CellPress
REVIEWS

Review

# Data Integration for Large-Scale Models of Species Distributions

Nick J.B. Isaac,[1,2,*] Marta A. Jarzyna,[3] Petr Keil,[4,5] Lea I. Dambly,[1,2] Philipp H. Boersch-Supan,[6,7] Ella Browning,[2,8] Stephen N. Freeman,[1] Nick Golding,[9] Gurutzeta Guillera-Arroita,[9] Peter A. Henrys,[10] Susan Jarvis,[10] José Lahoz-Monfort,[9] Jörn Pagel,[11] Oliver L. Pescott,[1] Reto Schmucki,[1] Emily G. Simmonds,[12] and Robert B. O'Hara[12]

**With the expansion in the quantity and types of biodiversity data being collected, there is a need to find ways to combine these different sources to provide cohesive summaries of species' potential and realized distributions in space and time. Recently, model-based data integration has emerged as a means to achieve this by combining datasets in ways that retain the strengths of each. We describe a flexible approach to data integration using point process models, which provide a convenient way to translate across ecological currencies. We highlight recent examples of large-scale ecological models based on data integration and outline the conceptual and technical challenges and opportunities that arise.**

## Species Distribution Models in Ecology

Large-scale ecological models of how species distributions and abundances vary over space and time are a critical tool in macroecology, biogeography, and conservation biology. They underpin our understanding of how biodiversity is shaped, how it is responding to anthropogenic activities, and how it might change in the future [1–3]. There is now a substantial literature on statistical tools for building **species distribution models (SDMs)** (see Glossary) and best practice in how to fit them [4–7]. SDMs also form a building block upon which more complex models, incorporating occupancy and/or abundance in space and time, can be built [8,9].

The information available for models of species' distributions is radically changing, thanks to a digital and technical revolution in data collection [10–12]. New technologies, such as camera traps, miniature geolocation devices, environmental DNA (eDNA), and passive acoustic monitoring [13–17], are creating new opportunities for surveying wildlife in space and time. These developments, allied with initiatives for data mobilization [16–18] and the rapid growth of citizen science [19,20], mean that ecological data are being generated at an unprecedented rate, in an ever-increasing number of formats and currencies.

## The Challenge of Data

All these data types are potentially informative about abundance and occurrence of species, and the processes that drive their dynamics in space and time. However, getting the most from the data revolution is challenging, since datasets are typically designed with a particular goal in mind, such that different data types have characteristic strengths and weaknesses. Conventional modeling approaches are each built around the properties of one particular data type; for example, the widely used SDM method MaxEnt [21] is designed to work with **presence-only data**, while **occupancy-detection models** [6] require that observations are replicated in order to estimate detection parameters.

Faced with a plethora of heterogeneous data types, it is now commonplace that more than one relevant dataset is available for any large-scale ecological question. Traditionally, modelers in this situation would be forced either to ignore any differences in how the datasets were collected or to choose between them. A common choice is between small quantities of **structured data** and large quantities of **unstructured data** [22]. Structured data derive from surveys with repeatable protocols and/or a stratified sampling design; these are expensive to collect and tend to be geographically restricted [23–25]. Unstructured data constitute the majority of available information [e.g., the Global

## Highlights

Integrated modeling of species distributions and abundance is emerging as a powerful tool in statistical ecology.

Point processes provide a flexible framework for developing integrated models, combining data representing the locations of individual organisms, local population abundance, and species–site occupancy.

These methods provide opportunities to make best use of existing and new data sources.

We expect that data integration will underpin the next generation of models predicting the current, future, and potential distributions of species.

[1]Centre for Ecology and Hydrology, Benson Lane, Crowmarsh Gifford, Wallingford, OX10 8BB, UK

[2]Centre for Biodiversity and Environment Research, Department of Genetics, Evolution and Environment, University College London, London, WC1E 6BT, UK

[3]Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, OH 43210, USA

[4]German Centre for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig, 04103 Leipzig, Germany

[5]Institute of Computer Science, Martin Luther University Halle-Wittenberg, 06120, Halle (Saale), Germany

[6]British Trust for Ornithology, Thetford, IP24 2PU, UK

[7]Department of Geography, University of Florida, Gainesville, FL 32611, USA

[8]Institute of Zoology, Zoological Society of London, London, NW1 4RY, UK

[9]School of BioSciences, University of Melbourne, Parkville, VIC 3010, Australia

Biodiversity Information Facility (GBIF) contains over a billion records] but are affected by numerous forms of bias [26–28].

An emerging alternative is to integrate the different data sources available into a single model in a way that retains the strengths of each. Several approaches to model-based data integration have been proposed [29,30]. In this review, we emphasize a specific formulation that explicitly separates the biological and data generation processes. We demonstrate how this approach can be applied to a wide spectrum of data types, spanning haphazard observations and systematic population counts. This is possible by harnessing a statistical framework that makes the translation between different types of biodiversity datasets clear mathematically, thus permitting the development of models to integrate the data.

[10]Centre for Ecology and Hydrology, Bailrigg, Lancaster, LA1 4AP, UK

[11]Institute of Landscape and Plant Ecology, University of Hohenheim, 70599 Stuttgart, Germany

[12]Department of Mathematical Sciences, Centre for Biodiversity Dynamics, Norwegian University of Science and Technology, N-7491 Trondheim, Norway

*Correspondence: njbi@ceh.ac.uk

## What Is Data Integration?

The process of bringing together data has many labels, for example, data fusion [31], assimilation [31,32], combination [33], or integration [34,35]. Their definitions are often context-specific, and some terms have several meanings [35,36]. Rather than disentangling the semantics, we simply distinguish two ways of bringing data together. Data pooling [30] assumes that any disparities between datasets are small enough to be ignored, or are degraded to a lowest common denominator (Figure 1A); for example, **presence–absence data** can be degraded into presence-only data in order to combine with GBIF data in a presence-only SDM. In this way, data pooling employs an **observation model** that is common to both datasets, although this is rarely explicit. A second approach – the main focus of this review – is more flexible as it can accommodate a wider range of data types, we call it integrated modeling, or model-based data integration (Figure 1B). Integrated modeling aims to describe explicitly the differences in how datasets were assembled, thus retaining the strengths of each and correcting, at least to some extent, their weaknesses. It accommodates the structure and potential biases in each source while propagating as much information as possible about the species' distribution.

Model-based data integration is not new to ecology. The field of **integrated population modeling (IPM)** has long recognized the benefits of using multiple data sources representing different aspects

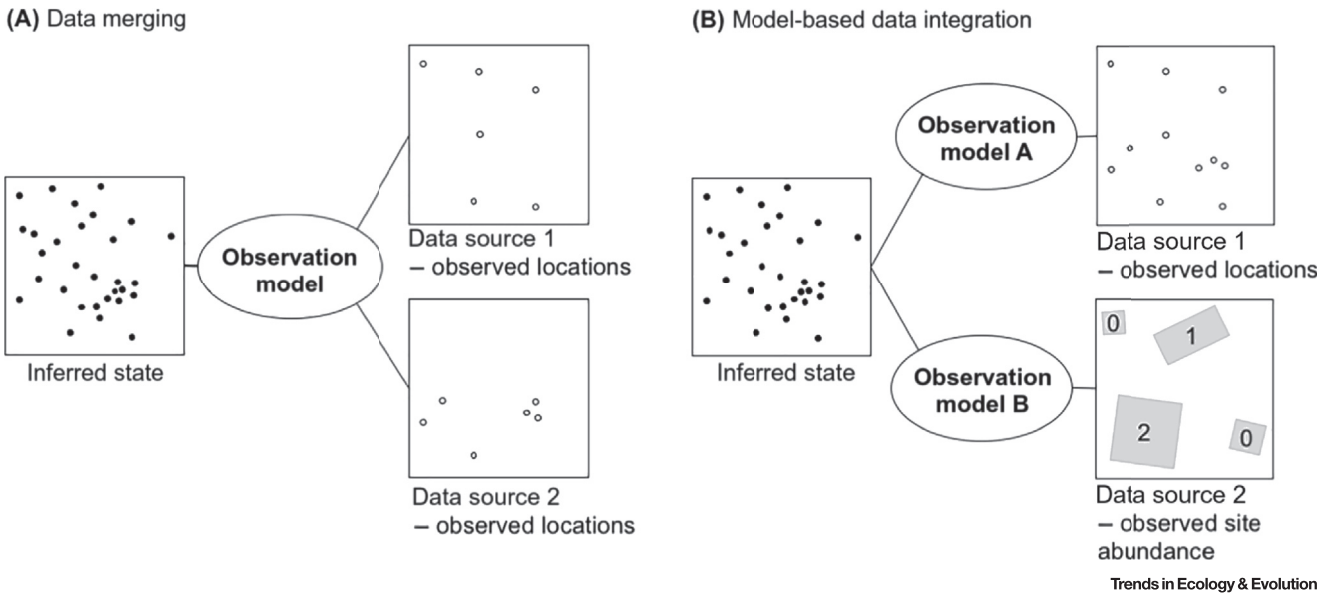


**Figure 1. Data Integration and Pooling.**
Data pooling (A) brings together observations from different sources that could be modeled with a single observation model, ignoring their disparities or reducing the data to a common denominator. In contrast, data integration (B) uses distinct observation submodels for each data source.

of an ecological process [37–39]. The strength of model-based data integration lies in sharing parameters across **submodels**, which allows demographic parameters to be estimated more precisely than through independent models, or actually to be estimated at all [37,39,40].

In contrast to IPM, **integrated distribution modeling** is a new and emerging field [29,41]. In part, this reflects the fact that integrated distribution models have been largely framed as advances in ecological statistics [29,30,42–47] (Box 1). However, we believe it is time for integrated models to leave the preserve of statisticians, and see greater uptake by ecologists. Data integration is facilitated by underappreciated links between species occurrence, abundance, and point locations of individuals, which we discuss next.

## State-Space and Point Process Models

In species distribution modeling, the task is to use our species observations to infer its actual geographical distribution (occurrence or abundance at sites). In models that account for the observation process, this involves two components. While we can generate a statistical description of the species' distribution as some function of environmental covariates and/or time. Unfortunately, the actual distribution cannot be observed directly, it is thus referred to as a **latent state**. We therefore also create a statistical description of how the data were produced using an observation model, for example, accounting for imperfect detection of the species in an occupancy-detection model [6]. In combination, these two (sub)models form a **state-space model**. State-space models are hierarchical, in that observations (i.e., the data) are conditional on the latent state (e.g., assuming it is only possible to observe the species where it truly occurs).

In state-space terminology, an integrated model can be defined by the existence of multiple observation submodels for the same latent state. The latent state, and the parameters describing it, are shared between datasets; alternatives to this joint-likelihood approach may be preferable in some circumstances [29,30,46,47], but lack the clear logical mapping of Figure 1B. Conceptualizing the latent state is relatively simple when all observation submodels refer to a common ecological currency. Examples would be: an occupancy-detection model in which the multiple datasets constitute survey types with differing sampling effort or errors [48]; or a model of species' abundance integrating data from point counts with transect walks. However, it is less clear how to proceed when multiple data types refer to different ecological currencies (e.g., presence–absence data and counts, related to occupancy and abundance). **Point process models** [49] provide a solution that reflects ecologists' intuitive understanding of how multiple data types can emerge from a single system.

## Observation Models for Point Processes

A point process is a statistical description of how points are distributed in space. In an ecological setting, the points can be thought of as the instantaneous location of individual organisms, or their activity centers. The number of points within a particular region is the site abundance, and the presence or absence of points within a region is site occupancy (Figure 2). This general framework thus encompasses a variety of data types and model structures that are commonly used within the ecological literature. The 'process' that describes the location of points is characterized by an intensity surface that represents density of points within a given area, and defines the latent state. The intensity is allowed to vary in space, so higher intensity means the species is more likely to occur at a particular location. For mobile species, the intensity can be interpreted as the distribution of locations over time. When using point processes in a state-space modeling framework, the intensity of the point process can be modeled in the conventional manner (e.g., as a function of rainfall, temperature, etc.).

The simplest situation, conceptually, is one in which the data consist solely of presence-only records (Figure 2, top row). As an example, consider a survey of plants within a meadow; the resulting data are point locations where individual plants were observed. Since detection is nearly always imperfect [6], the observations include only a subset of plants within the meadow. However, if the survey is unbiased, then the locations represent a random sample of the locations where plants actually occur.

### Glossary

**Abundance data**: these data may be in the form of direct counts (i.e., how many individuals were observed) or some index of abundance derived from the raw counts.
**Detection/nondetection data**: term sometimes used to acknowledge that species presence/absence is usually imperfectly observed. More specifically, often used for data that are collected in a way that are informative about the detection process (e.g., via repeat surveys to sites, multiple independent observers, times to detection, etc.).
**Integrated distribution modeling**: the practice of fitting species distribution models with more than one observation model.
**Integrated population modeling (IPM)**: the practice of simultaneously modeling population abundance and the demographic processes driving its variation, combining multiple sources of data into a single model (e.g., count or census-type data alongside mark–recapture and ring recoveries).
**Latent state**: an unobserved, and often practically unobservable, property of the modeled ecological system (e.g., the actual distribution or abundance of a species) that we are trying to estimate.
**Link function**: a function describing the relationship between the observations and the predicted mean of the latent state, to ensure that the predicted mean meets distributional criteria; for example, point counts are usually assumed to follow a Poisson or negative binomial distribution via a log link.
**Multispecies model**: a statistical model in which some parameters are shared among species, often by treating species-specific parameters as random effects.
**Observation model**: a statistical description of the data collection process. In a standard occupancy-detection model, the observation (sub)model characterizes the likelihood of detecting the species at a site where it is present.
**Occupancy-detection model**: a class of SDM where the data are collected so that they are informative about the detection

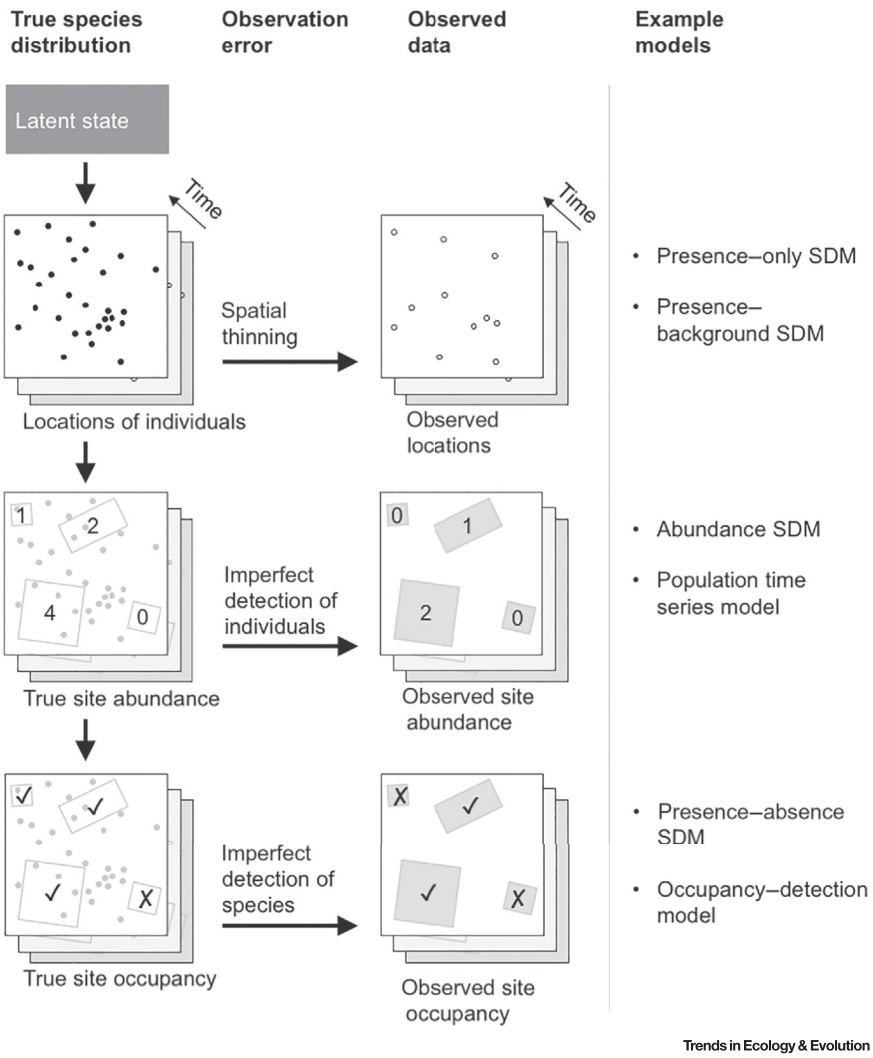


**Figure 2. Schematic Representation of How Different Types of Ecological Data and Models Are Interrelated.**

The panels in the left column demonstrate how data in multiple currencies emerge from a common set of ecological processes. The panels on the right illustrate the kinds of species data that are available to ecologists. Vertical arrows indicate how different ecological currencies are related to one another; horizontal arrows indicate types of observation processes by which data are an imperfect representation of the truth, although most real datasets contain multiple forms of observation error and bias. Abbreviation: SDM, species distribution model.

In this case, the observed pattern of points is a thinned-out version of the complete distribution of individuals, this is called a thinned point process (Figure 2, top row). Unfortunately, in most real datasets the thinning is not even and reflects biases in sampling effort, for example, we are more likely to record individuals that are nearer to major roads or field stations [27]. This sampling bias in geographic space is particularly a problem if it translates into a bias in environmental space. Ignoring it could lead to the erroneous conclusion that, for example, the habitat around roads are a species' preferred habitat. Spatial biases are particularly widespread and problematic in presence-only data [26,27], and these generally need to be accounted for in the observation model in order to estimate the true habitat preferences.

process (e.g., by several repeat surveys to at least some of the sampling sites). With such data, the model can separately estimate the probability of species' occurrence at a site, and the parameters driving the observation process, for example, the probability of detecting the species where present, or the probability that the species was present at a site where it was not observed.

**Point process model**: a statistical model that describes how points (e.g., individuals) are distributed in space. The stochastic process used for this description has a so-called intensity, in which points are more likely to be present at locations where the intensity is high. The most common implementation is a Poisson point process model, which assumes independence in the location of individuals, after accounting for the intensity.

**Presence–absence data**: records of whether a species is present or not at each of a number of sampling locations (e.g., quadrats or study sites). Detection does not have to be certain and its probability can be estimated with occupancy-detection models if there are multiple visits to a location. This term is often used for situations where detection/non-detection data would be more appropriate.

**Presence-only data**: records of the locations where a species was observed (e.g., from museum samples). These data lack information about where individuals were not observed, in contrast to presence–absence data.

**Species distribution model (SDM)**: generally refers to a statistical (correlative) model that relates environmental covariates to species' records over a geographic region. In practice, SDMs are often fitted to presence-only data, although are more robust when fitted to presence–absence or detection/non-detection data. SDMs include occupancy-detection models and abundance models.

**State-space model**: a model that combines a latent state with one or more observation models that describe how the data were generated from this latent state.

**Structured data**: data derived from a well-defined sampling

Structured **abundance data** require a sampling protocol to ensure that data are comparable across sites and time. The protocol defines the area within which individuals are observable (e.g., the radius around a point count or the length and width of a transect), and other aspects, such as how long is spent observing. The observed abundances in structured surveys can be modeled by assuming that they follow a Poisson distribution, which arises naturally from a Poisson point process model (Figure 2, middle row) as the number of individuals in an area follows a Poisson distribution whose mean is the integral of the intensity over that area (e.g., [50]). But observed abundances are often over dispersed (e.g., if there is variation at a finer scale than is being modeled), which can be modeled either with covariates or by assuming extra random variation, though, for example, a negative binomial distribution.

Some survey protocols record only whether or not the species was detected (e.g., recording from a checklist) within a defined site. These can be viewed as a degraded version of the abundance data; the data encodes whether there were zero or more than zero individuals observed. These are known as presence–absence data, although, acknowledging imperfect detection, it is more technically correct to refer to them as **detection/nondetection data**. Presence–absence data are typically modeled as a Bernoulli random variable (Figure 2, bottom row). Under the Poisson point process formulation, we consider that the underlying number of individuals follows a Poisson distribution and the observation model defines the probability of observing at least one individual via the complementary log-log **link function** [51,52].

## Discrete Alternatives

The vast majority of the large-scale ecological modeling literature is not based on using point processes, but using discrete representations of space, that is, in grid cells. This approach makes sense for data that are gathered in grids, where the survey protocol samples whole grid cells (e.g., most

protocol, such that observations are comparable in time and/or space (i.e., they can be described by a common observation model). Note that a structured survey protocol does not guarantee that the data are free from spatial bias; whilst some schemes select survey sites following a statistical sampling protocol (e.g., stratified random sampling; the UK Breeding Bird Survey), others do not (e.g., most butterfly monitoring schemes).

**Submodels**: components of a hierarchical state-space model. In the integrated models described here, there are separate observation submodels for each dataset and one state submodel for the latent state.

**Unstructured data**: data collected without formal protocol or sampling design, or where the protocols are unknown. Most unstructured data are in the form of presence-only data, for example, those arising when members of the public submit records of wildlife observations.

### Box 1. Case Studies of Integrated Models of Species Distributions

Dorazio [42] presents a (single species) model to combine presence-only records that suffer from sampling bias with abundance data from systematic surveys, using a point process model. The imperfect detection of individuals is estimated from the repeated point counts and mapped back to the point process by describing the actual abundance as a Poisson distribution with mean abundance defined by the point process intensity. The presence-only dataset is mapped back to the point process by modeling the sampling bias as a function of observation predictors, with the assumption that what influences observation bias does not influence environmental preferences (Figure IA).

Fithian *et al.* [43] model the distribution of 36 eucalyptus species in South-Eastern Australia. They combine biased presence-only records with presence–absence data from systematic surveys. They build a **multispecies model** linking both types of data through a common latent point process, assuming perfect detection in the presence–absence data. The presence-only dataset is mapped back to the point process by modeling sampling bias as a function of observation predictors so that all species are assumed to be exposed to the same pattern of bias. Sharing information across species improves the power to disentangle sampling bias from environmental preferences (Figure IB).

Pagel *et al.* [44] model spatiotemporal variation in abundance across the geographic range of a butterfly (*Pyronia tithonus*) in Great Britain. They combine structured abundance data (transect counts) with extensive unstructured presence–absence data (opportunistic species lists). In a hierarchical state-space model, both types of data are linked to a latent state variable representing abundance within grid cells. Detection in the presence–absence data is modeled as a function of abundance, made possible by the spatial overlap of the two datasets. This model enables information on abundance to be extracted from the proportion of species lists that report the species' presence in a given grid cell (Figure IC).

Guillera-Arroita *et al.* [45] model the occurrence of four Australian frog species based on environmental DNA (eDNA) surveys and aural survey. Their occupancy-detection model considers both false-negative and false-positive errors in the eDNA observation model. Detection parameters are not identifiable from the eDNA data alone, which is prone to false positives arising from sample contamination. To overcome these problems, the model incorporates calibration data for the eDNA analysis and presence–absence data from repeated aural surveys. Linking additional survey data to the occurrence state variable constraints alternative parameter solutions and thereby enables the estimation of false-positive detection probabilities of the eDNA surveys (Figure ID).

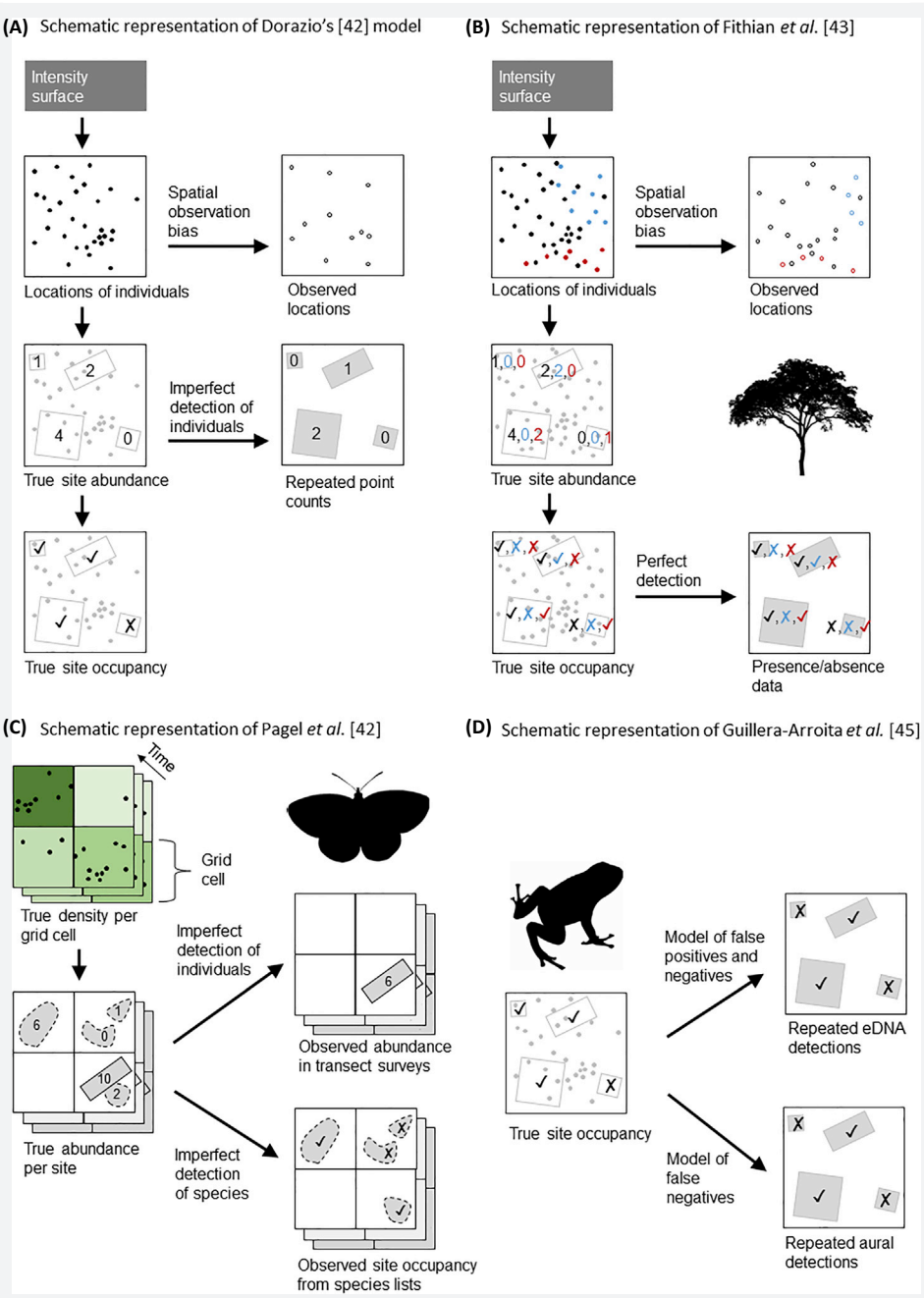


**Figure I. Integrated Models of Species Distributions.**

(A) Schematic representation of Dorazio's [42] model.
(B) Schematic representation of the model in Fithian et al. [44].
(C) Schematic representation of the model in Pagel et al. [43].
(D) Schematic representation of the model in Guillera-Arroita et al. [45].
Abbreviation: eDNA, environmental DNA.

European breeding bird schemes). Discrete space is also an attractive option when it is easy to conceptualize the observations as a direct realization of the latent state, for example, the latent state in a typical occupancy-detection model is the presence or absence of species within the whole site (or grid cells), and the observations (detected or not) are conditional upon the site being occupied [6]. However, constructing models based on discrete space is challenging for integrated models with data types in different currencies, especially if datasets differ in spatial coverage within the grid cell.

Another characteristic of the grid-based paradigm is that the resolution (i.e., grid size) is fixed, either constrained by the data or determined by some *a priori* choice or assumption about the most appropriate scale of investigation. This makes the inference scale-specific, since both occupancy and abundance scale nonlinearly with grid cell size, reflecting the fact that individuals are clumped in space [53,54]. Whilst methods do exist to translate across scales [55,56] and address within-grid heterogeneity [57,58], these are not always trivial to implement and can have limited predictive power. Scale problems become more difficult to manage when attempting to integrate datasets collected at different spatial resolutions, or from different locations within a grid cell [43]. This is a special case of a problem known in spatial statistics as 'change-of-support', for which promising solutions have recently been developed in the context of integrated distribution models [47].

Point processes provide a natural way of dealing with scale dependence, at least in the observed data, because the intensity of the point process (the latent variable) can vary in continuous space, which is natural to assume for ecological data. It is therefore relatively straightforward to make inferences at multiple scales from a single model [29]. For small regions, such as a few hectares within the range of a continentally distributed species, we can assume a constant (i.e., point) intensity, and the expected number of individuals within this region is the intensity multiplied by the area. For larger areas, the expected number of individuals is derived by integration of the intensity over the region (or a numerical approximation thereof [59]).

Note that the distinction between grid-based and point process models is not always obvious, in particular when environmental covariates are available on a spatial grid; for example, MaxEnt is presented to the user as a model in discrete space, although it is mathematically equivalent to a point process model under some circumstances [60].

## Implementation

For grids, the development and fitting of distribution models is a mature process; each grid cell can be considered on its own, and the main complication is usually whether spatial autocorrelation should be accounted for [61]. Integrating multiple datasets becomes complicated when issues of scale have to be considered, in which case including spatial autocorrelation may be essential to share information across datasets [29,46]. In practice, many implementations have been developed in a Bayesian framework, because this simplifies the integration of data and the handling of uncertainty. Bespoke models can be developed using the BUGS language and its derivatives, which are flexible enough to allow inclusion of the latent state in several likelihoods. Thus, Markov chain Monte Carlo (MCMC) software, such as WinBUGS [62] and JAGS [63], can be useful to fit these models, but they can be slow to sample and converge, especially for large and complex models in which parameters are strongly correlated. Newer MCMC software, such as Stan [64] and greta [65], are potentially much quicker, but few comparisons for large ecological datasets exist. Frequentist implementations are also possible for specific model formulations, as in IPM [37].

Several methods exist to fit point processes to data [66]. With multiple data types, Bayesian methods once again appear more attractive, but again the computational cost may be high. Recent developments in computational statistics have improved the situation; for instance, accurate numerical approximations of the posterior distribution have been developed [59,67]. These approximations have been combined with efficient methods for approximating the continuous space in the INLA software [67,68], which makes it possible to fit many complex ecological models efficiently using point processes [59], although this efficiency comes at a cost of less flexibility in what models can be fitted

compared with MCMC approaches. In addition, whilst it can be challenging to specify state-space models in INLA, we show in Box 2 how this can be done.

## Challenges and Opportunities

Development of data integration for large-scale ecological models has recently advanced both conceptually and practically. There is now an emerging literature of large-scale ecological models using data integration. Box 1 describes how four recent examples, all of which integrate across two datasets with different observation models, fit within the general framework described previously. As the availability of new data types grows, data integration will be an option for the majority of large-scale ecological modeling applications, so we anticipate that it will become routine for ecologists working at large spatial and temporal scales. The real power of data integration, however, becomes apparent when linking datasets where the observation models are too dissimilar to permit simple data pooling without losing substantial information (Figure 1), for example, using expert-drawn range maps [69,70].

Data integration increases the quantity of available data and makes it possible to translate across ecological currencies (e.g., using occurrence records to make inferences about abundance [40]). It also opens up opportunities to expand the scope of the investigation; for example, integrating structured datasets with unstructured presence-only data, such as those on GBIF, may allow for an estimation of species distributions beyond the extent of the structured dataset. Similarly, the temporal extent can be expanded by adding museum specimens or paleontological data (e.g., from lake cores) to contemporary observations. However, modeling across different extents raises questions about compatibility, for example, does integrating a small but highly structured dataset from Wales with a large unstructured dataset spanning Europe help us understand what is happening in Belgium?

### Box 2. Case Study of Data Integration

Integrated distribution models are difficult to specify and challenging to fit. Making them more accessible requires tools that are general and flexible enough to fit different observation model types. Here, we demonstrate data integration in a point process model.

Our case study is the black-throated blue warbler (*Setophaga caerulescens*), which Miller *et al*. [29] used to compare data integration approaches using BUGS. We demonstrate the joint-likelihood approach using INLA [67,68], which is well suited to model spatial point processes and is more computationally efficient than BUGS.

The species' true distribution is modeled as an inhomogeneous point process, whose intensity varies as a function of elevation, canopy cover, and a random spatial field. We use three data sources, each with a different observation process (Figure IA): eBird records, North American Breeding Bird Survey (BBS) data, and the subset of the Pennsylvania Breeding Bird Atlas (BBA) used by Miller *et al*. (Figure IB). We treat eBird data as presence-only records, emerging as a thinned version of the intensity surface, with human population density as a covariate on the observation process (i.e., records are more likely where people live). We use a simplified version of the BBS data, treating the 50-point samples as a replicate presence–absence data per site (known as 'routes' in BBS). We treat the BBA data as presence–absence and assume that the sites for all datasets are small enough to be represented as points. Each dataset has its own likelihood, with the true distribution as the (common) latent state.

The fitted model (Figure IC) integrates three datasets with different properties to produce a single distribution map accounting for variation and bias in sampling effort. A more complete analysis would need to verify that the species' distribution is reflected in covariates, rather than the random spatial field (which reflects unmodeled spatial autocorrelation). The observation model for eBird data could be refined to properly reflect the observation bias that generated it, for example, using an additional spatial field [44].

The core challenge of integrated distribution modeling is to ensure that the covariates and observation datasets are properly aligned. Functions to manipulate the data into a common format, and to fit the model in INLA, are available online [79]. Whilst preliminary, the model illustrates the potential of this framework. The code provided should be readily extensible to include other data types, and transferable to other systems.

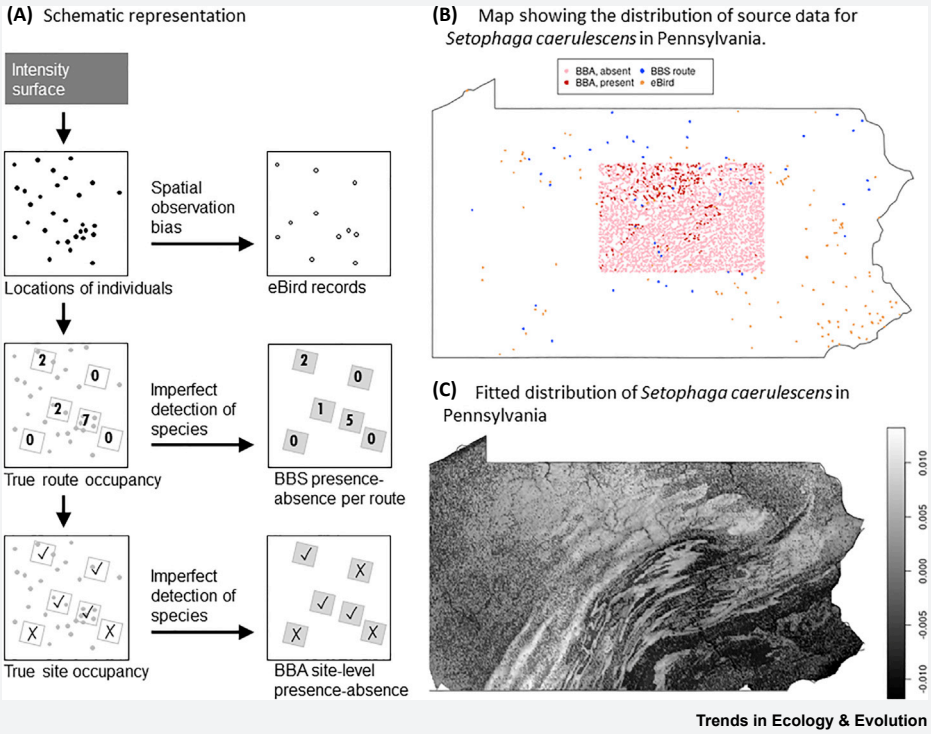


**Figure I. Integrated Model for Setophaga caerulescens in Pennsylvania.**

(A) Schematic representation of the integrated model (see Figure 2 in main text). (B) Map showing the distribution of source data. (C) Fitted distribution, derived from the integrated model in (A). The scale bar indicates the log of the intensity of the point process, such that light shades indicate a higher probability that the species is present. Abbreviations: BBA, Pennsylvania Breeding Bird Atlas; BBS, North American Breeding Bird Survey.

Although the ecological processes in Wales and Belgium may be different [71], the hope is that any differences will be captured by the dataset common to both countries. Exploring validity of these assumptions would be amenable to simulation studies, and these should be a priority for future research (see Outstanding Questions).

Prior to data integration, one must query the value of combining data. The choice between structured and unstructured data is typically seen as one between quality and quantity [22,26,72–74]; structured datasets are high quality but rare, whereas unstructured data are now plentiful but may contain biases. Most biodiversity data are in the form of unstructured records, so although using these data alone is attractive, the potential for bias means we may be on thin ice regarding the validity of predictions obtained from them [75,76]. Integrating structured with unstructured data has been shown to produce model parameter estimates that are both precise (on account of large sample sizes) and accurate (on account of the unbiased sample in the structured data) [42,44]. In other words, integrated models tend to inherit the best properties of the constituent datasets, not the worst. However, it is not clear whether this situation will be universally realized, for example, if bias in the unstructured data is small, there may be little to be gained (in terms of parameter estimates) by the addition of structured data, such that the complexity of an integrated model might not be justified. These issues become more challenging when all available datasets contain at least one form of bias. Can we adequately characterize each? It is conceivable that the act of integrating datasets could introduce new biases not found in the constituent datasets, such that the integrated model is 'worse' than either

## Outstanding Questions

When should complex integrated models be preferred over simple ones? Data integration has many advantages but is costly (e.g., in computational intensity). Do parameter-rich integrated models improve our ecological understanding? We need clear guidelines on when data integration is better than data pooling or discarding one dataset.

How do we quantify information gained by data integration? Information criteria (e.g., AIC, BIC) are not comparable between models with different sets of observations, so it's not obvious how to measure the added value of integrated modeling.

Under what circumstances does the joint-likelihood approach break down? This approach is conceptually appealing but performs poorly under certain conditions [29,30]. Understanding this trade-off should be a priority.

Can we be confident about working with biased data? How much structured data is enough to overcome bias in unstructured data? Can we detect biases if we don't know about them *a priori*? How should we evaluate whether biases have been adequately modeled? Is it possible to fit integrated models in which all datasets contain at least one form of bias?

How should we validate integrated distribution models? Model fit and prediction error are likely to be influenced by data quantity, quality, and the degree to which we capture biases. Moreover, the notion of 'fit' is not straightforward when there are multiple sets of observations, some of which contain known biases. Independent validation (e.g., test and training datasets) is at odds with the principle of using all available data, which has been an important motivation for integrated modeling. Guidelines for validation have been couched in terms of whether the contributing datasets should be

data pooling or proceeding with a single dataset. These new biases may be hard to detect, adding black ice to the thin ice of working with biased data [75,77]. Thus, it might be difficult to predict *a priori* whether data integration will be desirable for any particular application. Careful simulation studies might help with the development of some principles for when data integration is worthwhile.

considered equals, but this is not always obvious [29]. Several approaches to validation have been proposed (e.g., [40]) so there is a clear need to understand which of these works, under what circumstances, and how to apply them objectively.

## Concluding Remarks

The digital revolution has given us access to a growing volume of data about species, from a wide range of sources. We are making strides in developing methods to use these data, and are moving from ad hoc solutions for individual problems towards fully fleshing out a framework, based on point processes, to tackle a wide range of problems. These statistical advances need to be made readily available to ecologists, through the development of flexible and easy to use software. At the same time, conceptual and practical issues need to be addressed, such as exploring when data integration is worthwhile and how far we can go to combine different datasets. The potential that can be unleashed by solving these issues is huge; the biodiversity crisis is global [78], but so is the collection of biodiversity data, through the work of many local actors [17,18]. It is only by bringing together all of this effort that we can effectively use these disparate data in a coherent manner. The developments we have outlined should thus become the norm rather than the exception when investigating biodiversity over large spatial and temporal scales.

## Acknowledgments

This work was supported by Natural Environment Research Council grant NE/R005133/1 and award NE/R016429/1 as part of the UK-SCAPE programme delivering National Capability. We are grateful to Richard Chandler for stimulating and challenging conversations, and to Colin Beale, Diana Bowler, Corey Merow, Matt Farr, Elise Zipkin, and one anonymous reviewer for constructive feedback. G.G-A. and N.G. are supported by Discovery Early Career Researcher Awards from the Australian Research Council (DE160100904 and DE180100635).

## References

1. Kerr, J.T. *et al.* (2007) The macroecological contribution to global change solutions. *Science* 316, 1581–1584
2. Woodcock, B.A. *et al.* (2016) Impacts of neonicotinoid use on long-term population changes in wild bees in England. *Nat. Commun.* 7, 12459
3. Barbet-Massin, M. and Jetz, W. (2015) The effect of range changes on the functional turnover, structure and diversity of bird assemblages under future climate scenarios. *Glob. Chang. Biol.* 21, 2917–2928
4. Araújo, M.B. and Guisan, A. (2006) Five (or so) challenges for species distribution modelling. *J. Biogeogr.* 33, 1677–1688
5. Dormann, C.F. (2007) Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Glob. Ecol. Biogeogr.* 16, 129–138
6. MacKenzie, D.I. *et al.* (2005) *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence*, 1st edn (Academic Press)
7. Guillera-Arroita, G. *et al.* (2015) Is my species distribution model fit for purpose? Matching data and models to applications. *Glob. Ecol. Biogeogr.* 24, 276–292
8. Santini, L. *et al.* (2018) Global drivers of population density in terrestrial vertebrates. *Glob. Ecol. Biogeogr.* 27, 968–979
9. Guillera-Arroita, G. (2016) Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. *Ecography* 40, 281–295
10. Hampton, S.E. *et al.* (2013) Big data and the future of ecology. *Front. Ecol. Environ.* 11, 156–162
11. Soranno, P.A. and Schimel, D.S. (2014) Macrosystems ecology: big data, big ecology. *Front. Ecol. Environ.* 12, 3
12. Laurance, W.F. *et al.* (2016) Big data, big opportunities. *Front. Ecol. Environ.* 14, 347
13. Gibb, R. *et al.* (2019) Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods Ecol. Evol.* 10, 169–185
14. Hays, G.C. *et al.* (2016) Key questions in marine megafauna movement ecology. *Trends Ecol. Evol.* 31, 463–475
15. Bálint, M. *et al.* (2018) Environmental DNA time series in ecology. *Trends Ecol. Evol.* 33, 945–957
16. August, T. *et al.* (2015) Emerging technologies for biological recording. *Biol. J. Linn. Soc.* 115, 731–749
17. Kissling, W.D. *et al.* (2018) Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. *Biol. Rev.* 93, 600–625
18. Jetz, W. *et al.* (2012) Integrating biodiversity distribution knowledge: toward a global map of life. *Trends Ecol. Evol.* 27, 151–159
19. Silvertown, J. (2009) A new dawn for citizen science. *Trends Ecol. Evol.* 24, 467–471
20. Amano, T. *et al.* (2016) Spatial gaps in global biodiversity information and the role of citizen science. *Bioscience* 66, 393–400
21. Elith, J. *et al.* (2011) A statistical explanation of MaxEnt for ecologists. *Divers. Distrib.* 17, 43–57
22. Kamp, J. *et al.* (2016) Unstructured citizen science data fail to detect long-term population declines of common birds in Denmark. *Divers. Distrib.* 22, 1024–1035

23. Proença, V. *et al.* (2017) Global biodiversity monitoring: from data sources to Essential Biodiversity Variables. *Biol. Conserv.* 213, 256–263
24. Peterson, A.T. and Soberón, J. (2018) Essential Biodiversity Variables are not global. *Biodivers. Conserv.* 27, 1277–1288
25. Kindsvater, H.K. *et al.* (2018) Overcoming the data crisis in biodiversity conservation. *Trends Ecol. Evol.* 33, 676–688
26. Isaac, N.J.B. and Pocock, M.J.O. (2015) Bias and information in biological records. *Biol. J. Linn. Soc.* 115, 522–531
27. Meyer, C. *et al.* (2015) Global priorities for an effective information basis of biodiversity distributions. *Nat. Commun.* 6, 8221
28. Boakes, E.H. *et al.* (2010) Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biol.* 8, 11
29. Miller, D.A.W. *et al.* (2019) The recent past and promising future for data integration methods to estimate species' distributions. *Methods Ecol. Evol.* 10, 22–37
30. Fletcher, R.J. *et al.* (2019) A practical guide for combining data to model species distributions. *Ecology* 100, e02710
31. Peng, C. *et al.* (2011) Integrating models with data in ecology and palaeoecology: advances towards a model-data fusion approach. *Ecol. Lett.* 14, 522–536
32. Lahoz, W. *et al.* (2010) Data assimilation and information. In *Data Assimilation*, W. Lahoz *et al.*, eds. (Springer), pp. 3–12
33. Huelsenbeck, J.P. *et al.* (1996) Combining data in phylogenetic analysis. *Trends Ecol. Evol.* 11, 152–158
34. Reichman, O.J. *et al.* (2011) Challenges and opportunities of open data in ecology. *Science* 331, 703–705
35. Michener, W.K. and Jones, M.B. (2012) Ecoinformatics: supporting ecology as a data-intensive science. *Trends Ecol. Evol.* 27, 85–93
36. Ogle, K. and Barber, J.J. (2008) Bayesian data–model integration in plant physiological and ecosystem ecology. In *Progress in Botany*, U. Lüttge *et al.*, eds. (Springer), pp. 281–311
37. Besbeas, P. *et al.* (2002) Integrating mark-recapture-recovery and census data to estimate animal abundance and demographic parameters. *Biometrics* 58, 540–547
38. Buckland, S.T. *et al.* (2004) State-space models for the dynamics of wild animal populations. *Ecol. Model.* 171, 157–175
39. Fitsum, A. *et al.* (2010) An assessment of integrated population models: bias, accuracy, and violation of the assumption of independence. *Ecology* 91, 7–14
40. Schaub, M. *et al.* (2007) Use of integrated modeling to enhance estimates of population dynamics obtained from limited data. *Conserv. Biol.* 21, 945–955
41. Zipkin, E.F. *et al.* (2019) Innovations in data integration for modeling populations. *Ecology* 100, e02713
42. Dorazio, R.M. (2014) Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Glob. Ecol. Biogeogr.* 23, 1472–1484
43. Pagel, J. *et al.* (2014) Quantifying range-wide variation in population trends from local abundance surveys and widespread opportunistic occurrence records. *Methods Ecol. Evol.* 5, 751–760
44. Fithian, W. *et al.* (2015) Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods Ecol. Evol.* 6, 424–438
45. Guillera-Arroita, G. *et al.* (2017) Dealing with false-positive and false-negative errors about species occurrence at multiple levels. *Methods Ecol. Evol.* 8, 1081–1091
46. Pacifici, K. *et al.* (2017) Integrating multiple data sources in species distribution modeling: a framework for data fusion. *Ecology* 98, 840–850
47. Pacifici, K. *et al.* (2019) Resolving misaligned spatial data with integrated species distribution models. *Ecology* 100, e02709
48. van Strien, A.J. *et al.* (2010) Site-occupancy models may offer new opportunities for dragonfly monitoring based on daily species lists. *Basic Appl. Ecol.* 11, 495–503
49. Wiegand, T. and Moloney, K.A. (2013) *Handbook of Spatial Point-Pattern Analysis in Ecology*, 1st edn (CRC Press)
50. Banerjee, S. *et al.* (2003) *Hierarchical Modeling and Analysis for Spatial Data*, 1st edn (Chapman & Hall/CRC)
51. Kéry, M. and Royle, J.A. (2015) *Applied Hierarchical Modeling in Ecology: Analysis of Distribution, Abundance and Species Richness in R and BUGS* (*Prelude and Static Models, Vol. 1*) (Academic Press)
52. Royle, J.A. and Dorazio, R.M. (2008) *Hierarchical Modeling and Inference in Ecology*, 1st edn (Academic Press)
53. Barwell, L.J. *et al.* (2014) Can coarse-grain patterns in insect atlas data predict local occupancy? *Divers. Distrib.* 20, 895–907
54. McGill, B.J. (2011) Linking biodiversity patterns by autocorrelated random sampling. *Am. J. Bot.* 98, 481–502
55. Keil, P. *et al.* (2013) Downscaling of species distribution models: a hierarchical approach. *Methods Ecol. Evol.* 4, 82–94
56. Azaele, S. *et al.* (2015) Towards a unified descriptive theory for spatial ecology: predicting biodiversity patterns across spatial scales. *Methods Ecol. Evol.* 6, 324–332
57. McInerny, G.J. and Purves, D.W. (2011) Fine-scale environmental variation in species distribution modelling: regression dilution, latent variables and neighbourly advice. *Methods Ecol. Evol.* 2, 248–257
58. Graham, L.J. *et al.* (2019) Incorporating fine-scale environmental heterogeneity into broad-extent models. *Methods Ecol. Evol.* 10, 767–778
59. Illian, J.B. *et al.* (2013) Fitting complex ecological point process models with integrated nested Laplace approximation. *Methods Ecol. Evol.* 4, 305–315
60. Renner, I.W. and Warton, D.I. (2013) Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics* 69, 274–281
61. Beale, C.M. *et al.* (2010) Regression analysis of spatial data. *Ecol. Lett.* 13, 246–264
62. Lunn, D. *et al.* (2000) WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Stat. Comput.* 10, 325–337
63. Plummer, M. (2003) JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, Vienna, Austria, 20–22 March 2003
64. Carpenter, B. *et al.* (2017) Stan: a probabilistic programming language. *J. Stat. Softw.* 76, 1–32
65. Golding, N. (2019) greta: simple and scalable statistical modelling in R. *J. Open Source Softw.* 4, 1601
66. Renner, I.W. *et al.* (2015) Point process models for presence-only analysis. *Methods Ecol. Evol.* 6, 366–379
67. Rue, H. *et al.* (2009) Approximate Bayesian inference for latent Gaussian models by using integrated

nested Laplace approximations. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 71, 319–392
68. Martins, T.G. *et al.* (2013) Bayesian computing with INLA: new features. *Comput. Stat. Data Anal.* 67, 68–83
69. Merow, C. *et al.* (2017) Integrating occurrence data and expert maps for improved species range predictions. *Glob. Ecol. Biogeogr.* 26, 243–258
70. Domisch, S. *et al.* (2016) Model-based integration of observed and expert-based information for assessing the geographic and environmental distribution of freshwater species. *Ecography* 39, 1078–1088
71. Powney, G.D. *et al.* (2014) Can trait-based analyses of changes in species distribution be transferred to new geographic areas? *Glob. Ecol. Biogeogr.* 23, 1009–1018
72. Bayraktarov, E. *et al.* (2019) Do big unstructured biodiversity data mean more knowledge? *Front. Ecol. Evol.* 6, 239
73. Kelling, S. *et al.* (2019) Using semistructured surveys to improve citizen science data for monitoring biodiversity. *BioScience* 69, 170–179
74. van Strien, A.J. *et al.* (2013) Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. *J. Appl. Ecol.* 50, 1450–1458
75. Kéry, M. (2011) Towards the modelling of true species distributions. *J. Biogeogr.* 38, 617–618
76. Isaac, N.J.B. *et al.* (2014) Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods Ecol. Evol.* 5, 1052–1060
77. Royle, J.A. (2006) Site occupancy models with heterogeneous detection probabilities. *Biometrics* 62, 97–102
78. Díaz, S. *et al.* (2019) *Summary for Policymakers of the Global Assessment Report on Biodiversity and Ecosystem Services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services* (IPBES)
79. Dambly, L. *et al.* (2019) *Integrated model of the black-throated blue warbler in Pennsylvania.* https://zenodo.org/record/3363936#.XUw_F-NKhFE