions across the region. For sample selection bias in eBird data, we consider distance to urban areas (Fig. 3c). For imperfect detection, we considered a linear effect of date (1 = Jan 1; 365 = Dec 31). Prior analysis of point transect data provided little support for nonlinear effects of date on detectability (only 1 of 31 species in Gottlieb et al. 2017 had a nonmonotonic relationship of date), so we only included a linear effect here. All covariates were centered and scaled for analysis.

We evaluated the utility of models with the use of threefold block validation (Wenger and Olden 2012), where we built models with two of the geographic regions considered and predicted onto a third region not considered in model building (where we partitioned both data sources). We assessed predictive accuracy using the area under the ROC curve (AUC) statistic and block-validated log-likelihoods (Fithian et al. 2015). AUC is a commonly used statistic for model discrimination, whereas block-validated log-likelihoods better capture model calibration (Lawson et al. 2014). Models were implemented using the code provided in Koshkina et al. (2017), extended to account for data weighting (Eq. 10).

Results.—When contrasting parameter estimates among models for 24 species, four general patterns emerged. First, eBird data showed strong evidence of sample selection bias, where coefficients for distance from urban areas were significantly negative for all 24 species (Fig. 4a; Appendix S1: Table S3). Second, numerical maximization algorithms did not converge for site-occupancy models of four species (blue grosbeak, pileated woodpecker, downy woodpecker, mourning dove; see Appendix S1: Table S2 for scientific names), although integrated models did converge for these species (Appendix S1: Tables S4, S5), thereby allowing inference on these species. Third, integrating data sets improved the precision of parameter estimates relative to occupancy models (Fig. 4b), although precision did not increase relative to presence-only models (Fig. 4a). In addition, for parameters that did not have shared information between data sets, such the use of distance to urban areas as a spatial sample bias covariate in presence-only data and date as a detectability covariate, estimates and associated uncertainty tended to be similar between integrated models and each independent model (Fig. 4). Fourth, estimates from the integrated models were largely driven by the abundant presence-only data when weighting was not explicitly considered (Fig. 4a). This effect became more severe as the amount of eBird data increased, where parameter estimates of the ISDM tended to shrink toward the estimates from the presence-only IPP model (Fig. 4 inset).

Given that the integrated joint log-likelihood function is additive (Eq. 9), the component from the eBird IPP

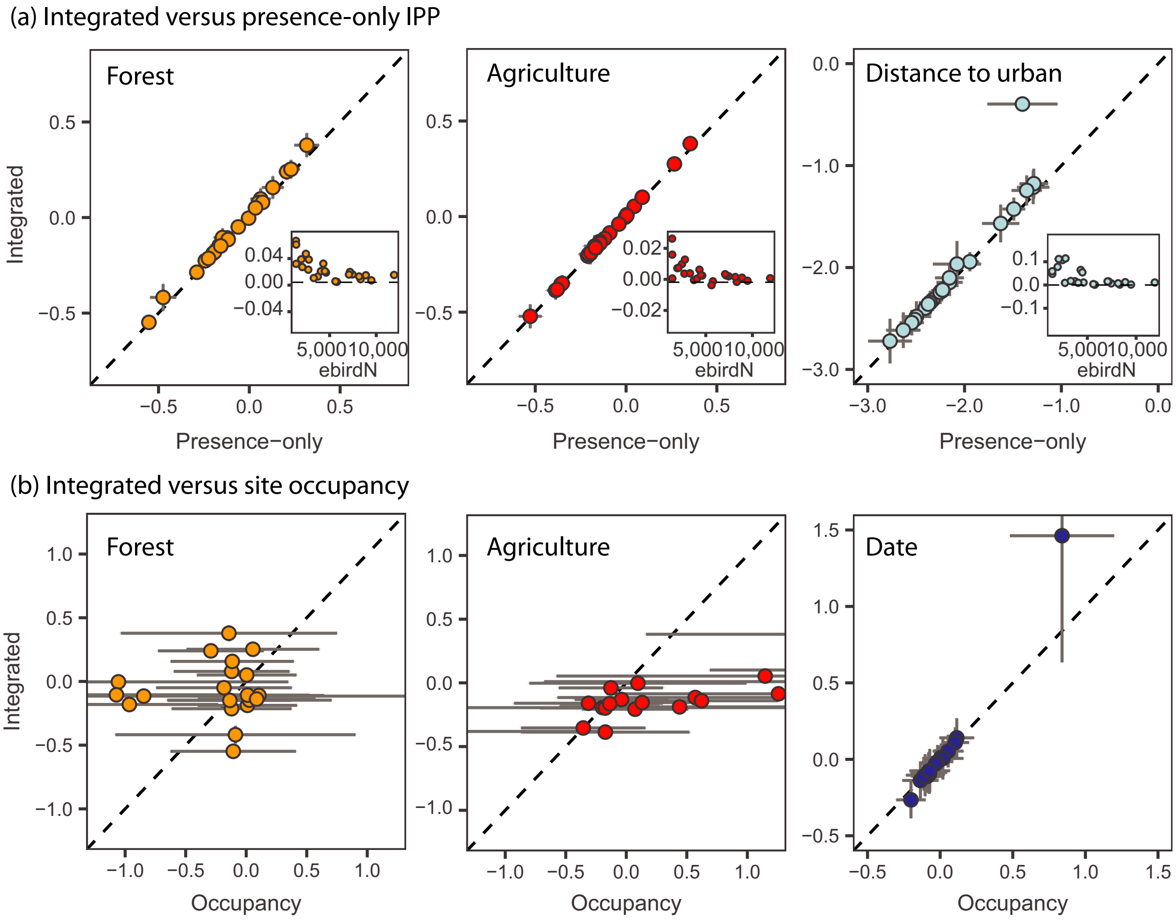

FIG. 4. Parameter estimates (±95% CI) for integrated models are largely driven by abundant presence-only data. Shown are comparisons of (a) presence-only point process (IPP) and (b) site-occupancy models to integrated models (ISDMs), focusing on differences in parameter estimates for the proportion of forest and agriculture within 1 km, distance to urban areas (a sampling bias covariate), and date (a detectability covariate). Insets show the difference between integrated and presence-only models as a function of the amount of eBird data used, illustrating that as sample size increases, parameter estimates converge.

model appeared to be contributing much more to the joint likelihood, driving the results in the integrated model. When using a weighted joint likelihood (Eq. 10), we found that, for most species, providing greater weight to occupancy data increased predictive performance (Fig. 5a), based on both AUC and block-validated log-likelihoods. Note that for the few species where weighting presence-only data was preferred (e.g., $w > 0.5$), these species tended to be those for which site-occupancy models did not converge (e.g., blue grosbeak, pileated woodpecker). Parameter estimates were relatively stable with changes in w until w reached relatively small values (<0.1); at these small values, parameter estimates and associated uncertainty were pulled toward estimates from occupancy (Appendix S1: Figs. S1, S2, Table S6). In general, there was a weak negative correlation with the best w selected and the amount of eBird data used for each species ($r_s = -0.34$ based on AUC; $r_s = -0.14$ based on block-validated log-likelihoods).

Using the most supported weight for joint likelihoods, we found that ISDMs tended to predict species distributions similar to or better than using presence-only data, with ISDMs having higher AUC and block-validated log-likelihoods for 21 and 23 species, respectively (Fig. 5b). When contrasting ISDMs to occupancy models, there was more mixed support, where 14 species of 20 species (where algorithms converged for occupancy models) were better predicted with ISDMs based on AUC, while ISDMs tended to predict distributions better than occupancy models based on block-validated log-likelihoods for all species considered.

Discussion

Our review and synthesis identified that there are several ways to combine data for modeling species distributions. These approaches vary in the ability to account for a variety of issues. Integrated models based on the IPP framework that formally link different types of data have several desirable properties, providing perhaps the most flexible framework for combining data (Table 1). Our illustration with 24 bird species across the southeastern United States provides a comprehensive example of these benefits.

Multiple data sources and the value of data integration

Data sources vary in the amount and reliability of information for modeling species distributions. We contrasted detailed planned surveys that allowed for estimating imperfect detection, but were limited in the number of samples, with presence-only data from eBird that was plentiful but suffered from biases associated with distance from urban areas. We note that checklist data can also be used from eBird, which provides potential information on species absence but does not provide a clear means for estimating imperfect detection (Pacifici et al. 2017). We found that there was much greater precision in estimated environmental relationships from eBird data, which was largely driven by the greater sample size of those data. Yet the ability of models based on presence-only data to predict to new regions (i.e., transferability; Wenger and Olden 2012) was relatively limited. Integrated models generally improved predictions, but evaluation of data weighting emphasized that the occupancy data should be weighted more than presence-only data for improving model predictions (Fig. 5).

Data integration has generally been shown to improve distribution models (Appendix S1: Table S1) in two major ways. First, precision of estimated environmental relationships generally increases. Second, predictive performance of models can also increase. Data

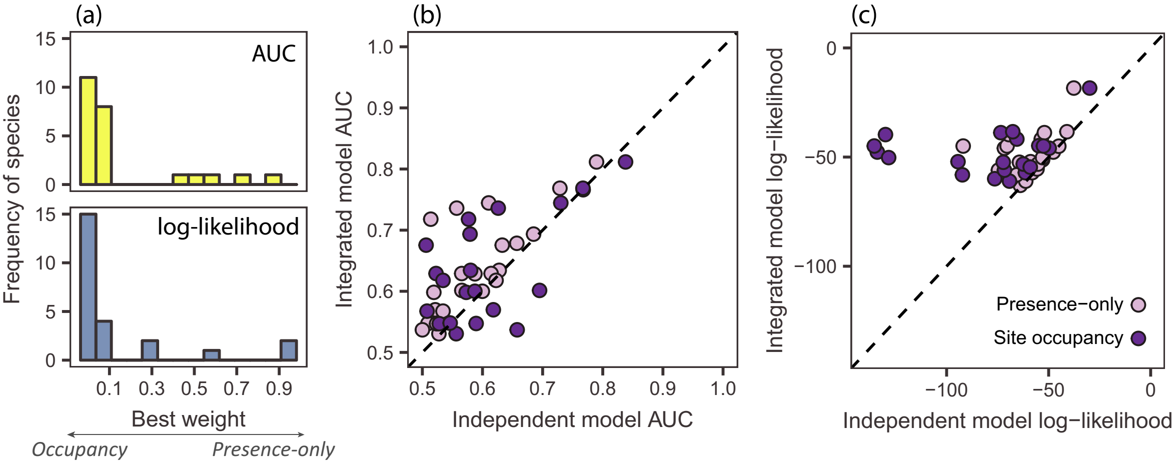

FIG. 5. Weighting of joint likelihoods and block validation. (a) Distribution of best weights, w, based on AUC Area under the ROC curve; (a metric of discrimination) and block-validated log-likelihoods (a metric of calibration) for 24 species, where smaller values indicate greater weights to the occupancy component of the integrated model (ISDM). (b, c) Differences in predictive performance of presence-only, occupancy, and integrated distribution models (using best weighting scheme) using (b) AUC and (c) block-validated log-likelihoods.

integration can also solve several long-standing problems with modeling distributions. For instance, the development of ISDMs was largely motivated by the need to improve presence-only data and resulting models (Dorazio 2014, Fithian et al. 2015, Bradter et al. 2018) where presence-only data do not provide information on species prevalence, limiting the potential to estimate the probability of occurrence (Guillera-Arroita et al. 2015). Data integration has been argued to improve modeling through the potential ability to better account for sample selection bias (Fithian et al. 2015) and for estimating abundance or occurrence across regions through integrated intensity functions (Hefley and Hooten 2016).

Big data, citizen science, and data weighting

Big data are increasingly used for modeling species distributions. The term “big data” is broadly used to describe large digital datasets arising from recent advancements in information technology and data acquisition (Jin et al. 2015). These big data are increasingly available for ecologists as biological observations are now being generated, captured, and processed at unprecedented scales and rates through citizen science (such as eBird), remote sensing, and other means (Hashem et al. 2015, La Sorte et al. 2018). Although big data provide a wealth of information, the use of big data presents a number of challenges such as massive volumes, high dimensionality, and sampling biases (Dickinson et al. 2010, Tye et al. 2017). Statistical advances have been developed for dealing with these issues, for example, by using species distribution models applied to spatiotemporally restricted extents and adapting to variation in sampling intensity (Fink et al. 2010). ISDMs, however, provide untapped potential for combining big data with data generated from more targeted surveys to balance the tradeoffs between different sources of bias and improve model performance and inference.

Despite the potential advantages of integrating different data sources for species distribution modeling, several challenges remain. When integrating data for species distribution modeling, disparity in the amount of data from different sources can lead to one data source contributing substantially more to the joint likelihood than another data source. The problem of data weighting for integrated population models has been emphasized in fisheries (Francis 2017, Punt 2017, see also Saunders et al. 2019), and a variety of data weighting schemes have been considered. For example, individual data points have been weighted based on measures of variability (e.g., SD, CV of data used to generate abundance indices) or sample size (e.g., number of fish caught for compositional data) for each sampling unit (Francis 2011, Punt 2017). Entire data sources, rather than individual data points, have also been weighted based on sample size and measures of variability of the data sources (Francis 2011). Some types of weighting have been shown have desirable asymptotic properties in maximum-likelihood estimation (Wang and Zidek 2005), but more work is needed on this issue for ISDMs particularly because there can be precision-bias tradeoffs in estimation when integrating data sources. Two practical issues for data weighting are (1) the weighting scheme used and (2) the values of weights.

In our example, we applied a simple proportional weighting scheme to each component likelihood. This approach provides a practical way to weight two data sources differentially, causing parameter estimates of the integrated model to shift toward estimates of the independent model that has greater weight (Appendix S1: Figs. S1, S2). However, other considerations are relevant. For example, the spatial resolution of the observational data will often vary, as will the models used to integrate such data. Likelihood functions for presence-only data are at a point level when using IPP models, whereas occupancy data occur at a coarser resolution. This difference in sampling resolution could be considered in weighting schemes. An alternative way to view the problem is to consider formally that the data sources vary in reliability and formulate likelihood functions that accommodate variation in reliability (Lele and Das 2000, Lele and Allen 2006).

To determine weight values, we used a cross-validation technique (Wang and Zidek 2005). This post hoc approach was helpful for guiding decisions for predictive accuracy. Other approaches to assigning weights could be considered based on the knowledge of the system. For instance, one may wish to downweight presence-only data, given its known limitations relative to data from planned surveys (cf. Francis 2017). Similarly, data with known sampling selection bias could be downweighted to reduce—but not eliminate—the contribution of those data to modeling efforts. In such situations, determining the magnitude of downweighting may still require either an objective technique, such as cross-validation, or could be driven by subjective criteria (such as expert opinion). Evaluating the impacts of different weighting schemes and advancing better ways to increase the efficacy of data integration are needed.

Guidance

The rapid growth of combining data for modeling species distributions (Fig. 1) and the flexibility of integrated distribution models make it tempting to apply these techniques. Here we provide some guidance regarding prudent applications of combining data.

1) When combining data, avoid simple pooling and use caution with combining independent models. Simple pooling of different data sources masks differences in sampling designs (spatial, temporal, and observational differences). Although independent models

can be useful for qualitative comparisons, quantitative integration can be inappropriate because such models often focus on subtly different quantities and vary in spatial and temporal support.

2) Use observation models that honor sampling designs of different data sources whenever possible. The use of observation models is highly relevant to many problems in ecology (Royle and Dorazio 2008), and it has been repeatedly shown that SDMs that account for imperfect detection improve inferences and model predictions (Rota et al. 2011, Lahoz-Monfort et al. 2014). Given that different sources of data will likely vary in the degree and type of observation errors (Dorazio 2014, Hefley and Hooten 2016, Ruiz-Gutierrez et al. 2016), using observation models may be particularly important when integrating disparate data sources. In a similar way, accounting for sampling bias and design can help improve integration of data sources (Fithian et al. 2015).
3) When combining data, match data temporally. Most integrated modeling approaches that have been applied have implicitly assumed a temporally constant environment (Schank et al. 2017). Here we truncated eBird data to match the sampling time frame of planned surveys. Although these models can be extended to spatio-temporal frameworks (Hefley and Hooten 2016), in practice it may be challenging to couple data collected on very different timescales, such as presence-only data from museum specimens spanning decades and planned surveys spanning a short period of time (e.g., 1 yr).
4) Consider variation in the amount and reliability of different data sources. Because data sources can vary tremendously in data amount, caution should be used when combining data. Weighting joint likelihood functions is a practical way to place greater emphasis on one data source.
5) Broad-scale programs, such as eBird, should consider incorporating planned surveys as part of their sampling strategy. This could be particularly useful for monitoring programs (e.g., biological atlases, camera trap surveys) that have semipermanent sampling locations or grids. In these cases, having a secondary sampling strategy focused on different types of data collection would open the possibilities of using ISDMs for more reliable inferences.

Finally, we note that there may be many situations where integrating data sources may not be needed. Integrating planned survey data is expected to frequently improve presence-only modeling, both in terms of identifying environmental relationships and mapping species distributions, because of the added information in regards to species prevalence and the potential to isolate sample selection bias better (Dorazio 2014). However, in situations where planned survey data are plentiful and cover the geographic extent of interest, there may be little gained in combining it with presence-only information that may be less reliable.

Conclusions

Combining data has great potential to improve understanding of species distributions and predictive models. We focused on linking distributional data only, but other types of formal data integration could occur to better predict the dynamics of species distribution (Miller et al. 2019, Van Schmidt et al. 2019). By integrating multiple sources of information into the modeling process, greater insights into environmental relationships and the mechanisms driving species distributions can occur. With the rapid rise in data availability, we expect integrated models will provide an increasingly common and powerful approach for addressing problems of species distributions and ongoing environmental change.

Acknowledgments

This work was supported by the U.S. Department of Agriculture, USDA-NIFA Initiative Grant No. 2012-67009-20090 and the University of Florida's School of Natural Resource and Environment. We thank B. Brosi, H. Ober, L. Smith, I. Gottlieb, M. Nunez-Regueiro, and numerous field technicians for their assistance. We thank many landowners and land managers who allowed access to their properties, especially Loncala, Inc., and Resource Management Service LLC. T. Arnold, C. Merow, and two anonymous reviewers provided useful suggestions, which greatly improved and clarified the ideas presented here.

Literature Cited

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

Araújo, M. B., and M. Luoto. 2007. The importance of biotic interactions for modelling species distributions under climate change. Global Ecology and Biogeography 16:743–753.

Araújo, M. B., and M. New. 2007. Ensemble forecasting of species distributions. Trends in Ecology & Evolution 22:42–47.

Bradter, U., L. Mair, M. Jonsson, J. Knape, A. Singer, and T. Snall. 2018. Can opportunistically collected citizen science data fill a data gap for habitat suitability models of less common species? Methods in Ecology and Evolution 9:1667–1678.

Case, M. J., and J. J. Lawler. 2017. Integrating mechanistic and empirical model projections to assess climate impacts on tree species distributions in northwestern North America. Global Change Biology 23:2005–2015.

Chandler, R., and J. Hepinstall-Cymerman. 2016. Estimating the spatial scales of landscape effects on abundance. Landscape Ecology 31:1383–1394.

Cressie, N., and C. K. Wikle. 2011. Statistics for spatio-temporal data. John Wiley and Sons, Inc., Hoboken, New Jersey, USA.

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

Domisch, S., A. M. Wilson, and W. Jetz. 2016. Model-based integration of observed and expert-based information for assessing the geographic and environmental distribution of freshwater species. Ecography 39:1078–1088.

Dorazio, R. M. 2012. Predicting the geographic distribution of a species from presence-only data subject to detection errors. Biometrics 68:1303–1312.

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.

Douma, J. C., J. P. M. Witte, R. Aerts, R. P. Bartholomeus, J. C. Ordonez, H. O. Venterink, M. J. Wassen, and P. M. van Bodegom. 2012. Towards a functional basis for predicting vegetation patterns; incorporating plant traits in habitat distribution models. Ecography 35:294–305.

Elith, J., and J. R. Leathwick. 2009. Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology Evolution and Systematics 40:677–697.

Ellison, A. M. 1996. An introduction to Bayesian inference for ecological research and environmental decision-making. Ecological Applications 6:1036–1046.

Evans, J. M., R. J. Fletcher, Jr., and J. Alavalapati. 2010. Using species distribution models to identify suitable areas for biofuel feedstock production. Global Change Biology Bioenergy 2:63–78.

Evans, J. M., R. J. Fletcher, Jr., J. R. R. Alavalapati, A. L. Smith, D. Geller, P. Lal, M. Acevedo, D. Vasudev, J. Calabria, and T. Upadhyay. 2013. Forestry bioenergy in the southeast United States: implications for wildlife habitat and biodiversity. National Wildlife Federation, Merrifield, Virginia, USA.

Fink, D., W. M. Hochachka, B. Zuckerberg, D. W. Winkler, B. Shaby, M. A. Munson, G. Hooker, M. Riedewald, D. Sheldon, and S. Kelling. 2010. Spatiotemporal exploratory models for broad-scale survey data. Ecological Applications 20:2131–2147.

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

Fletcher, R. J., R. A. McCleery, D. U. Greene, and C. A. Tye. 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.

Fletcher, R. J., Jr., and M. J. Fortin. 2018. Spatial ecology and conservation modeling: applications with R. Springer, Switzerland.

Fourcade, Y., J. O. Engler, D. Roedder, and J. Secondi. 2014. Mapping species distributions with MAXENT using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. PLoS ONE 9:e97122.

Francis, R. 2011. Data weighting in statistical fisheries stock assessment models. Canadian Journal of Fisheries and Aquatic Sciences 68:1124–1138.

Francis, R. 2017. Revisiting data weighting in fisheries stock assessment models. Fisheries Research 192:5–15.

Galik, C. S., and R. C. Abt. 2016. Sustainability guidelines and forest market response: an assessment of European Union pellet demand in the southeastern United States. Global Change Biology Bioenergy 8:658–669.

Gaston, A., and J. I. Garcia-Vinas. 2011. Modelling species distributions with penalised logistic regressions: A comparison with maximum entropy models. Ecological Modelling 222:2037–2041.

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

Gottlieb, I. G. W., R. J. Fletcher, M. M. Nunez-Regueiro, H. Ober, L. Smith, and B. J. Brosi. 2017. Alternative biomass strategies for bioenergy: implications for bird communities across the southeastern United States. Global Change Biology Bioenergy 9:1606–1617.

Graham, C. H., S. Ferrier, F. Huettman, C. Moritz, and A. T. Peterson. 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology & Evolution 19:497–503.

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

Guisan, A. et al. 2013. Predicting species distributions for conservation decisions. Ecology Letters 16:1424–1435.

Hashem, I. A. T., I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan. 2015. The rise of “big data” on cloud computing: Review and open research issues. Information Systems 47:98–115.

Hefley, T. J., D. M. Baasch, A. J. Tyre, and E. E. Blankenship. 2015. Use of opportunistic sightings and expert knowledge to predict and compare Whooping Crane stopover habitat. Conservation Biology 29:1337–1346.

Hefley, T. J., B. M. Brost, and M. B. Hooten. 2017. Bias correction of bounded location errors in presence-only data. Methods in Ecology and Evolution 8:1566–1573.

Hefley, T. J., and M. B. Hooten. 2016. Hierarchical species distribution models. Current Landscape Ecology Reports 1:87–97.

Hefley, T. J., A. J. Tyre, D. M. Baasch, and E. E. Blankenship. 2013. Nondetection sampling bias in marked presence-only data. Ecology and Evolution 3:5225–5236.

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

Jin, X. L., B. W. Wah, X. Q. Cheng, and Y. Z. Wang. 2015. Significance and challenges of big data research. Big Data Research 2:59–64.

Keil, P., A. M. Wilson, and W. Jetz. 2014. Uncertainty, priors, autocorrelation and disparate data in downscaling of species distributions. Diversity and Distributions 20:797–812.

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

La Sorte, F. A., C. A. Lepczyk, J. L. Burnett, A. H. Hurlbert, M. W. Tingley, and B. Zuckerberg. 2018. Opportunities and challenges for big data ornithology. Condor 120:414–426.

Lahoz-Monfort, J. J., G. Guillera-Arroita, and B. A. Wintle. 2014. Imperfect detection impacts the performance of species distribution models. Global Ecology and Biogeography 23:504–515.

Lawson, C. R., J. A. Hodgson, R. J. Wilson, and S. A. Richards. 2014. Prevalence, thresholds and the performance of presence–absence models. Methods in Ecology and Evolution 5:54–64.

Lele, S. R., and K. L. Allen. 2006. On using expert opinion in ecological analyses: a frequentist approach. Environmetrics 17:683–704.

Lele, S. R., and A. Das. 2000. Elicited data and incorporation of expert opinion for statistical inference in spatial studies. Mathematical Geology 32:465–487.

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

Marcantonio, M., M. Metz, F. Baldacchino, D. Arnoldi, F. Montarsi, G. Capelli, S. Carlin, M. Neteler, and A. Rizzoli. 2016. First assessment of potential distribution and dispersal capacity of the emerging invasive mosquito *Aedes koreicus* in Northeast Italy. Parasites & Vectors https://doi.org/10.1186/s13071-016-1340-9

McCarthy, K. P., R. J. Fletcher, C. T. Rota, and R. L. Hutto. 2012. Predicting species distributions from samples collected along roadsides. Conservation Biology 26:68–77.

Merow, C., J. M. Allen, M. Aiello-Lammens, and J. A. Silander. 2016. Improving niche and range estimates with Maxent and point process models by integrating spatially explicit information. Global Ecology and Biogeography 25:1022–1036.

Merow, C., A. M. Wilson, and W. Jetz. 2017. Integrating occurrence data and expert maps for improved species range predictions. Global Ecology and Biogeography 26:243–258.

Miller, D. A., J. D. Nichols, B. T. McClintock, E. H. C. Grant, L. L. Bailey, and L. A. Weir. 2011. Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. Ecology 92:1422–1428.

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

Norris, K. 2004. Managing threatened species: the ecological toolbox, evolutionary theory and declining-population paradigm. Journal of Applied Ecology 41:413–426.

Pacifici, K., B. Reich, D. Miller, and B. Pease. 2019. Resolving misaligned spatial data with integrated distribution models. Ecology 100:e02709.

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

Palaoro, A. V., M. M. Dalosto, G. C. Costa, and S. Santos. 2013. Niche conservatism and the potential for the crayfish *Procambarus clarkii* to invade South America. Freshwater Biology 58:1379–1391.

Pawitan, Y. 2001. In all likelihood: statistical modeling and inference using likelihood. Oxford University Press, Oxford, UK.

Petitpierre, B., K. McDougall, T. Seipel, O. Broennimann, A. Guisan, and C. Kueffer. 2016. Will climate change increase the risk of plant invasions into mountains? Ecological Applications 26:530–544.

Phillips, S. J., M. Dudik, J. Elith, C. H. Graham, A. Lehmann, J. Leathwick, and S. Ferrier. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications 19:181–197.

Punt, A. E. 2017. Some insights into data weighting in integrated stock assessments. Fisheries Research 192:52–65.

Raxworthy, C. J., E. Martinez-Meyer, N. Horning, R. A. Nussbaum, G. E. Schneider, M. A. Ortega-Huerta, and A. T. Peterson. 2003. Predicting distributions of known and unknown reptile species in Madagascar. Nature 426: 837–841.

Regos, A., M. D'Amen, N. Titeux, S. Herrando, A. Guisan, and L. Brotons. 2016. Predicting the future effectiveness of protected areas for bird conservation in Mediterranean ecosystems under climate change and novel fire regime scenarios. Diversity and Distributions 22:83–96.

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

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

Rota, C. T., R. J. Fletcher Jr., J. M. Evans, and R. L. Hutto. 2011. Does accounting for detectability improve species distribution models? Ecography 34:659–670.

Royle, J. A. 2004. N-mixture models for estimating population size from spatially replicated counts. Biometrics 60:108–115.

Royle, J. A., R. B. Chandler, C. Yackulic, and J. D. Nichols. 2012. Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. Methods in Ecology and Evolution 3:545–554.

Royle, J. A., and R. M. Dorazio. 2008. Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations, and communities. Academic Press, Cambridge, Massachusetts. USA.

Ruiz-Gutierrez, V., M. B. Hooten, and E. H. C. Grant. 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.

Saunders, S. P., M. T. Farr, A. D. Wright, C. A. Bahlai, J. W. Ribeiro, S. Rossman, A. L. Sussman, T. W. Arnold, and E. F. Zipkin. 2019. Disentangling data discrepancies with integrated population models. Ecology 100:e02714.

Schank, C. J. et al. 2017. Using a novel model approach to assess the distribution and conservation status of the endangered Baird's tapir. Diversity and Distributions 23:1459–1471.

Sullivan, B. L. et al. 2014. The eBird enterprise: An integrated approach to development and application of citizen science. Biological Conservation 169:31–40.

Talluto, M. V. et al. 2016. Cross-scale integration of knowledge for predicting species ranges: a metamodelling framework. Global Ecology and Biogeography 25:238–249.

Trainor, A. M., and O. J. Schmitz. 2014. Infusing considerations of trophic dependencies into species distribution modelling. Ecology Letters 17:1507–1517.

Tye, C. A., R. A. McCleery, R. J. Fletcher, Jr., D. U. Greene, and R. S. Butryn. 2017. Evaluating citizen vs. professional data for modelling distributions of a rare squirrel. Journal of Applied Ecology 54:628–637.

Underwood, J. G., C. D'Agrosa, and L. R. Gerber. 2010. Identifying conservation areas on the basis of alternative distribution data sets. Conservation Biology 24:162–170.

Van Schmidt, N. D., T. Kovach, A. M. Kilpatrick, J. L. Oviedo, L. Huntsinger, T. V. Hruska, N. L. Miller, and S. R. Beissinger. 2019. Integrating social and ecological data to model metapopulation dynamics in coupled human and natural systems. Ecology 100:e02711.

Wang, X. G. and J. V. Zidek. 2005. Selecting likelihood weights by cross-validation. Annals of Statistics 33:463–500.

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

Wenger, S. J., and J. D. Olden. 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. Methods in Ecology and Evolution 3:260–267.

Wiens, J. J. et al. 2010. Niche conservatism as an emerging principle in ecology and conservation biology. Ecology Letters 13:1310–1324.

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

Supporting Information

Additional supporting information may be found in the online version of this article at http://onlinelibrary.wiley.com/doi/10.1002/ecy.2710/suppinfo