Received: 6 October 2022 | Accepted: 20 February 2023

DOI: 10.1111/2041-210X.14091

Methods in Ecology and Evolution

APPLICATION

PointedSDMs: An R package to help facilitate the construction of integrated species distribution models

Philip S. Mostert[1,2] | Robert B. O'Hara[1,2]

[1]Department of Mathematical sciences, Norwegian University of Science and Technology, Trondheim, Norway

[2]Centre for Biodiversity Dynamics, Norwegian University of Science and Technology, Trondheim, Norway

Correspondence
Philip S. Mostert
Email: philip.s.mostert@ntnu.no

Handling Editor: Luis Cayuela

Abstract

1. Ecological data are being collected at a large scale from a multitude of different sources, each with their own sampling protocols and assumptions. As a result, the integration of disparate datasets is a rapidly growing area in quantitative ecology, and is subsequently becoming a major asset in understanding the shifts and trends in species' distributions.
2. However, the tools and software available to construct statistical models to integrate these disparate datasets into a unified framework is lacking. This has made these methods inaccessible to general practitioners and has stagnated the growth of data integration in more applied settings.
3. We therefore present *PointedSDMs*: an easy to use *R* package used to construct integrated species distribution models. It provides functions to easily format the data, fit the models in a computationally efficient way and presents the output in a format that is convenient for additional work.
4. This paper illustrates the different uses and functions available in the package, which are designed to simplify the modelling of integrated models. A case study using the package is also presented: combining three datasets coming from different sampling protocols, all containing records of *Setophaga caerulescens* across Pennsylvania state.

KEYWORDS
count data, data integration, integrated nested Laplace approximation, integrated species distribution models, presence absence, presence only, R package

1 | INTRODUCTION

Ecological research in the 21st century has been characterized by the accumulation of species occurrence data, due mainly to the advancements in digital technology and online data repositories (LaDeau et al., 2017). While this accumulation of data has expanded the potentials of ecological analysis on the spread, range shifts and relationship species have with the underlying environment, a multitude of challenges have arisen. In particular, the data are likely to have come from disparate sources, resulting in heterogeneous attributes, assumptions and sampling protocols inherent in each (Fletcher Jr et al., 2019).

Typically, analysis of such data uses species distribution models (SDMs), which model the relationship between species' distributions and the underlying environment. They are fitted to data using a variety of different estimation procedures and software packages (examples of such provided in Norberg et al., 2019). However, when

several datasets were available, the standard approach was to favour one dataset and discard the others, or use them in some form of secondary analysis (Simmonds et al., 2020).

As a result, a myriad of statistical methods to perform analysis, make predictions and efficiently use all available data have been produced—the so-called integrated species distribution models (ISDMs; see Miller et al., 2019, for a detailed review). A common result among research on the topic is that integrated data models appear to not only expand the spatial scope of a study, but also appear to be superior to models with a single data source by providing improvements to the results and estimates in comparison to using only a single dataset (see e.g. Bowler et al., 2019; Fithian et al., 2015; Miller et al., 2019).

Despite the development of ISDMs, a significant problem is the lack of general software and tools to make inference with them; thus, the overall uptake has been generally slow. Here we introduce *PointedSDMs*, an easy to use *R* (R Core Team, 2022) package designed to fit SDMs using data obtained from heterogeneous sources, and integrate them all together in a unified statistical framework. It does so using a hierarchical state space formulation—in which we link a process model (which provides a description of the true distribution of the model) with observation models for each dataset, dependent on their underlying sampling protocols (Isaac et al., 2020).

The integrated model for this package is fitted using integrated nested Laplace approximation (INLA)—a computationally efficient method used by Bayesian statisticians to fit latent Gaussian models. The theory behind the INLA methodology is discussed in detail in Rue et al. (2009), and estimating models with this methodology is made simple with the now established *R-INLA* package (Martins et al., 2013). The *PointedSDMs* package constructs a wrapper around the *R* package *inlabru* (Bachl et al., 2019), which, in turn, builds on the *R-INLA* package to help provide a user-friendly method to simplify the modelling of spatial process models.

1.1 | Statistical model

The aim of our state-space point process model is to use the available species' location data to make inference about the 'true' distribution of the population of the species'; since this distribution cannot be directly observed, it is referred to as a latent state (Isaac et al., 2020). To do inference, we use a hierarchical modelling structure with an underlying process model which provides a statistical description of how points are distributed in space; the role of such is a reflection of how multiple data types emerge from the same system (Isaac et al., 2020). This process has a spatially varying intensity function (denoted here by $\lambda(s)$) which is some function of environmental covariates $\mathbf{X}$ and parameters $\boldsymbol{\phi}$ such that a higher intensity implies that the species is more abundant in a location. A visual representation of the hierarchical setup of this model is presented in Figure 1.

For this model, we assume that the underlying process model is a log-Gaussian Cox process (LGCP) with an intensity function given as $\lambda(s) = exp\{\eta(s)\}$, which describes the expected number of species at some location, s. The log of this intensity function is thus given as:

$$\eta(s) = \alpha + \sum_{u=1}^{k} \beta_u X_u(s) + \zeta(s), \quad (1)$$

where α is a dataset-specific intercept term, β_u is the coefficient associated with the uth environmental covariate and $\zeta(s)$ is a zero-mean spatially continuous Gaussian random field (GRF), included in the model to account for potential spatial autocorrelation and the effects of all the environmental covariates not included in the model. Therefore, the expected number of species' presences within a region Ω is given by the integral of the intensity function across the entire region:

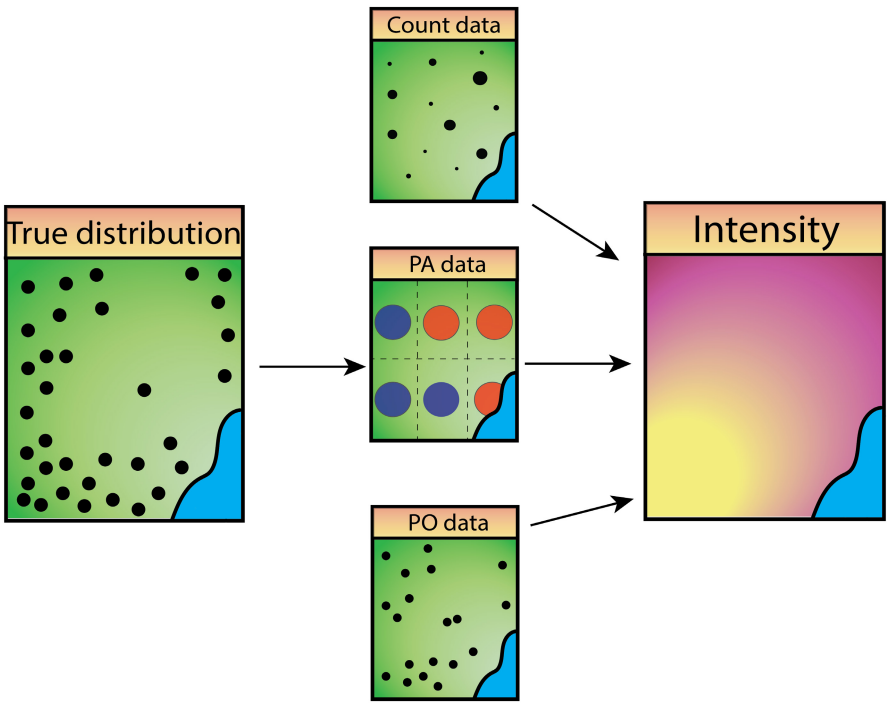

FIGURE 1 Representation of the structure of the integrated species distribution model, where each dataset is a separate realization of the 'true' species distribution. This is done by assuming each dataset has its own observation process, with a common latent, which is described by ecological covariates and parameters.

$$\mu(s) = \int_{\Omega} \lambda(s) ds. \quad (2)$$

Next, we assume that each dataset process $(Y_i, i = 1, 2, \ldots, n)$ has its own sub-model (observation model), which provide a statistical description on the data-collection process (Isaac et al., 2020). These models link the intensity function to the dataset's assumed likelihood, given by, $\mathcal{L}(Y_i \mid \lambda(s), \theta_i)$, where θ_i are the parameters for the i^{th} observation model. Table 1 provides a description of the three types of datasets allowed in *PointedSDMs*: presence-only (modelled as a thinned Poisson random variable), presence–absence (modelled as a Bernoulli random variable with a *cloglog* link function (see Kéry & Royle, 2016) and counts (modelled as a Poisson random variable) datasets.

Then, by combining the process model with the observation models, the full likelihood for the data processes $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_n\}$ is given by:

$$\mathcal{L}(\mathbf{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\phi}) \propto p(\lambda((s), \boldsymbol{X}, \phi) \cdot \prod_{i=1}^{n} \mathcal{L}(Y_i \mid \lambda(s), \theta_i), \quad (3)$$

that is, the model component for the latent state of the model, multiplied by the product of the individual likelihoods for the data processes.

In addition to the species location data, datasets sometimes include additional trait variables (often referred to as marks). These data may also be included in the point-process modelling framework to supplement the amount of information in the SDM through the joint-likelihood method described above, by treating each mark as its own observation model. That is, we assume within the datasets there are marks $(M_l, l = 1, 2, \ldots, p)$ with associated observation models $\mathcal{L}(M_l \mid \lambda(s), \theta_l)$, which results in the full likelihood:

$$\mathcal{L}(\boldsymbol{Y}, \boldsymbol{M} \mid \boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\phi}) \propto p(\lambda((s), \boldsymbol{X}, \phi) \cdot \prod_{i=1}^{n} \mathcal{L}(Y_i \mid \lambda(s), \theta_i) \cdot \prod_{l=1}^{p} \mathcal{L}(M_l \mid \lambda(s), \theta_l). \quad (4)$$

2 | PACKAGE FUNCTIONALITY

PointedSDMs was developed to streamline the modelling process and provide a general framework for integrated SDMs for ecologists who have a collection of heterogeneous datasets at hand. It does so by re-formatting and assigning appropriate metadata to the species' location and covariate data, and then constructing the relevant objects required by *R-INLA* (Martins et al., 2013) to do the model fitting. The package contains four primary functions for model pre-preparations (intModel), fitting and inference (fitISDM) and cross-validation (datasetOut and blockedCV), as well as several generic functions related to plotting, printing and predicting the results of the model.

intModel is the first function used in the integrated modelling process, and is built using *R*'s *R6* (Chang, 2021) object-orientated system. Here, the user adds the species location data, environmental covariates, as well as additional *R-INLA* and *sp* (Pebesma & Bivand, 2005) objects required; most of the other arguments for this function are used to define variable names and terms to be included in the model. Since this is an *R6* object, there are a handful of slot functions which allow further specification and adjustments of the components in the model. A description of each of these slot functions and their intended use is available in Table 2. *PointedSDMs* allows datasets from three sampling schemes: presence-only, presence–absence and count data, where the latter two are defined in the model through their response variable names, using the intModel's arguments *responsePA* and *responseCounts*, respectively.

If the user defines a spatial partitioning of their data points using intModel's slot function, '.\$spatialBlock', spatial cross-validation may be performed using the function, blockedCV: which iteratively calculates a cross-validation score by leaving a certain block of data out of the model based on their spatial location.

fitISDM is used for the modelling and estimation of the integrated model. The *data* argument of the function is an object created by the function intModel, which contains the necessary information and metadata required in the model. The second argument, *options* is used to control any additional *R-INLA* or *inlabru* options.

After the model has been estimated, another form of spatial cross-validation may be completed using the function, datasetOut. The function works by calculating a cross-validation score from the following steps:

1. Running a new model with one less dataset (from the main model)—resulting in a reduced model,
2. Predicting the intensity function at the locations of the left-out dataset with the reduced model,
3. Using the predicted values as an offset in a new model,

TABLE 1 Details on the observation models for the species location data which may be used in the *PointedSDMs* R package.

Dataset type	Statistical family	Link function	Dataset description
Presence-only	Thinned *Poisson*	*log*()	Typically opportunistically collected data with only information on the species presence available, treated as a thinned point-process to reflect sampling biases
Presence–absence	*Binomial*	*cloglog*()	Information on both the presence and absence of a species at a sampling location. Sometimes referred to as detection/non-detection data
Counts	*Poisson*	*log*()	The number of species located at each sampling location obtained through direct counts or some other index of abundance Sometimes referred to as abundance data

TABLE 2 The main slot functions available in intModel. A demonstration of the different functions in use is presented in the Setophaga vignette in the package.

Slot function	Description
.$plot()	Create a plot of the points classified by either dataset or species
.$addData()	Include species location data in the model
.$addBias()	Add an additional bias spatial field to a selected dataset
.$updateFormula()	Update the formula for selected observation models
.$changeComponents()	Add or remove specific inlabru components in the model
.$priorsFixed()	Specify priors for the fixed effects in the model
.$specifySpatial()	Specify arguments for the spatial field construction
.$spatialBlock()	Spatially block the data for cross-validation
.$addSamplers()	Add integration domain for the presence-only datasets

4. Finding the difference between the marginal-likelihood of the main model (i.e. the model with all the datasets considered) and the marginal-likelihood of the offset model.

Installation of the package may be done directly from *CRAN* servers using the following *R* script:

```
install.packages('PointedSDMs')
```

Any concerns and questions regarding the use of the package may be asked on issues board of the package's GitHub repository: https://github.com/PhilipMostert/PointedSDMs

3 | WORKED EXAMPLE

3.1 | Introduction

The example below illustrates the use of *PointedSDMs* in a worked example, using datasets from a variety of distinct sources. The datasets used include observations of the black-throated blue warbler *Setophaga caerulescens* (genus) from both structured and unstructured sampling schemes, obtained from various locations around Pennsylvania state (41°12′N, 77°11′W) on the eastern side of the United States of America (USA), which were collected between 2005 and 2009. Similar studies using these data were presented by Miller et al. (2019) and Isaac et al. (2020), who used the *WinBUGS* (Lunn et al., 2000) and the *R-INLA* package, respectively, to obtain results.

3.2 | Description of spatial covariates

Two standardized and continuous spatial covariates describing the study area were used in this analysis. The first, elevation, describes the height in metres above sea level, obtained from the package, *elevatr* (Hollister et al., 2021), and the second, canopy, describes the percentage of tree canopy covered in the area, obtained from the package, *FedData* (Bocinsky, 2022), which, in turn, accesses the data from the *National Land Cover Database* (NLCD). Both of these packages produced spatial covariates in the form of *Raster* objects, which we stacked into a single *RasterBrick* object before analysis.

3.3 | Description of datasets

The data used in this analysis come from three heterogeneous sources, where we assumed the underlying sampling protocol for each dataset is unique to that dataset (we considered datasets representing: presence-only, presence–absence and count data). These datasets and their unique sampling protocols are displayed graphically in Figure 2. We see that the three datasets combined have a better spatial representation of Pennsylvania compared to any dataset individually. However, each has a different sampling protocol, which implies an ISDM is appropriate to use for this example.

The citizen science presence-only data were obtained from *eBird* (Sullivan et al., 2009), a citizen science project launched by the *Cornell Lab of Ornithology* where amateur birders are able to submit checklists of avian detections to an online data repository, which has grown significantly since its inception, and has established itself as a significant tool in scientific research. Since the *eBird* data are collected by non-scientists, as so we expect the biases typically found in citizen science data. Given the nature of such data, we modelled these data as a thinned version of the intensity surface, with an additional spatial random field to account for biases in the collection process (Simmonds et al., 2020). Mathematically, this is given by:

$$\begin{aligned} Y_{\text{eBird}} &\sim \text{Poisson}\left(\omega(s)\cdot e^{\eta_{\text{eBird}}(s)}\right) \\ \eta(s) &= \alpha_{\text{eBird}} + \beta_{\text{elevation}} + \beta_{\text{canopy}} + \zeta_{\text{shared}}(s) + \zeta_{\text{bias}}(s), \end{aligned} \tag{5}$$

where $\omega(s)$ is the thinning parameter of the intensity function since we assume imperfect detection from these data (and as a result, estimate relative abundance instead of true abundance). This parameter cannot be directly estimated, and is therefore confounded within the intercept term of the model.

The other two datasets used come from structured survey data. The first comes from the *North American Breeding Bird Survey* (BBS; Pardieck et al., 2018), a long-term birding project designed to monitor changes in North American breeding bird populations for numerous species (Sauer et al., 2017). These data are collected alongside roadside survey routes composed of 50 independent stops 800 m apart from one another (Sauer et al., 2013). Surveys are conducted at each stops annually by an observer for a duration of 3 min, where avian species are identified by both sight and audition around a 400 m radius surrounding the stop. For their analysis, Isaac et al. (2020) treated the BBS data as a replicate presence–absence data per sight; however, for illustrative purposes we treat it as a count datasets, with a response variable denoting the number of species observed at each sight. Mathematically, this is given by:

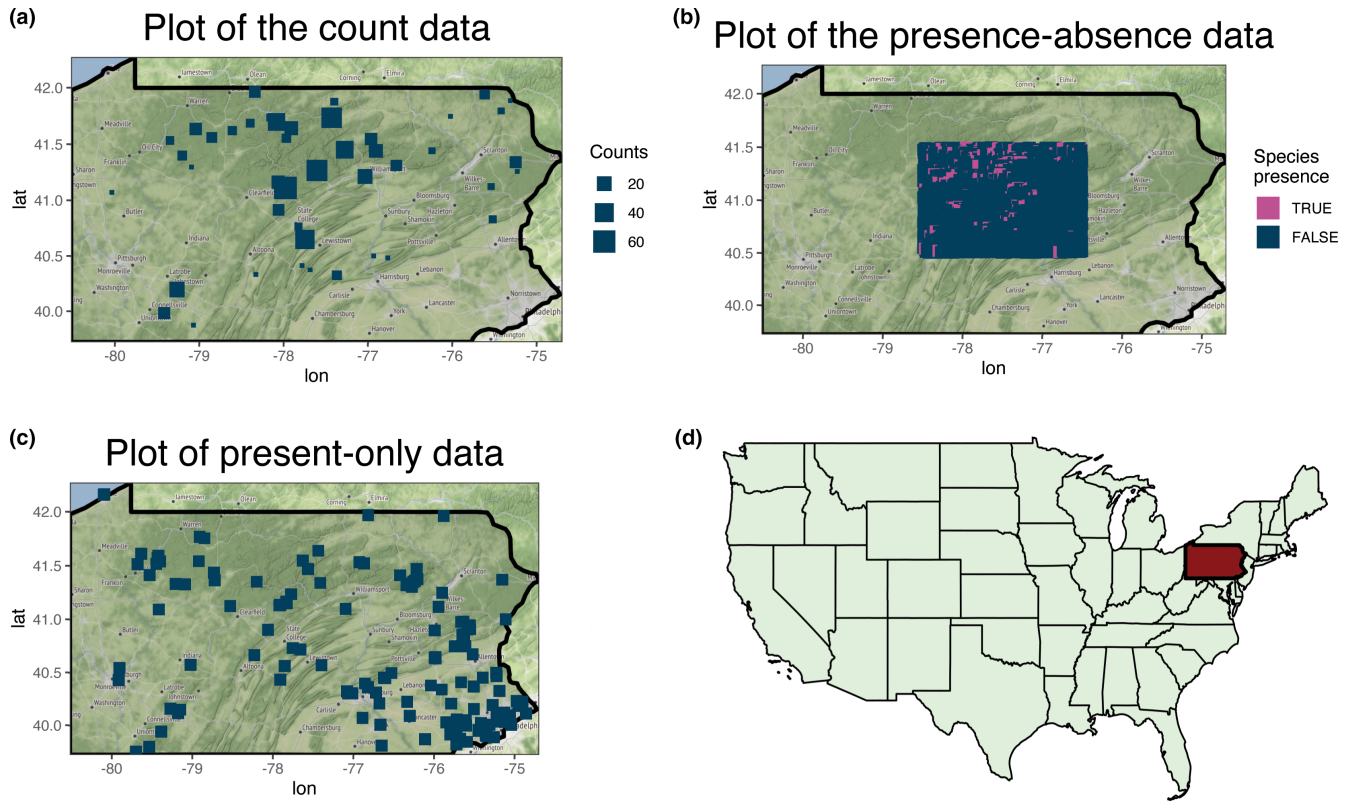

FIGURE 2 (a–c) Plots of the three datasets considered in the case study for species Setophaga caerulescens. (d) Map of the United States of America, highlighting Pennsylvania state.

$$\begin{aligned} Y_{BBS} &\sim \text{Poisson}\left(e^{\eta_{BBS}(s)}\right) \\ \eta_{BBS}(s) &= \alpha_{BBS} + \beta_{\text{elevation}} + \beta_{\text{canopy}} + \zeta_{\text{shared}}(s). \end{aligned} \tag{6}$$

The *Pennsylvania Breeding Bird Atlas* (BBA; obtained from Paton et al., 2019) is another long-term avian project following a standardized collection process (Wilson et al., 2012). These data comprise of more than 34,000 point counts (5165 points across the state of Pennsylvania) collected over a period of 5 years (2005–2009). Observers recorded observations of singing black-throated blue warbler around a 150 m radius for 5 equal intervals, each 75 s in duration. Following Isaac et al. (2020), we treat this data as detection/non-detection data (*1* indicating the presence of a species; *0* indicating the absence of a species) and assume that the sights are small enough to be represented as points. Mathematically, this is given by:

$$\begin{aligned} Y_{BBA} &\sim \text{Binomial}\left(p_i\right) \\ \text{cloglog}\left(p_i\right) &= \alpha_{BBA} + \beta_{\text{elevation}} + \beta_{\text{canopy}} + \zeta_{\text{shared}}(s), \end{aligned} \tag{7}$$

where p_i represents the probability of presence in location i.

3.4 | Model preparations

The first step to running an integrated model with *PointedSDMs* is to organize and assign appropriate metadata to the individual datasets, using the intModel function, which is used to initiate and prepare the statistical model before any inference is made; and so the arguments it takes are used to assign the relevant metadata to the datasets and covariates as well as set up all the objects required by *R-INLA*.

```
# SpatialPolygons object of PA state for the boundary for the mesh
PA <- USAboundaries::us_states(states = "Pennsylvania")
PA <- PA$geometry[1]
PA <- as(PA, "Spatial")

mesh <- INLA::inla.mesh.2d(boundary = inla.sp2segment(PA),
                           cutoff = 0.2,
                           max.edge = c(0.1, 0.24),
                           offset = c(0.1, 0.4))

proj <- sp::CRS("+proj=longlat +datum=WGS84
                +no_defs +ellps=WGS84 +towgs84=0,0,0")

# Stack covariates together into one Raster object
covariates <- scale(stack(elev_raster, NLCD_canopy_raster))
names(covariates) <- c('elevation', 'canopy')

spatial_data <- intModel(eBird_caerulescens, BBS, BBA,
                         Coordinates = c('X', 'Y'),
                         Projection = proj, Mesh = mesh,
                         responsePA = 'NPres', responseCounts = 'Counts',
                         spatialCovariates = covariates)
```

We would also like to account for spatial autocorrelation in the model through a GRF with a Matérn covariance function, which may be computationally expensive for large point process models. *R-INLA* counters this issue by approximating these GRFs via the stochastic partial differential equation (SPDE) approach (Lindgren et al., 2011), which requires the construction of a Delaunay triangulated mesh (interested readers who would like further details on the mesh construction are referred to Krainski et al., 2018; Lindgren & Rue, 2015). The mesh for this example was created with the inla.mesh.2d function by supplying a *SpatialPolygons* boundary of the study region as well as the *max.edge*, *offset*, and *cutoff* arguments. Furthermore, the SPDE models for this example were specified using penalizing complexity (PC) priors (Simpson et al., 2017), which are designed to control the spatial range and standard deviation in the GRF's Matérn covariance function to reduce over-fitting in the model.

```
spatial_data$specifySpatial(sharedSpatial = TRUE,
                            prior.sigma = c(5, 0.01),
                            prior.range = c(1, 0.01))
```

Simmonds et al. (2020) demonstrated in a simulation study that running a second spatial field for opportunistically collected presence-only data is a useful method to account for bias when knowledge of the sources of bias is scant or when covariates to adjust for bias are unavailable. Therefore, we use the `.\$addBias` function to add a second spatial field to our citizen science data to account for potential biases not reflected in the shared field.

```
spatial_data$addBias('eBird_caerulescens')
```

3.5 | Results

The integrated model is easily fit using the fitISDM function as below, which takes two arguments: *data* (which is an *intModel* object created above) and options (which is a list of *R-INLA* and *inlabru* options used to configure the model). In this model, the two fixed covariates and separate intercept terms for the three datasets were considered. In addition to the bias field for *eBird_caerulescens*, a common spatial field was used across the datasets; and to speed up computation time, *R-INLA*'s empirical Bayes' integration strategy was used.

```
spat_model <- fitISDM(data = spatial_data,
    options = list(control.inla = list(int.strategy='eb')))
```

PointedSDMs also includes the function datasetOut to carry out a form leave-one-out cross validation, which iteratively omits one dataset (and its associated marks) out of the full model. Table 3 illustrates the results of omitting one dataset out of the model at a time, where the mean change in fixed effects appears to vary significantly between the datasets.

```
data_out <- datasetOut(model = spat_model,
            dataset = c('BBS', 'BBA', 'eBird_caerulescens'),
            predictions = TRUE)
```

Setting predictions = TRUE allows the user to calculate a cross-validation score obtained by leaving out a dataset. In this case leaving out the BBS dataset causes the greatest difference in marginal likelihood between the main model and the reduced (without BBS) model, suggesting that this dataset provides the most information in our integrated model.

3.6 | Predictions

A crucial part of the process of making SDMs is creating prediction maps (such as those in Figure 3 to help researchers understand the species' spread. Predictions of the ISDMs from fitISDM are made easy using the predict function. The function will automatically create individual formulas to predict per dataset after the user has specified which components they would like to predict (with the arguments: *covariates*, *spatial* and *intercept*); and all components used

TABLE 3 Leave-one-out cross-validation score as well as the changes in fixed effects as a result of leaving a dataset out. The cross-validation score is calculated by finding the difference between the marginal likelihood of the full model, and the marginal likelihood of the model with the dataset left out.

Dataset left out	Δ Elevation	Δ Canopy	Cross-validation score
BBS	0.1974	0.1074	3.4660
BBA	−0.5537	−0.1786	2.7370
eBird	0.0951	0.0867	2.3187

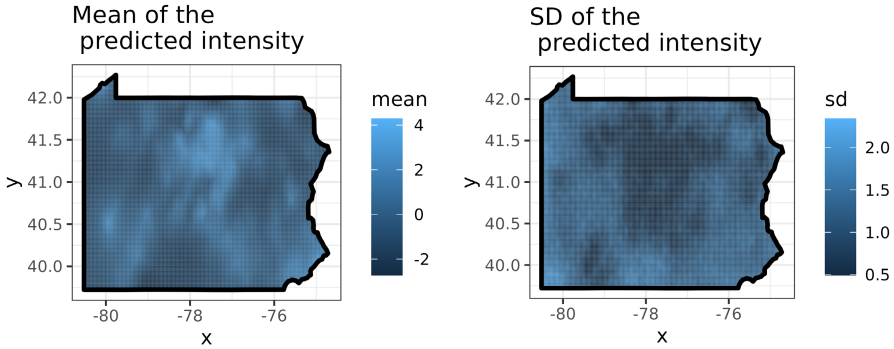

FIGURE 3 Mean and standard deviation of the predicted intensity ($\log(\lambda(s))$) of the integrated species distribution model, which gives a reflection of relative abundance across the spatial map.

in the model may be included by setting the *predictor* argument to TRUE. However, any formula may be predicted by using the function's *formula* argument.

```
projections <- predict(spat_model, mesh = mesh,
                       mask = PA,
                       predictor = TRUE,
                       fun = 'linear',
                       n.samples = 1000)
projection_means <- plot(projections,
                         plot = FALSE)
```

PointedSDMs also provides methods to plot basic predictive maps for a variety of statistics. By setting the *plot* argument to FALSE, the *ggplot* (Wickham, 2016) object of the predicted statistic is given, which would allow for more custom plotting functionality.

4 | CONCLUSIONS

PointedSDMs is an *R* package that provides the tools to make the most of the vast volume of species location data available today, by promoting and facilitating the integrated modelling of marked point process SDMs in a convenient way. It does so using the now well-established INLA methodology, and by constructing wrapper functions around the *R* package, *inlabru*.

4.1 | Opportunities for future work

A multitude of different R packages have been developed in the past to assist with the construction of SDMs (see: *dismo* [Hijmans et al., 2022], *sdm* [Naimi & Aráujo, 2016], *biomod2* [Thuiller et al., 2023], *HMSC* [Tikhonov et al., 2020] to mention a few); however, none of them have methods to create ISDMs—thus providing a novelty of *PointedSDMs*. Despite this, there are still extensions to the *PointedSDMs* framework which should considered to extend the project further.

Different groups of species influence each other through a multitude of processes (such as predation and competition), thereby affecting each other's distribution across space and time. A method to account for these processes would be to add inter-species interactions between different species, therefore changing the model framework to a joint species distribution model (JSDM).

A limitation of this model is that it only incorporates a small subset of the types of data used in ecology (presence-only, presence–absence and count data). Therefore, there is an opportunity to incorporate other types of data (such as biomass and movement data) into this framework, which would thereby extend the possibilities of research within a project.

Furthermore, providing tools to assist users in adding more custom components into the model, for example being able to both add random effects and change their precision matrix, should be considered. Incorporating this into the package would allow *PointedSDMs* to be used in additional analyses, such as studying phylogenetics.

Finally, constructing the necessary tools and data pipelines to move species and environmental data from online repositories to create a complete workflow in a way that is not only reproducible, but also easy enough to use for ecologists and policymakers with minimal basic skills would allow a package like this to show off its full potential (a reflection of the steps to develop such a workflow is discussed in Mostert et al., 2022). Before this is completed, tools required to simplify the sharing and standardization of ecological data on a large scale need to be developed.

AUTHOR CONTRIBUTIONS

Philip S. Mostert wrote the script for the R package, *PointedSDMs*, provided the graphics and led the writing for the first draft of the manuscript. Robert B. O'Hara provided conceptualization of the project, supervision and reviewed the manuscript.

ACKNOWLEDGEMENTS

We want to thank Walter Jetz and Petr Keil for help and discussions early in this project. We would also like to thank Ben Moore for providing us with the koala dataset used in the *marked point process* vignette in the package.

CONFLICT OF INTEREST STATEMENT

We declare no conflicts of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.14091.

DATA AVAILABILITY STATEMENT

All data are freely available for the reader in the *R* package, *PointedSDMs*, which is available on the Comprehensive R Archive Network: https://cran.r-project.org/web/packages/PointedSDMs/index.html, the *R* code is also open source and available on GitHub: https://github.com/PhilipMostert/PointedSDMs. The version of the package used for the example in this manuscript (v1.2.0) is archived at https://doi.org/10.5281/zenodo.7688454 (Mostert, 2023).

ORCID

Philip S. Mostert https://orcid.org/0000-0001-8017-3435
Robert B. O'Hara https://orcid.org/0000-0001-9737-3724

REFERENCES

Bachl, F. E., Lindgren, F., Borchers, D. L., & Illian, J. B. (2019). Inlabru: An R package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, *10*(6), 760–766. https://doi.org/10.1111/2041-210X.13168

Bocinsky, R. K. (2022). FedData: Functions to automate downloading geospatial data available from several federated data sources. R package version 3.0.0.9000.

Bowler, D. E., Nilsen, E. B., Bischof, R., O'Hara, R. B., Yu, T. T., Oo, T., Aung, M., & Linnell, J. D. (2019). Integrating data from different survey types for population monitoring of an endangered species: The case of the eld's deer. *Scientific Reports*, *9*(1), 1–14. https://doi.org/10.1038/s41598-019-44075-9

Chang, W. (2021). R6: Encapsulated classes with reference semantics. R package version 2.5.1.

Fithian, W., Elith, J., Hastie, T., & Keith, D. A. (2015). Bias correction in species distribution models: Pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, *6*(4), 424–438. https://doi.org/10.1111/2041-210X.12242

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

Hijmans, R. J., Phillips, S., Leathwick, J., & Elith, J. (2022). dismo: Species Distribution Modeling. R package version 1.3.9.

Hollister, J., Shah, T., Robitaille, A. L., Beck, M. W., & Johnson, M. (2021). elevatr: Access elevation data from various APIs. R package version 0.4.2.

Isaac, N. J., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., Freeman, S. N., Golding, N., Guillera-Arroita, G., Henrys, P. A., Jarvis, S., Lahoz-Monfort, J., Pagel, J., Pescott, O. L., Schmucki, R., Simmonds, E. G., & O'Hara, R. (2020). Data integration for large-scale models of species distributions. *Trends in Ecology & Evolution*, *35*(1), 56–67. https://doi.org/10.1016/j.tree.2019.08.006

Kéry, M., & Royle, J. A. (2016). *Applied hierarchical modeling in ecology: Analysis of distribution, abundance and species richness in R and BUGS* (1st ed.). Academic Press.

Krainski, E., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., & Rue, H. (2018). *Advanced spatial modeling with stochastic partial differential equations using R and INLA* (1st ed.). Chapman and Hall/CRC.

LaDeau, S., Han, B., Rosi-Marshall, E., & Weathers, K. (2017). The next decade of big data in ecosystem science. *Ecosystems*, *20*(2), 274–283. https://doi.org/10.1007/s10021-016-0075-y

Lindgren, F., & Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, *63*, 1–25.

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fieldsand Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(4), 423–498. https://doi.org/10.1111/j.1467-9868.2011.00777.x

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*(4), 325–337. https://doi.org/10.1023/A:1008929526011

Martins, T. G., Simpson, D., Lindgren, F., & Rue, H. (2013). Bayesian comput-ing with INLA: New features. *Computational Statistics & Data Analysis*, *67*, 68–83. https://doi.org/10.1016/j.csda.2013.04.014

Miller, D. A., Pacifici, K., Sanderlin, J. S., & Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, *10*(1), 22–37. https://doi.org/10.1111/2041-210X.13110

Mostert, P. S. (2023). PhilipMostert/PointedSDMs (version 1.2.0) [computer software]. https://doi.org/10.5281/zenodo.7688454

Mostert, P. S., Bjørkås, R., Bruls, A. J., Koch, W., & Martin, E. C. (2022). intSDM: A reproducible framework for integrated species distribution models. *bioRxiv*. https://doi.org/10.1101/2022.09.15.507996

Naimi, B., & Aráujo, M. B. (2016). sdm: A reproducible and extensible r platform for species distribution modelling. *Ecography*, *39*(4), 368–375. https://doi.org/10.1111/ecog.01881

Norberg, A., Abrego, N., Blanchet, F. G., Adler, F. R., Anderson, B. J., Anttila, J., Aráujo, M. B., Dallas, T., Dunson, D., Elith, J., Foster, S. D., Fox, R., Franklin, J., Godsoe, W., Guisan, A., O'hara, B., Hill, N. A., Holt, R. D., Francis, H. C. K., … Ovaskainen, O. (2019). A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, *89*(3), e01370. https://doi.org/10.1002/ecm.1370

Pardieck, K. L., Ziolkowski, D. J., Lutmerding, M., & Hudson, M.-A. (2018). *North American breeding bird survey dataset 1966–2017, version 2017.0*. U.S. Geological Survey, Patuxent Wildlife Research Center. https://doi.org/10.5066/F76972V8

Paton, G. D., Shoffner, A. V., Wilson, A. M., & Gagné, S. A. (2019). Data from: The traits that predict the magnitude and spatial scale of forest bird responses to urbanization intensity. *Dryad Digital Repository*. https://doi.org/10.5061/dryad.t4g871v

Pebesma, E. J., & Bivand, R. S. (2005). Classes and methods for spatial data in R. *R News*, *5*(2), 9–13. https://doi.org/10.1007/978-1-4614-7618-4_2

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(2), 319–392. https://doi.org/10.1111/j.1467-9868.2008.00700.x

Sauer, J. R., Link, W. A., Fallon, J. E., Pardieck, K. L., & Ziolkowski, D. J. (2013). The north American breeding bird survey 1966–2011: Summary analysis and species accounts. *North American Fauna*, *79*, 1–32. https://doi.org/10.3996/nafa.79.0001

Sauer, J. R., Pardieck, K. L., Ziolkowski, D. J., Jr., Smith, A. C., Hudson, M.-A. R., Rodriguez, V., Berlanga, H., Niven, D. K., & Link, W. A. (2017). The first 50 years of the north american breeding bird survey. *The Condor: Ornithological Applications*, *119*(3), 576–593. https://doi.org/10.1650/CONDOR-17-83.1

Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J., & O'Hara, R. B. (2020). Is more data always better? A simulation study of benefits and limitations of integrated distribution models. *Ecography*, *43*(10), 1413–1422. https://doi.org/10.1111/ecog.05146

Simpson, D., Rue, H., Riebler, A., Martins, T. G., & Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, *32*(1), 1–28. https://doi.org/10.1214/16-STS576

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, *142*(10), 2282–2292. https://doi.org/10.1016/j.biocon.2009.05.006

Thuiller, W., Georges, D., Gueguen, M., Engler, R., Breiner, F., Lafourcade, B., & Patin, R. (2023). biomod2: Ensemble platform for species distribution modeling. R package version 4.2.-2.

Tikhonov, G., Opedal, Ø. H., Abrego, N., Lehikoinen, A., de Jonge, M. M., Oksanen, J., & Ovaskainen, O. (2020). Joint species distribution modelling with the r-package hmsc. *Methods in Ecology and Evolution*, *11*(3), 442–447. https://doi.org/10.1111/2041-210X.13345

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer-Verlag.

Wilson, A. M., Brauning, D. W., & Mulvihill, R. S. (2012). *Second atlas of breeding birds in Pennsylvania* (2nd ed.). Pennsylvania State University Press.

How to cite this article: Mostert, P. S., & O'Hara, R. B. (2023). PointedSDMs: An R package to help facilitate the construction of integrated species distribution models. *Methods in Ecology and Evolution*, *14*, 1200–1207. https://doi.org/10.1111/2041-210X.14091