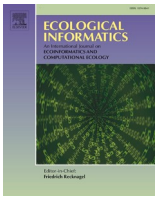
Ecological Informatics 76 (2023) 102127

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



An integrated species distribution modelling framework for heterogeneous biodiversity data

Martin Jung

Biodiversity, Ecology and Conservation research group, International Institute for Applied Systems Analysis (IIASA), Schlossplatz 1, A-2361 Laxenburg, Austria

ARTICLE INFO

Keywords:
Species distribution model
Data integration
Environmental niche
Offset
Point-process-model
Bayesian
R-package

ABSTRACT

Most knowledge about species and habitats is in-homogeneously distributed, with biases existing in space, time and taxonomic and functional knowledge. Yet, controversially the total amount of biodiversity data has never been greater. A key challenge is thus how to make effective use of the various sources of biodiversity data in an integrated manner. Particularly for widely used modelling approaches, such as species distribution models (SDMs), the need for integration is urgent, if spatial and temporal predictions are to be accurate enough in addressing global challenges.

Here, I present a modelling framework that brings together several ideas and methodological advances for creating integrated species distribution models (iSDM). The ibis.iSDM R-package is a set of modular convenience functions that allows the integration of different data sources, such as presence-only, community survey, expert ranges or species habitat preferences, in a single model or ensemble of models. Further it supports convenient parameter transformations and tuning, data preparation helpers and allows the creation of spatial-temporal projections and scenarios. Ecological constraints such as projection limits, dispersal, connectivity or adaptability can be added in a modular fashion thus helping to prevent unrealistic estimates of species distribution changes.

The ibis.iSDM R-package makes use of a series of methodological advances and is aimed to be a vehicle for creating more realistic and constrained spatial predictions. Besides providing convenience functions for a range of different statistical models as well as an increasing number of wrappers for mechanistic modules, ibis.iSDM also introduces several innovative concepts such as sequential or weighted integration, or thresholding by prediction uncertainty. The overall framework will be continued to be improved and further functionalities be added.

1. Introduction

Species distribution models (SDM) are the most widely used ecological modelling approaches when the aim is to infer, predict and project species assets (or other biodiversity features) in space and time (Elith and Leathwick, 2009). These models usually rely on statistical relationships between species occurrences and environmental covariates based on the niche concept (Araújo and Guisan, 2006; Blonder et al., 2014; Guisan and Thuiller, 2005). Measures and indicators derived from SDM outputs are for example commonly used to inform biodiversity survey efforts (Fois et al., 2018), identify areas of potential conservation value (Jung et al., 2021) or project the impact of changes in land-use, management intensity or climate (Leclère et al., 2020; Leitão et al., 2022; Santini et al., 2021). Nevertheless there are calls that inferences made by SDMs should be more critically interrogated in terms of the processes and responses they are able to capture (Evans et al., 2016; Hannemann et al., 2016; Lee-Yaw et al., 2022; Weber et al., 2017), especially since - as a data-driven method - SDMs are heavily dependent on the good availability and quality of data at adequate scales.

Accurate estimation of changes in biodiversity requires sufficient monitoring, which however can be financially and taxonomically (e.g. required expertise to survey a species) costly. Most biodiversity occurrence data are collected opportunistically, often by citizen scientists, and which has resulted in spatial, environmental and temporal biases (Hughes et al., 2021; Meyer et al., 2015). Modelling approaches such as SDMs usually reach better performance with well curated or systematically collected datasets as response functions stabilize and spurious correlations with some covariates are minimized (Hannemann et al., 2016; Smith and Santos, 2020). Yet, the reality is that complete or unbiased sampling coverage for any given species and data source is rarely,

E-mail address: jung@iiasa.ac.at.

https://doi.org/10.1016/j.ecoinf.2023.102127
Received 21 February 2023; Received in revised form 10 April 2023; Accepted 12 May 2023
Available online 1 June 2023

if ever achieved. Instead, scientists and landscape managers usually are left with multiple heterogeneous data sources, such as range maps, citizen-science data, structured surveys and checklists or species traits (Isaac et al., 2020; Jetz et al., 2019). This has subsequently lead to renewed calls for better data integration in biodiversity syntheses across scales (Heberling et al., 2021).

Species distribution models are particular sensitive to geographical or environmental biases in underlying biodiversity data (Baker et al., 2022; Botella et al., 2020). And although several methods have been developed to account to some extent for sampling biases (Chauvier et al., 2021; Warton et al., 2013), it can be argued that more information on the biology of a species is usually known (for example where a species broadly persist), that what is usually provided as input to an ecological model. Historically, SDM approaches have mostly relied on only single data sources (e.g. presence-only records from databases such as GBIF). New modelling approaches and frameworks have been developed to integrate different data sources into one combined prediction (Fletcher et al., 2019; Isaac et al., 2020; Miller et al., 2019). These so-called 'integrated' SDMs have the promise of providing in many cases more accurate, less biased representations of a species niche while also accounting for some of the biases that plague biodiversity datasets.

Integrated SDMs were originally proposed as a method to integrate presence-only and presence-absence information to account for biases in either (Koshkina et al., 2017). The promise of such an approach is that a "high quality" or multiple datasets combined with abundant, but often biased or faulty data, such as citizen science records, can improve overall parameter estimation by balancing opposing strengths (quantity against quality). Previous work has shown that integrating additional data can improve the precision of species trend estimates (Hertzog et al., 2021), account for biases in underlying biodiversity data (Fithian et al., 2015; Pacifici et al., 2019), help the prediction of species distributions (Koshkina et al., 2017; Merow et al., 2017; Peel et al., 2019) and modify response functions by accounting for prior knowledge of species-environment relationships (Hofner et al., 2011). And although there is some evidence that integrated SDMs do not necessarily always perform better than standard SDMs using a single data source (Ahmad Suhaimi et al., 2021; Simmonds et al., 2020), it is beyond doubt that the necessity for data or model-based integration will only increase in the SDM literature in the coming years.

Much of the development of integrated SDMs has been enabled by thinking of them as regression formulations. Assuming exclusively presence-only information about a species is available, a species distribution can be inferred through a Poisson process (Renner et al., 2015), which is statistically equivalent to the popular Maxent framework (Renner and Warton, 2013). A particular advantage of this modelling paradigm is that – rather than creating "pseudo-absence" points of a species as required for example by logistic regressions – modellers are able to estimate and project the distribution using randomized (or targeted) "background" samples that can be used to infer the relative intensity of occurrence (Guillera-Arroita et al., 2014; Warton and Shepherd, 2010), which comes with fewer assumptions about the true absence of a species, while being congruent to logistic regressions (Warton and Shepherd, 2010). Additionally, SDMs inferred from a Poisson process easily allow the integration of spatial-explicit priors through offsets (Merow et al., 2017, 2016), priors (Fletcher et al., 2019) or model-based bias controls through integration of other datasets or by forcing a certain value (such as maximum sampling bias) during the projection phase only (Fithian et al., 2015; Phillips et al., 2009; Warton et al., 2013). The paradigm of formulating a SDM as a regression formulation has furthermore facilitated the development of methods where properties of individual datasets (e.g. presence-only vs presence-absence) are taken explicitly into account. These types of model-based integration, theoretically based on joint likelihood estimation, are among the most elegant but also computationally demanding types of integrated SDMs currently in existence (Doser et al., 2021; Fithian et al., 2015; Isaac et al., 2020; Miller et al., 2019). Given these developments, there is a need for an adaptable SDM framework that easily allows to integrate the various types of biodiversity information that are out there.

At this point readers might wonder of the exact gap that yet another statistical SDM package is trying to fill, especially given the wealth of software already available to researchers (Sillero et al., 2023; Thuiller et al., 2009). Although new R-packages for joint inference using multiple likelihoods have become recently available (Doser et al., 2021; Mostert et al., 2022), they do not offer all the flexibility of integration outlined by Fletcher et al., such as for the ability to add offsets, priors or ensembles (Fletcher et al., 2019). In addition, there does not yet exist a software solution that situates a PPM modelling framework in the context of integrated modelling while also allowing for scenario projections with typical constraints such as dispersal (Seaborn et al., 2020). With the *ibis.iSDM* package (https://iiasa.github.io/ibis.iSDM/) I intend to fill this gap, providing a generic wrapper package to integrate various types of biodiversity information, and in a way that is modular and easily expandable with additional functionalities in the future. The package is presented here in terms of its design, structure and key functionality as well as through a series of different exemplary use cases for constructing integrated SDMs and scenarios. Less emphasis is given here to different parameters and supporting modules since those will be incrementally added, and in depth detailed on the help pages of the pages as well as the online website.

2. Modelling framework

2.1. Design philosophy

The Integrated model for BiodIversity distribution projectionS (or *ibis.iSDM*, https://iiasa.github.io/ibis.iSDM/) aims to provide a series of convenience functions for fitting integrated SDMs. It captures in functionality all the different types of integration, such as ensembles, offsets and covariates, priors or joint modelling, outlined by Fletcher et al. (2019), while also being specific to the biodiversity type to be estimated. For example presence-only biodiversity datasets added to a distribution object are estimated by default through an inhomogeneous Poisson point process model (PPM), which assumes that the true number of individuals $N(y)$ can be approximated as relative observation intensity λ integrated over an area A, e.g. $N(y) \approx Poisson\left(\int_A \lambda(s) ds\right)$. The intensity λ can be estimated as $log(\lambda_s) = \beta_0 + \beta_k x_s + \varepsilon_s$ based on thinned observations s, β being the 1 to k coefficients in the model including an intercept (β_0), x being the covariate values in given area and ε being the model error. Inferring environmental suitability through PPMs is usually preferable way if only presence-only data is available (Renner et al., 2015; Warton and Shepherd, 2010), although the ibis.iSDM package also supports the common practice of adding "pseudo-absence" points to datasets (Fig. 4).

Most code in the *ibis.iSDM* package is highly modular as the main functionalities have been created in an object-oriented way by making use of a object structure inspired by the tidyverse (Wickham, 2016), allowing to retain data and functions contained within each object to facilitate reuse through other functions (Fig. 1, SI Fig. 1). Not only does this facilitate cleaner coding overall, it also makes the code more modular with regards to adding datasets or integrating other methods. For example, the existing implementation allows to directly add two different dispersal simulators, KissMiG (Nobis and Normand, 2014) and MigClim (Engler et al., 2012) to constrain future projections (see also scenario section below).

A typical *ibis.iSDM* workflow begins with defining a modelling background (e.g. the area over which a SDM is to be created) to which biodiversity data or covariates can then be added (SI Fig. 1). It should be noted that preparation of input data is left to the users and can be easily achieved through a range of external packages (Sillero et al., 2023; Zizka et al., 2019). Additionally, any other information on biodiversity-relevant data, such as priors and offsets for habitat preferences or

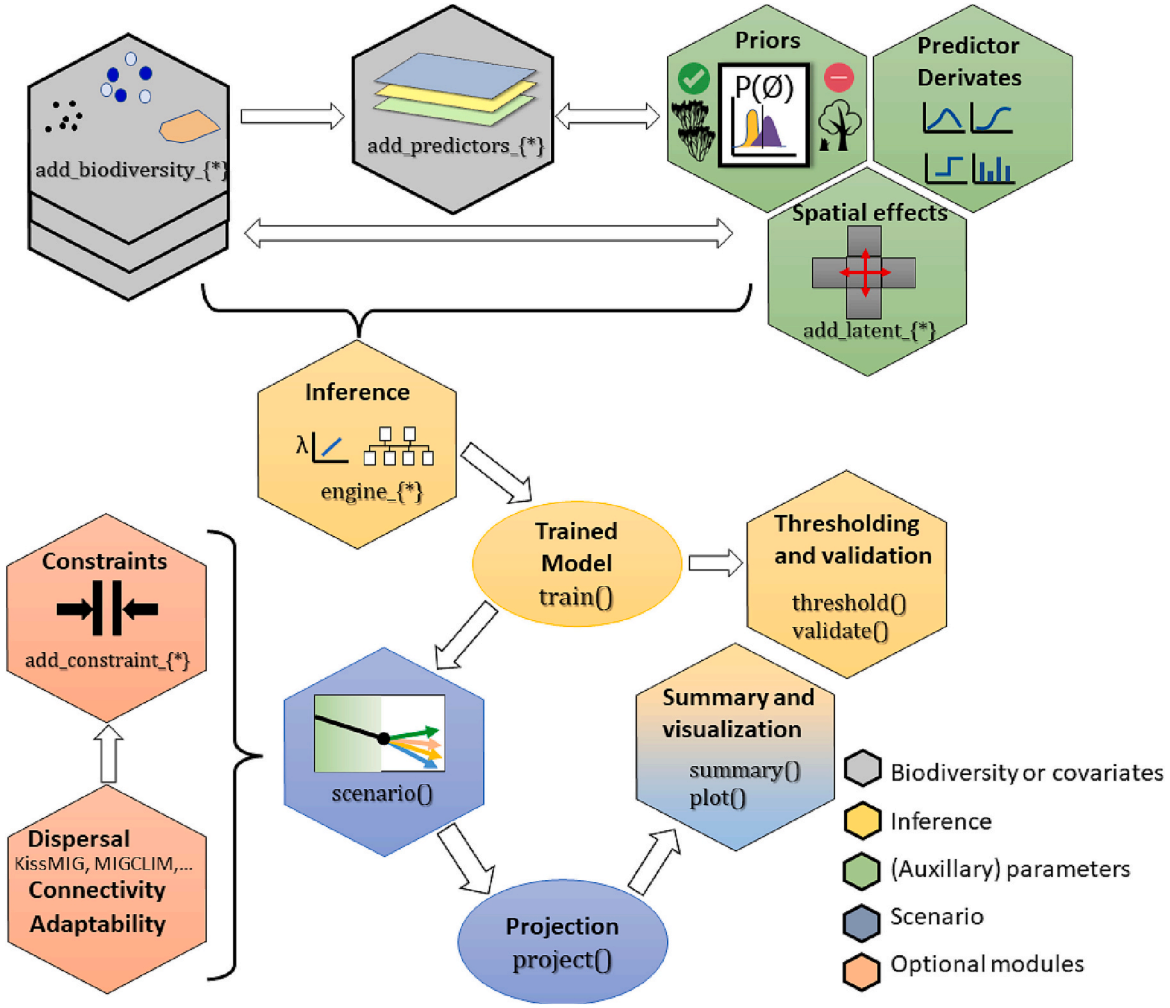


Fig. 1. Schematic and typical workflow of the *ibis.iSDM* package, where biodiversity and covariates datasets and combined with a series of auxiliary or optional modules. Through the use of different engines, response functions towards certain covariates and species distributions can be inferred. Each individual entry (hexagon) has its own function and stores internal data that can be accessed in a modular way. Many of the function have multiple variants (indicated by the {*}) allowing different data or parameter types to be added. A full list of all functions and examples can be found online (https://iiasa.github.io/ibis.iSDM/) and example code can be found in SI Fig. 1. Icons are created by the authors or are under public domain (CC-0).

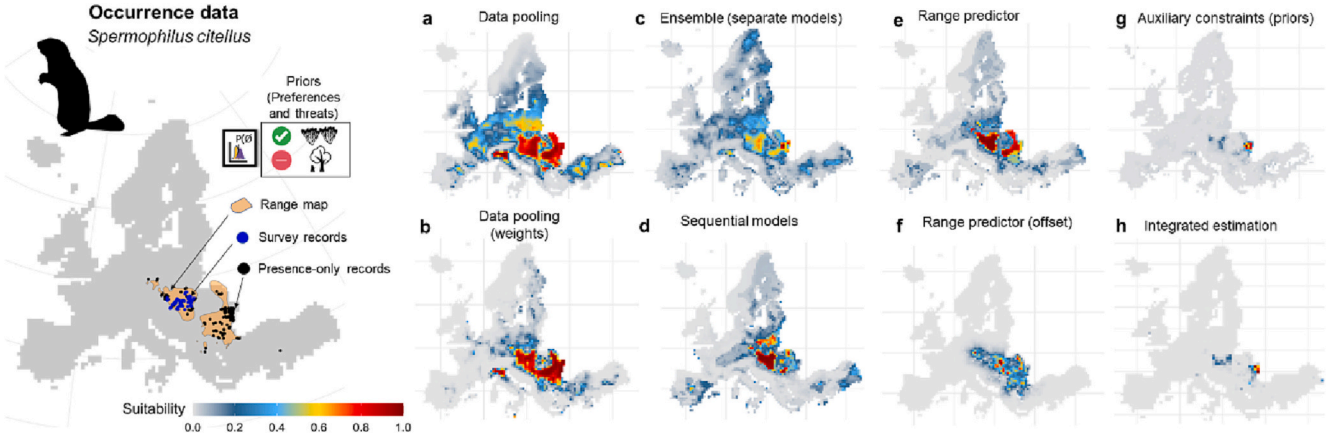


Fig. 2. The suitable habitat estimated with a SDM can vary depending on how different datasets are integrated as shown for the European ground squirrel (*Spermophillus citellus*). The available information for the species is combined either by a) data pooling, b) data pooling but with dataset specific weights, c) mean ensemble of different models, d) sequential estimation, e) inclusion of its range as predictor or f) as an offset, g) use of auxiliary climatic limits and priors or h) integrated estimation through joint likelihoods. All code and data with covariates to recreate the figures can be found in the supplementary materials.

known areas of occurrence, can also be added to the same object (SI Fig. 1). Finally, after specifying an engine and training the model, the resulting fit can then be visually interrogated, summarized and validated (Fig. 1) or passed on to construct a 'scenario' with different (temporal) predictors. The sections below highlights the package functionalities in more depth and also include demonstrations with example code and data for each.

2.2. Integration

The *ibis.iSDM* package supports all types of integration outlined by Fletcher et al. (2019), some even in multiple different ways (Fig. 2). The decision on which type of integration is preferable is specific to the types of data available in a given modelling problem. The easiest form of integration is to simply combine all point datasets ("pooling") and the package supports pooling with and without weights (Fig. 2a-b), the latter can for example give higher weight to potentially fewer, but more accurate records (Fig. 2b). Besides data pooling there is support for creating model ensembles ("ensemble(…)") for instance through means weighted by performance statistics (e.g., AUC) from independent data (Guisan and Thuiller, 2005; Valavi et al., 2021). Ensembles can also be constructed for model projections (e.g., scenarios up to 2050) as well as for response functions ("ensemble_partial(…)"). However often there are not enough data available to reliably fit every type of model, especially given the demanding nature of some machine learning approaches, and computation time can be a considerable limitation as well, such as for more demanding Bayesian models. The package will raise warnings and highlighted messages in case the provided information is not sufficient for inferring a species distribution.

Not always are there multiple point occurrence datasets available for a given species, although rarely are they the only information known about the biology of a species. In many cases expert information on habitat preferences, or a broad delineation of a species range can also provide contextual information about a species (Brooks et al., 2019; Merow et al., 2017). *Ibis.iSDM* supports as another type of integration the addition of expert delineated - or previous created model predictions - as covariates to model objects (Domisch et al., 2016), for example for species ranges ("add_predictor_range()") or elevational limits which transforms an elevational covariate into lower and upper bounded variables("add_predictor_elevationpref()"). Alternatively, such information could also be added through offsets that affect a regression fit and similar methods (e.g. "add_offset_range()" or "add_offset_elevation()") have been implemented in the package (Merow et al., 2017, 2016). Specific to each individual engine (defined as algorithmic approach for inference and projection, see below) there is also support for adding priors on the coefficients towards certain covariates via "add_priors(…)". Priors are usually specified either directly on the coefficients (magnitude and sign) or their direction, using for example monotonicity constraints (e.g. specifying that a certain variable have to be positive, Fig. 2g). Many priors can be particularly useful to avoid non-sensical response functions (Hofner et al., 2011), for example when owing to differences in grain a known forest-associated species the intended directional response towards this variable tends towards a particular trend.

Extending Fletcher et al., there are also options to use dataset specific weights or factor interactions to account for differences in included datasets (Leung et al., 2019). All these types of integration are also supported for inference on single datasets or can be used in sequential estimation. For example a potential use case easily enabled by *ibis.iSDM* could be to first fit a model using one biodiversity data source and a specific set of covariates such as broad climatic data, and then use the output of the resulting prediction as an offset to estimate the distribution with a different biodiversity or covariate data. Lastly, integration is also possibly through a dedicated model that combines multiple presence-only and presence-absence datasets together through a joint likelihood in a Bayesian setting (Fithian et al., 2015; Fletcher et al., 2019; Koshkina et al., 2017). These models are usually the most computationally intensive, but also the most elegant as all integration is done through dataset specific likelihoods (Fig. 2h).

2.3. Different engines

The backbone of any SDM modelling are the algorithm used for inference which in ibis.iSDM are called *"engines"*. To this date *ibis.iSDM* supports a total of 7 different engines for inferring or projecting the relative habitat suitability of biodiversity features. Those can broadly be classified into engines using either regressions and or non-parametric machine learning approaches and being frequentist or Bayesian in nature. Engines supported are regularized elastic net regressions through the *glmnet* package as also used by the *maxnet* package (Friedman et al., 2010; Phillips et al., 2017), Bayesian regularized "Spike-and-Slab" regressions with the *BoomSpikeSlab* package (Scott, 2022), Bayesian additive regression trees through *dbarts* (Carlson, 2020; Dorie, 2022), monotonic gradient descent boosting via *mboost* (Hofner et al., 2011; Hothorn et al., 2022), Extreme Gradient Boosting through *xgboost* (Chen et al., 2023), Bayesian spatial regressions with *INLA* and *inlabru* (Bachl et al., 2019; Lindgren and Rue, 2015) and general Bayesian regressions with *stan* (Gabry and Češnovar, 2022; Stan Development Team, 2022). The glmnet, stan and Bayesian regularized regressions only support linear response functions, while the other engines can also make use of non-linear estimation.

Although some engines support only linear response functions, non-linearity can be introduced through specific transformations of covariates such as hinge, threshold, quadratic or product derivates, as done in the popular maxent/maxnet modelling approach (Merow et al., 2013; Phillips et al., 2017). Functionalities to create such derivates are readily available when adding covariates to a distribution model (see SI Fig. 1 and code examples in the supplementary materials). Each of the different engines support different types of integration, with some engines being more flexible than others. For example, priors on coefficients can in some cases only constrain the directionality of response functions (Hofner et al., 2011), and in other cases also the magnitude of expected changes in relation to environmental covariates. An comparative overview of the capacities of each engine can be found online (https://iiasa.github.io/ibis.iSDM/l).

2.4. Model evaluation

Model evaluation through independent or withhold data is a critical part of the construction of species distribution models (Elith and Leathwick, 2009; Valavi et al., 2021). SDMs can be 'validated' in both a discrete and continuous way, with the former having been criticized for being dependent on thresholds applied to predictions of suitable habitat (Lawson et al., 2014; Liu et al., 2013). The *ibis.iSDM* package supports both continuous and discrete validation methods via the "validate()" function. Continuous validations use error metrics (e.g. RMSE) to infer prediction precision (Jung, 2022), while discrete validations can be calculated on a-priori mapped thresholded distributions with a range of different options from binary to normalized estimation (Fig. 4c). The identification of best thresholds for discrete validation can be achieved through heuristic searches for local optima in prediction performance measures (Márcia Barbosa et al., 2013). Estimated distributions can thus be validated ("validate()", SI Fig. 1) with independent or withheld data in a wide range of settings. The ibis.iSDM package does not yet support standard approaches such as spatial or spatial-temporal cross splitting (using for example the blockCV package, (Roberts et al., 2017)) directly in the modelling framework, and users should consider this aspect separately in their individual cases as part of the data preparation.

Lastly it should be highlighted that many commonly applied validation approaches are not necessarily appropriate when several different sources of information exist and best practices in the validation of integrated SDMs are still an open research topic as also highlighted by

Isaac et al. (2020). This is since (a) the consideration of all available data is one of the main points of model-based integration, (b) appropriate validation metrics are less straight-forward than for single datasets as biases and sampling methods can differ, and (c) any validation dataset might not represent the niche and environmental parameters estimated by the integrated model. For example, the standard practice of withholding parts of the training data for validating a model often means that both training and testing data suffer from the same spatial and environmental biases (Baker et al., 2022). If, however prior knowledge of the biology of a species is integrated in a SDM through a prior or offset, thus "nudging" or constraining response functions towards a more sensible outcome and ultimately different prediction, the use of any (biased) withheld data would likely indicate a reduced predictive performance compared to a model without such priors. One idea could be to validate SDMs not only based on their spatial predictions, but also on the magnitude and direction of their response functions (Smith and Santos, 2020). Certainly, more conceptual work is needed to design appropriate validation schemes for integrated SDMs.

2.5. Fitting and constraining projections in space and time

One of the objectives of species distribution modelling is to project the likely distribution or suitable habitat of a species into presence, past and future. In the simplest case SDM projections are usually made by multiplying the coefficients obtained from a previously fitted model with a matrix of (future) predictors (Elith et al., 2010; Thuiller et al., 2009). Such projections can be useful for making future projections and often show acceptable realism in independent assessments (Morán-Ordóñez et al., 2017; Soultan et al., 2022). Yet, such naïve projections assume that species are in equilibrium with their environment and often – but not always – neglect factors such as biotic interactions, adaptation and dispersal (Araújo and Guisan, 2006; Elith et al., 2010).

The *ibis.iSDM* package can project the distribution of biodiversity assets to different time periods, by supplying future covariates as multi-dimensional array using the "stars" R-package (Pebesma, 2022). Future projections can be defined via the "scenario(model)" function which requires a previously fitted *ibis.iSDM* model. After a scenario of projections has been created it can be summarized through a range of metrics (Fig. 3a). Similar as during the model inference, predictor transformations and thresholds can be flexibly added (see supplementary materials). After a scenario has been created, different summary methods and metrics of change can be obtained which are useful in model-based projections of biodiversity indicators (Leclère et al., 2020). As with other functions of the package, users should understand the implications of adding certain constraints to a model projection and apply reasoning and biological knowledge as appropriate.

Most SDMs tend to either overfit (leading to a prediction that reproduces the data) or indicate areas as suitable habitat that might be unreachable for the species or not suitable owing to other non-considered factors (see Fig. 2). A common and practical way to partly address such issues is to constrain the projection to a certain area or neighbourhood, although model-based integration can also act as a constraint on the parameter space (Miller et al., 2019; Peel et al., 2019). Besides the incorporation of spatial constraints during the model parametrization, such as by adding projection limits ("distribution(…, limits = layer)") (Cooper et al., 2018) or the inclusion of spatial

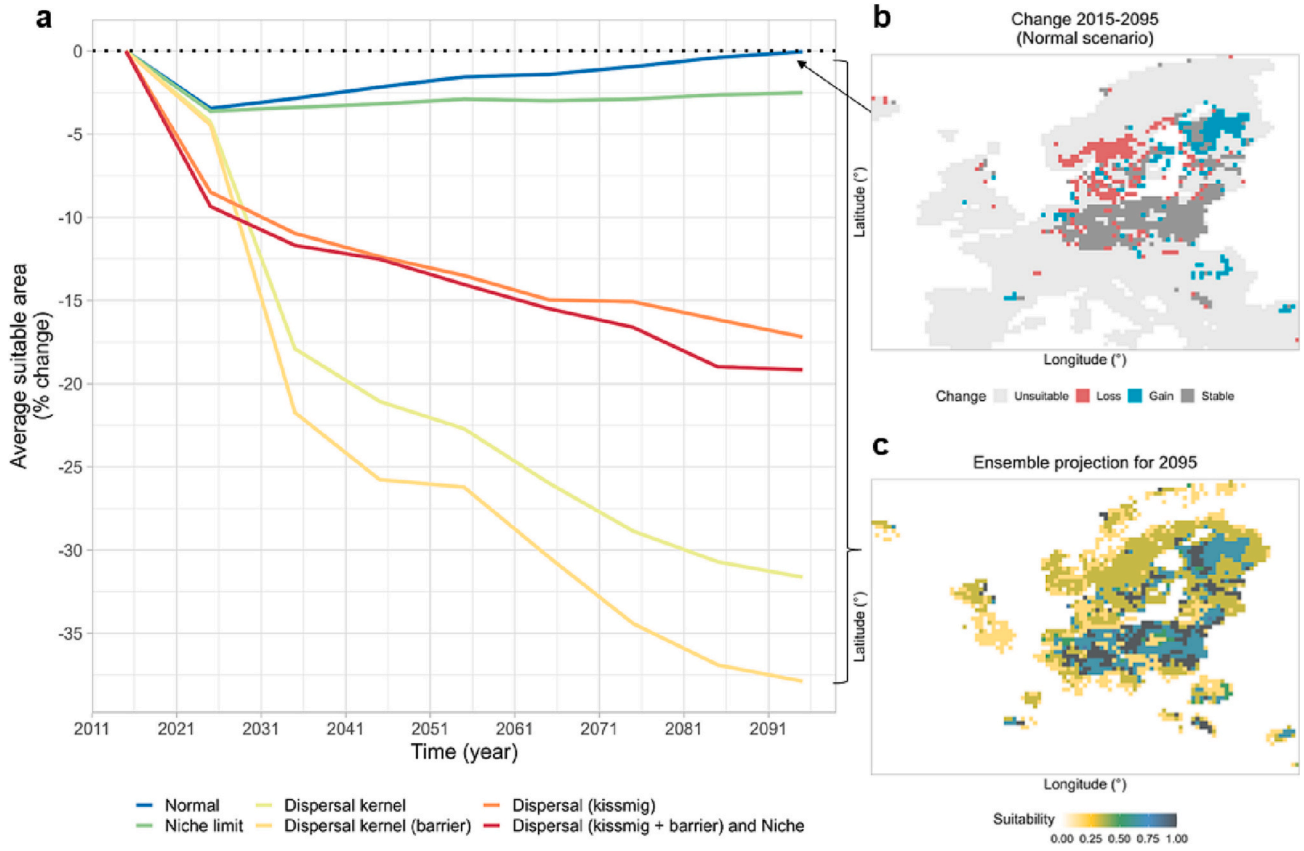


Fig. 3. Future projections of suitable habitat for a virtual species up to the year 2095, with each scenario being run with or without certain constraints related to dispersal, barriers or niche limitations. (a) Shows the projected average suitable habitat from 2015 to 2095 (10 year steps) for various scenarios that include constraints. (b) Change in thresholded suitable habitat between 2015 and 2095 for a scenario without any constraints (blue line in a). The colour of grid cells indicates which areas have been gained, lost or remained stable between the start and end date. (c) Shows an ensemble of all projections in a) for the year 2095, with higher values indicating higher suitability. All code and data with covariates to recreate the figures can be found in the supplementary materials. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

covariates or autocorrelation ("add_spatial_latent()") (Domisch et al., 2019), there are furthermore ways to specifically constrain future projections. The *ibis.iSDM* package here currently considers dispersal, barrier and adaptability constraints that can be added to a projection scenario.

Adding biological informed constraints to projections of correlative SDMs can be seen as another form of data integration, and the resulting "hybrid" SDMs have been shown to perform well compared to non-constrained SDM when projecting to novel conditions (Zurell et al., 2016). The most common constrains added to SDMs are those that limit or enable the dispersal of populations at the margins of a distribution emulating distinct colonization events (Seaborn et al., 2020). The *ibis.iSDM* package supports simple linear and negative exponential dispersal kernel that limit dispersal events to certain distances per time step (Fig. 3a), as well as more sophisticated simulators based on cellular automata such as the popular MIGClM (Engler et al., 2012) or KISSMig R packages (Nobis and Normand, 2014). Constraints can also be added on suitable habitats, corridors or known boundaries that prevent an expansion of a species (Cooper et al., 2018) or on the extent to which a species is able to adapt its niche (Bush et al., 2016). Similar as for inference, the modular structure of scenario objects and ability to add constraints enables convenient expansion of the package (see also development plans).

2.6. Other innovations in the ibis R-package

There are several other smaller innovations in the ibis R-package, which to our knowledge have never been considered or provided in similar form in a SDM framework. Besides having an object-based specification for integrated SDMs (Fig. 1), the use of Bayesian SDMs for estimation also allows for example to visualize not only the mean predicted suitability of a species, but also the pixel-based uncertainty as calculated from a single model posterior, which can be summarized in statistical moments such as standard deviation or the coefficient of variation (Fig. 4). Traditionally, uncertainty has been assessed as variation among different models in an ensemble (Thuiller et al., 2019) as also supported by the "ensemble()" function in ibis.iSDM. This however captures mainly uncertainty among models, opposed to the uncertainty introduced by the data and inferred response function (Hao et al., 2020; Thuiller et al., 2019), which is usually in the investigator's main interest when capturing uncertainty. Here the *ibis.iSDM* provides some plotting functionalities to visualize more than one moment from a posterior of a single model (Fig. 4b).

Similarly, having a pixel-based uncertainty for individual models also allows to create novel types of thresholds. For example, the *ibis.iSDM* package allows with the option 'min.cv' to identify those grid cells that have a high mean suitability, but also low uncertainty (Fig. 4b). A number of other threshold methods are available, for example by maximizing validation statistics such as the Area under the Curve (AUC)

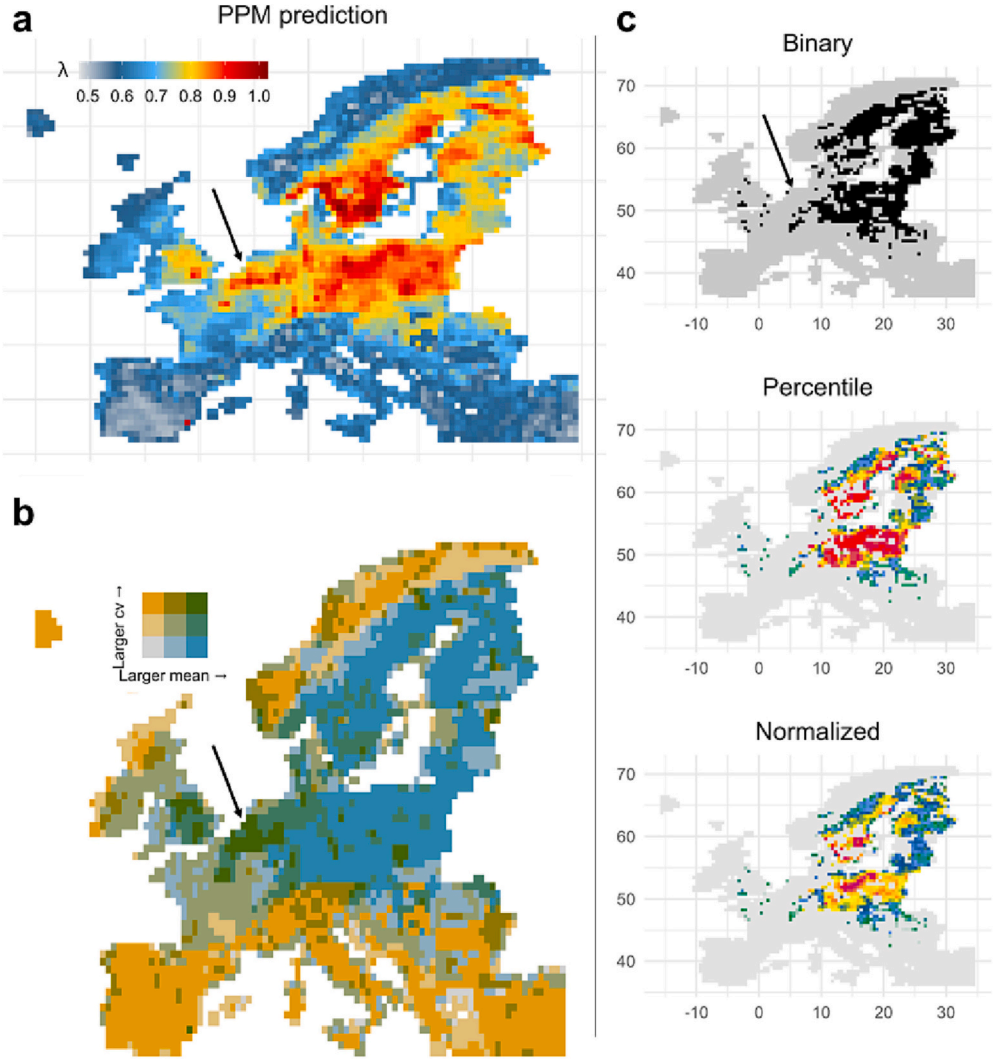


Fig. 4. Single Poisson process model (PPM) of a virtual Scandinavian species using Bayesian regularized regression. (a) Shows the predicted λ of the PPM summarized as mean from the posterior. (b) Bivariate visualization of both the mean and the coefficient of variation from the model posterior. Areas shown in blue have large suitability (expressed as λ) while also having low relative variation. (c) Predictions from b) that have been thresholded to maximize the mean and minimize the coefficient of variation. This form of threshold avoids the separation of areas that are too uncertain to be considered suitable (indicated by arrows). Shown are three different output formats where the remaining values have either been threshold, binned into percentiles or normalized. All code and data with covariates to recreate the figures can be found in the supplementary materials. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

or True Skill Statistics (TSS) using the "modEvA" R-package (Márcia Barbosa et al., 2013), or by thresholding with the minimum presence values (e.g. the minimum value across occurrence points), fixed or percentile values. Finally, all suitability predictions subject to thresholds can be created in binary, categorical percentile and normalized outputs (Fig. 4c). Thresholding to a normalized or percentile characterization of the distribution retains some of the detail of the projected suitability distribution, while also removing uncertain areas and noise.

A general paradigm of the *ibis.iSDM* framework is to support data type specific modelling, e.g. presence-only records are by default always inferred as originating from a Poisson point process. However, there might be use cases where it is more convenient, faster or better explainable to create pseudo-absences points similar as in most of the SDM literature (Phillips et al., 2009; Valavi et al., 2021). Functionalities have been added to specify how pseudo-absences should be added to available occurrence records, such as by sampling them randomly, within a buffer, outside a zonal layer or expert range, or by using a target background (Phillips et al., 2009; Ranc et al., 2017) using the occurrence of other, closely related species (a common practice that can be considered as an integration of external information as well). In a simple comparison of different approaches using presence-only records of the Iberian frog *Discoglossus galganoi* (Fig. 5), I find that sampling pseudo-absences outside an expert-range and using human population density as bias correction performs best (AUC = 0.989, TSS = 0.978), outperforming even targeted background sampling (AUC = 0.940, TSS = 0.88). Although this simple demonstration should not serve as a comprehensive assessment, it again demonstrates the value of using additional sources of biodiversity information for the construction of SDMs.

3. Next steps and further development plans

New advances and literature on how to integrate different data in SDM frameworks continue to be published every year. This R-package aims to offer support for multiple types of data integration, but it does not claim to be the single modelling framework to integrate all different approaches, and other packages to fit SDMs might be more useful for specific use cases (Sillero et al., 2023). Yet, the package is in continuous development and will be gradually improved as time allows. Since many of the functions to fit or project SDMs in this package are designed as modular in nature, there are imminent opportunities for expanding the package with new constraints and integration options.

There are many methodological ways to integrate different data in (spatial) regression model and projections. For example, in a public health context Arambepola *et al.* have developed methods to combine polygon and point estimates via disaggregation regressions so as to downscale critical health related indicators in the absence of finer resolved information (Arambepola et al., 2022). Such approaches naturally connect to the design philosophy of the ibis.iSDM package and similar approaches could be applied to range maps and presence-only records. Other newly developed R-packages allow to infer species occupancy by integrating structured survey with presence-only records, innovatively also making use of nearest-neighbour gaussian process regressions for spatially constrained occupancy models (Doser et al., 2021). Integrated modelling could also be used to incorporate occurrence of multiple different species using for example factor interactions (Leung et al., 2019), multi-nominal predictions using for example convolutional neural networks (Deneu et al., 2021) or co-occurrences through jSDM frameworks where feasible in the context of data integration (Ovaskainen et al., 2017, 2016). Integrated SDMs are likely the most useful in situations where only limited high quality data exist, as most more advanced modelling techniques are quite demanding with regards to the minimum amount of data required (Merow et al., 2014). Nevertheless, further work is necessary to comparatively assess the performance and accuracy of different types of integration such as those outlined in this work.

Integrating data into SDMs can be beneficial to increase the biological realism of predictions. However, especially when making future predictions, SDMs have a number of short-comings, for example by relying on the assumption that species or habitats are in equilibrium with their environment (Elith et al., 2010). One way to account for such conditions is to make SDMs temporally explicit, so that response functions are spatially and temporally varying (Soriano-Redondo et al., 2019), which can help to make better short to medium term forecasts. Another option is to make explicit assumptions through pre-defined processes in mechanistic SDMs, where specific species-environment relationships and the demographic structure and spatial placement of current and future populations can be simulated (Briscoe et al., 2019).

Mechanistic SDM approaches have long been recognized as being particular useful for projections into unknown and non-equilibrium environments (Briscoe et al., 2019; Kearney and Porter, 2009), or for estimating factors related to demography or the dispersal of individuals, which makes them particularly useful for conservation management problems that go beyond the conservation of suitable habitats (Zurell

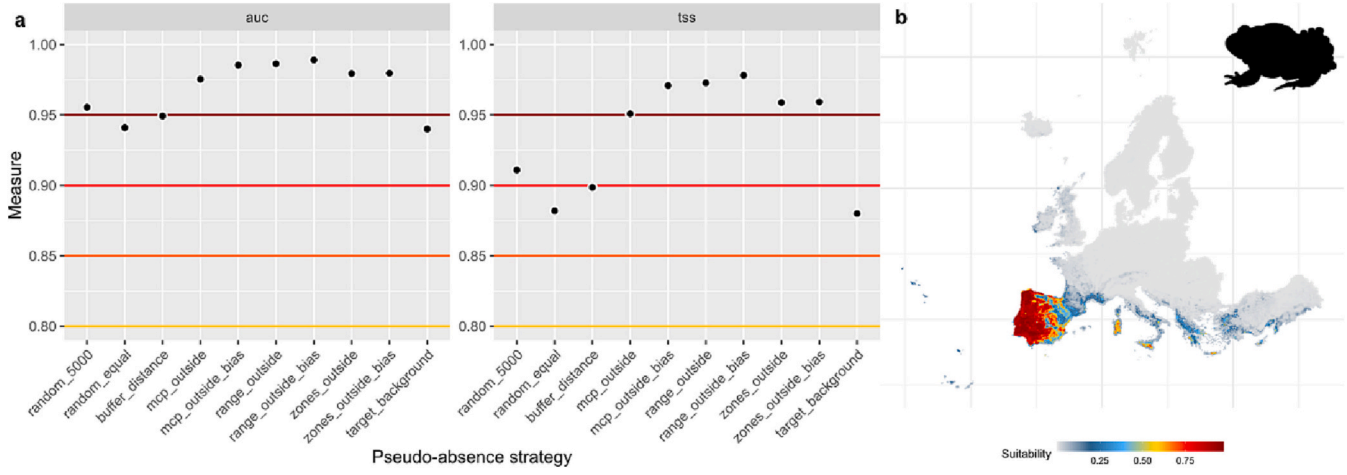


Fig. 5. Validating different practices of pseudo-absence generation using the Iberian frog *Discoglossus galganoi* as model species. (a) Showing measures of the area under the curve (AUC) and true skill statistic (TSS) calculated on withheld data for models using different practices of pseudo-absence generation in *ibis.iSDM*. Horizontal lines indicate 5% improvement steps. Simulations include pseudo-absence generation through random, distance, minimum convex polygons, zonal, range and co-generic targeted background creation. (b) Weighted mean ensemble prediction of individual models, with larger values indicating higher habitat suitability for the species. All code and data with covariates to recreate the figures can be found in the supplementary materials.

et al., 2022). In the *ibis.iSDM* package there are already a few dispersal simulators implemented (see scenario section above) and there furthermore plans to allow for seamless integration with the rangeShifter eco-evolutionary platform (Bocedi et al., 2021). Another idea is to enable support for dedicated equations, for example for population growth or microclimatic thresholds (Schouten et al., 2020), and integrate them into inference and projections (Talluto et al., 2016). Yet, given the data needs and parameter demands for most mechanistic SDMs, and the influence they can have on simulation outcomes, the use of fully mechanistic SDMs will likely remain to limited to specific case studies and model species. Nevertheless, the consideration of further mechanistic modelling approaches can be seen as an important step towards more integrated models.

Code and data availability

The ibis.iSDM R-package can be openly downloaded at https://github.com/iiasa/ibis.iSDM. A R CRAN release is planned in the future. All code and example data used to create the figures in this work is made openly available in the Supplementary Materials (https://osf.io/a6w2k/).

Declaration of Competing Interest

None.

Data availability

Data and code to reproduce the figures in the article are made available in the Supplementary Materials

Acknowledgements

This work has received funding through an European Commission Service contract (07.0202/2020/836131/SER/ENV.D.2) for the project "BIOCLIMA: Assessing Land use, Climate and Biodiversity impacts of National Energy and Climate Plans (NECPs) and National Biodiversity Strategies and Action Plans (NBSAPs) from the EU and its Member States

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2023.102127.

References

Ahmad Suhaimi, S.S., Blair, G.S., Jarvis, S.G., 2021. Integrated species distribution models: a comparison of approaches under different data quality scenarios. Divers. Distrib. 27, 1066–1075. https://doi.org/10.1111/ddi.13255.

Arambepola, R., Lucas, T.C.D., Nandi, A.K., Gething, P.W., Cameron, E., 2022. A simulation study of disaggregation regression for spatial disease mapping. Stat. Med. 41, 1–16. https://doi.org/10.1002/sim.9220.

Araújo, M.B., Guisan, A., 2006. Five (or so) challenges for species distribution modelling. J. Biogeogr. 33, 1677–1688. https://doi.org/10.1111/j.1365-2699.2006.01584.x.

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 Ecol. Evol. 10, 760–766. https://doi.org/10.1111/2041-210X.13168.

Baker, D.J., Maclean, I.M.D., Goodall, M., Gaston, K.J., 2022. Correlations between spatial sampling biases and environmental niches affect species distribution models. Glob. Ecol. Biogeogr. 31, 1038–1050. https://doi.org/10.1111/geb.13491.

Blonder, B., Lamanna, C., Violle, C., Enquist, B.J., 2014. The n-dimensional hypervolume. Glob. Ecol. Biogeogr. 23, 595–609.

Bocedi, G., Palmer, S.C.F., Malchow, A., Zurell, D., Watts, K., Travis, J.M.J., 2021. RangeShifter 2.0: an extended and enhanced platform for modelling spatial eco-evolutionary dynamics and species' responses to environmental changes. Ecography ecog.05687. https://doi.org/10.1111/ecog.05687.

Botella, C., Joly, A., Monestiez, P., Bonnet, P., Munoz, F., 2020. Bias in presence-only niche models related to sampling effort and species niches: lessons for background point selection. PLoS One 15, e0232078. https://doi.org/10.1371/journal.pone.0232078.

Briscoe, N.J., Elith, J., Salguero-Gómez, R., Lahoz-Monfort, J.J., Camac, J.S., Giljohann, K.M., Holden, M.H., Hradsky, B.A., Kearney, M.R., McMahon, S.M., Phillips, B.L., Regan, T.J., Rhodes, J.R., Vesk, P.A., Wintle, B.A., Yen, J.D.L., Guillera-Arroita, G., 2019. Forecasting species range dynamics with process-explicit models: matching methods to applications. Ecol. Lett. 22, 1940–1956. https://doi.org/10.1111/ele.13348.

Brooks, T.M., Pimm, S.L., Akçakaya, H.R., Buchanan, G.M., Butchart, S.H.M., Foden, W., Hilton-Taylor, C., Hoffmann, M., Jenkins, C.N., Joppa, L., Li, B.V., Menon, V., Ocampo-Peñuela, N., Rondinini, C., 2019. Measuring terrestrial area of habitat (AOH) and its utility for the IUCN red list. Trends Ecol. Evol. 34, 977–986. https://doi.org/10.1016/j.tree.2019.06.009.

Bush, A., Mokany, K., Catullo, R., Hoffmann, A., Kellermann, V., Sgrò, C., McEvey, S., Ferrier, S., 2016. Incorporating evolutionary adaptation in species distribution modelling reduces projected vulnerability to climate change. Ecol. Lett. 19, 1468–1478. https://doi.org/10.1111/ele.12696.

Carlson, C.J., 2020. embarcadero: species distribution modelling with Bayesian additive regression trees in R. Methods Ecol. Evol. 11, 850–858. https://doi.org/10.1111/2041-210X.13389.

Chauvier, Y., Zimmermann, N.E., Poggiato, G., Bystrova, D., Brun, P., Thuiller, W., 2021. Novel methods to correct for observer and sampling bias in presence-only species distribution models. Glob. Ecol. Biogeogr. https://doi.org/10.1111/geb.13383 geb.13383.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., Yuan, J., 2023. xgboost: Extreme Gradient Boosting.

Cooper, J.C., Soberón, J., Morueta-Holme, N., 2018. Creating individual accessible area hypotheses improves stacked species distribution model performance. Glob. Ecol. Biogeogr. 27, 156–165. https://doi.org/10.1111/geb.12678.

Deneu, B., Servajean, M., Bonnet, P., Botella, C., Munoz, F., Joly, A., 2021. Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. PLoS Comput. Biol. 17, e1008856 https://doi.org/10.1371/journal.pcbi.1008856.

Domisch, S., Wilson, A.M., Jetz, W., 2016. Model-based integration of observed and expert-based information for assessing the geographic and environmental distribution of freshwater species. Ecography 39, 1078–1088. https://doi.org/10.1111/ecog.01925.

Domisch, S., Friedrichs, M., Hein, T., Borgwardt, F., Wetzig, A., Jähnig, S.C., Langhans, S. D., 2019. Spatially explicit species distribution models: a missed opportunity in conservation planning? Divers. Distrib. 25, 758–769. https://doi.org/10.1111/ddi.12891.

Dorie, V., 2022. dbarts: discrete Bayesian additive regression trees sampler.

Doser, J.W., Finley, A.O., Kéry, M., Zipkin, E.F., 2021. spOccupancy: An R Package for Single Species, Multispecies, and Integrated Spatial Occupancy Models, pp. 1–31.

Elith, J., Leathwick, J.R., 2009. Species distribution models: ecological explanation and prediction across space and time. Annu. Rev. Ecol. Evol. Syst. 40, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159.

Elith, J., Kearney, M., Phillips, S., 2010. The art of modelling range-shifting species. Methods Ecol. Evol. 1, 330–342. https://doi.org/10.1111/j.2041-210X.2010.00036.x.

Engler, R., Hordijk, W., Guisan, A., 2012. The MIGCLIM R package - seamless integration of dispersal constraints into projections of species distribution models. Ecography 35, 872–878. https://doi.org/10.1111/j.1600-0587.2012.07608.x.

Evans, M.E.K., Merow, C., Record, S., McMahon, S.M., Enquist, B.J., 2016. Towards process-based range modeling of many species. Trends Ecol. Evol. 31, 860–871. https://doi.org/10.1016/j.tree.2016.08.005.

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 Ecol. Evol. 6, 424–438. https://doi.org/10.1111/2041-210X.12242.

Fletcher, R.J., 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, e02710. https://doi.org/10.1002/ecy.2710.

Fois, M., Cuena-Lombraña, A., Fenu, G., Bacchetta, G., 2018. Using species distribution models at local scale to guide the search of poorly known species: review, methodological issues and future directions. Ecol. Model. 385, 124–132. https://doi.org/10.1016/j.ecolmodel.2018.07.018.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1–22. https://doi.org/10.18637/jss.v033.i01.

Gabry, J., Češnovar, R., 2022. cmdstanr: R Interface to "CmdStan.".

Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., 2014. Maxent is not a presence-absence method: a comment on Thibaud et al. Methods Ecol. Evol. 5, 1192–1197. https://doi.org/10.1111/2041-210X.12252.

Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecol. Lett. 8, 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x.

Hannemann, H., Willis, K.J., Macias-Fauria, M., 2016. The devil is in the detail: unstable response functions in species distribution models challenge bulk ensemble modelling. Glob. Ecol. Biogeogr. 25, 26–35. https://doi.org/10.1111/geb.12381.

Hao, T., Elith, J., Lahoz-Monfort, J.J., Guillera-Arroita, G., 2020. Testing whether ensemble modelling is advantageous for maximising predictive performance of species distribution models. Ecography 43, 549–558. https://doi.org/10.1111/ecog.04890.

Heberling, J.M., Miller, J.T., Noesgaard, D., Weingart, S.B., Schigel, D., 2021. Data integration enables global biodiversity synthesis. Proc. Natl. Acad. Sci. U. S. A. 118, e2018093118 https://doi.org/10.1073/pnas.2018093118.

Hertzog, L.R., Frank, C., Klimek, S., Röder, N., Böhner, H.G.S., Kamp, J., 2021. Model-based integration of citizen science data from disparate sources increases the

precision of bird population trends. Divers. Distrib. https://doi.org/10.1111/ddi.13259 ddi.13259.
Hofner, B., Müller, J., Hothorn, T., 2011. Monotonicity-constrained species distribution models. Ecology 92, 1895–1901.
Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M., Hofner, B., 2022. mboost: Model-Based Boosting.
Hughes, A.C., Orr, M.C., Ma, K., Costello, M.J., Waller, J., Provoost, P., Yang, Q., Zhu, C., Qiao, H., 2021. Sampling biases shape our view of the natural world. Ecography 44, 1259–1269. https://doi.org/10.1111/ecog.05926.
Isaac, N.J.B., 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.B., 2020. Data integration for large-scale models of species distributions. Trends Ecol. Evol. 35, 56–67. https://doi.org/10.1016/j.tree.2019.08.006.
Jetz, W., McGeoch, M.A., Guralnick, R., Ferrier, S., Beck, J., Costello, M.J., Fernandez, M., Geller, G.N., Keil, P., Merow, C., Meyer, C., Muller-Karger, F.E., Pereira, H.M., Regan, E.C., Schmeller, D.S., Turak, E., 2019. Essential biodiversity variables for mapping and monitoring species populations. Nat. Ecol. Evolut. 3, 539–551. https://doi.org/10.1038/s41559-019-0826-1.
Jung, M., 2022. Predictability and transferability of local biodiversity environment relationships. PeerJ 10, e13872. https://doi.org/10.7717/peerj.13872.
Jung, M., Arnell, A., de Lamo, X., García-Rangel, S., Lewis, M., Mark, J., Merow, C., Miles, L., Ondo, I., Pironon, S., Ravilious, C., Rivers, M., Schepaschenko, D., Tallowin, O., van Soesbergen, A., Govaerts, R., Boyle, B.L., Enquist, B.J., Feng, X., Gallagher, R., Maitner, B., Meiri, S., Mulligan, M., Ofer, G., Roll, U., Hanson, J.O., Jetz, W., Di Marco, M., McGowan, J., Rinnan, D.S., Sachs, J.D., Lesiv, M., Adams, V. M., Andrew, S.C., Burger, J.R., Hannah, L., Marquet, P.A., McCarthy, J.K., Morueta-Holme, N., Newman, E.A., Park, D.S., Roehrdanz, P.R., Svenning, J.-C., Violle, C., Wieringa, J.J., Wynne, G., Fritz, S., Strassburg, B.B.N., Obersteiner, M., Kapos, V., Burgess, N., Schmidt-Traub, G., Visconti, P., 2021. Areas of global importance for conserving terrestrial biodiversity, carbon and water. Nat. Ecol. Evolut. 5, 1499–1509. https://doi.org/10.1038/s41559-021-01528-7.
Kearney, M., Porter, W., 2009. Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. Ecol. Lett. 12, 334–350. https://doi.org/10.1111/j.1461-0248.2008.01277.x.
Koshkina, V., Wang, Y., Gordon, A., Dorazio, R.M., White, M., Stone, L., 2017. Integrated species distribution models: combining presence-background data and site-occupancy data with imperfect detection. Methods Ecol. Evol. 8, 420–430. https://doi.org/10.1111/2041-210X.12738.
Lawson, C.R., Hodgson, J.A., Wilson, R.J., Richards, S.A., 2014. Prevalence, thresholds and the performance of presence-absence models. Methods Ecol. Evol. 5, 54–64. https://doi.org/10.1111/2041-210X.12123.
Leclère, D., Obersteiner, M., Barrett, M., Butchart, S.H.M., Chaudhary, A., De Palma, A., DeClerck, F.A.J., Di Marco, M., Doelman, J.C., Dürauer, M., Freeman, R., Harfoot, M., Hasegawa, T., Hellweg, S., Hilbers, J.P., Hill, S.L.L., Humpenöder, F., Jennings, N., Krisztin, T., Mace, G.M., Ohashi, H., Popp, A., Purvis, A., Schipper, A. M., Tabeau, A., Valin, H., van Meijl, H., van Zeist, W.-J., Visconti, P., Alkemade, R., Almond, R., Bunting, G., Burgess, N.D., Cornell, S.E., Di Fulvio, F., Ferrier, S., Fritz, S., Fujimori, S., Grooten, M., Harwood, T., Havlík, P., Herrero, M., Hoskins, A. J., Jung, M., Kram, T., Lotze-Campen, H., Matsui, T., Meyer, C., Nel, D., Newbold, T., Schmidt-Traub, G., Stehfest, E., Strassburg, B.B.N., van Vuuren, D.P., Ware, C., Watson, J.E.M., Wu, W., Young, L., 2020. Bending the curve of terrestrial biodiversity needs an integrated strategy. Nature 585, 551–556. https://doi.org/10.1038/s41586-020-2705-y.
Lee-Yaw, A., McCune, L., Pironon, J., Sheth, N., 2022. Species distribution models rarely predict the biology of real populations. Ecography 2022. https://doi.org/10.1111/ecog.05877.
Leitão, P.J., Toraño Caicoya, A., Dahlkamp, A., Guderjan, L., Griesser, M., Haverkamp, P. J., Nordén, J., Snäll, T., Schröder, B., 2022. Impacts of forest management on forest bird occurrence patterns—a case study in Central Europe. Front. Forests Global Change 5. https://doi.org/10.3389/ffgc.2022.786556.
Leung, B., Hudgins, E.J., Potapova, A., Ruiz-Jaen, M.C., 2019. A new baseline for countrywide α-diversity and species distributions: illustration using >6,000 plant species in Panama. Ecol. Appl. 29, 1–13. https://doi.org/10.1002/eap.1866.
Lindgren, F., Rue, H., 2015. Bayesian spatial modelling with R-INLA. J. Stat. Softw. 63, 1–25.
Liu, C., White, M., Newell, G., 2013. Selecting thresholds for the prediction of species occurrence with presence-only data. J. Biogeogr. 40, 778–789. https://doi.org/10.1111/jbi.12058.
Márcia Barbosa, A., Real, R., Muñoz, A.-R., Brown, J.A., 2013. New measures for assessing model equilibrium and prediction mismatch in species distribution models. Divers. Distrib. 19, 1333–1338. https://doi.org/10.1111/ddi.12100.
Merow, C., Smith, M.J., Silander, J.A., 2013. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. Ecography 36, 1058–1069. https://doi.org/10.1111/j.1600-0587.2013.07872.x.
Merow, C., Smith, M.J., Edwards, T.C., Guisan, A., McMahon, S.M., Normand, S., Thuiller, W., Wüest, R.O., Zimmermann, N.E., Elith, J., 2014. What do we gain from simplicity versus complexity in species distribution models? Ecography 37, 1267–1281. https://doi.org/10.1111/ecog.00845.
Merow, C., Allen, J.M., Aiello-Lammens, M., Silander, J.A., 2016. Improving niche and range estimates with Maxent and point process models by integrating spatially explicit information. Glob. Ecol. Biogeogr. 25, 1022–1036. https://doi.org/10.1111/geb.12453.
Merow, C., Wilson, A.M., Jetz, W., 2017. Integrating occurrence data and expert maps for improved species range predictions. Glob. Ecol. Biogeogr. 26, 243–258. https://doi.org/10.1111/geb.12539.
Meyer, C., Kreft, H., Guralnick, R., Jetz, W., 2015. Global priorities for an effective information basis of biodiversity distributions. Nat. Commun. 6, 8221. https://doi.org/10.1038/ncomms9221.
Miller, D.A.W., Pacifici, K., Sanderlin, J.S., Reich, B.J., 2019. The recent past and promising future for data integration methods to estimate species' distributions. Methods Ecol. Evol. 10, 22–37. https://doi.org/10.1111/2041-210X.13110.
Morán-Ordóñez, A., Lahoz-Monfort, J.J., Elith, J., Wintle, B.A., 2017. Evaluating 318 continental-scale species distribution models over a 60-year prediction horizon: what factors influence the reliability of predictions? Glob. Ecol. Biogeogr. 26, 371–384.
Mostert, P., Bjørkås, R., Bruls, A.J.H.M., Koch, W., Martin, E.C., 2022. $_{int}$ SDM: a reproducible framework for integrated species distribution models (preprint). Ecology. https://doi.org/10.1101/2022.09.15.507996.
Nobis, M.P., Normand, S., 2014. KISSMig - a simple model for R to account for limited migration in analyses of species distributions. Ecography 37, 1282–1287. https://doi.org/10.1111/ecog.00930.
Ovaskainen, O., Roy, D.B., Fox, R., Anderson, B.J., 2016. Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. Methods Ecol. Evol. 7, 428–436. https://doi.org/10.1111/2041-210X.12502.
Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T., Abrego, N., 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. Ecol. Lett. 20, 561–576. https://doi.org/10.1111/ele.12757.
Pacifici, K., Reich, B.J., Miller, D.A.W., Pease, B.S., 2019. Resolving misaligned spatial data with integrated species distribution models. Ecology 100, 1–15. https://doi.org/10.1002/ecy.2709.
Pebesma, E., 2022. stars: Spatiotemporal Arrays, Raster and Vector Data Cubes.
Peel, S.L., Hill, N.A., Foster, S.D., Wotherspoon, S.J., Ghiglione, C., Schiaparelli, S., 2019. Reliable species distributions are obtainable with sparse, patchy and biased data by leveraging over species and data types. Methods Ecol. Evol. 10, 1002–1014. https://doi.org/10.1111/2041-210X.13196.
Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol. Appl. 19, 181–197.
Phillips, S.J., Anderson, R.P., Dudík, M., Schapire, R.E., Blair, M.E., 2017. Opening the black box: an open-source release of Maxent. Ecography. https://doi.org/10.1111/ecog.03049.
Ranc, N., Santini, L., Rondinini, C., Boitani, L., Poitevin, F., Angerbjörn, A., Maiorano, L., 2017. Performance tradeoffs in target-group bias correction for species distribution models. Ecography 40, 1076–1087. https://doi.org/10.1111/ecog.02414.
Renner, I.W., Warton, D.I., 2013. Equivalence of MAXENT and poisson point process models for species distribution modeling in ecology. Biometrics 69, 274–281. https://doi.org/10.1111/j.1541-0420.2012.01824.x.
Renner, I.W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S.J., Popovic, G., Warton, D.I., 2015. Point process models for presence-only analysis. Methods Ecol. Evol. 6, 366–379. https://doi.org/10.1111/2041-210X.12352.
Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929. https://doi.org/10.1111/ecog.02881.
Santini, L., Benítez-López, A., Maiorano, L., Čengić, M., Huijbregts, M.A.J., 2021. Assessing the reliability of species distribution projections in climate change research. Divers. Distrib. 27, 1035–1050. https://doi.org/10.1111/ddi.13252.
Schouten, R., Vesk, P.A., Kearney, M.R., 2020. Integrating dynamic plant growth models and microclimates for species distribution modelling. Ecol. Model. 435, 109262 https://doi.org/10.1016/j.ecolmodel.2020.109262.
Scott, S., 2022. _BoomSpikeSlab: MCMC for Spike and Slab Regression_.
Seaborn, T.J., Goldberg, C.S., Crespi, E.J., 2020. Integration of dispersal data into distribution modeling: what have we done and what have we learned? Front. Biogeogr. 12, 80. https://doi.org/10.21425/F5FBG43130.
Sillero, N., Campos, J.C., Arenas-Castro, S., Barbosa, A.M., 2023. A curated list of R packages for ecological niche modelling. Ecol. Model. 476, 110242 https://doi.org/10.1016/j.ecolmodel.2022.110242.
Simmonds, E.G., Jarvis, S.G., Henrys, P.A., Isaac, N.J.B., O'Hara, R.B., 2020. Is more data always better? A simulation study of benefits and limitations of integrated distribution models. Ecography ecog.05146. https://doi.org/10.1111/ecog.05146.
Smith, A.B., Santos, M.J., 2020. Testing the ability of species distribution models to infer variable importance. Ecography 43, 1801–1813. https://doi.org/10.1111/ecog.05317.
Soriano-Redondo, A., Jones-Todd, C.M., Bearhop, S., Hilton, G.M., Lock, L., Stanbury, A., Votier, S.C., Illian, J.B., 2019. Understanding species distribution in dynamic populations: a new approach using spatio-temporal point process models. Ecography 42, 1092–1102. https://doi.org/10.1111/ecog.03771.
Soultan, A., Pavón-Jordán, D., Bradter, U., Sandercock, B.K., Hochachka, W.M., Johnston, A., Brommer, J., Gaget, E., Keller, V., Knaus, P., Aghababyan, K., Maxhuni, Q., Vintchevski, A., Nagy, K., Raudonikis, L., Balmer, D., Noble, D., Leitão, D., Øien, I.J., Shimmings, P., Sultanov, E., Caffrey, B., Boyla, K., Radišić, D., Lindström, Å., Velevski, M., Pladevall, C., Brotons, L., Karel, Š., Rajković, D.Z., Chodkiewicz, T., Wilk, T., Szép, T., van Turnhout, C., Foppen, R., Burfield, I., Vikstrøm, T., Mazal, V.D., Eaton, M., Vorisek, P., Lehikoinen, A., Herrando, S., Kuzmenko, T., Bauer, H.-G., Kalyakin, M.V., Voltzit, O.V., Sjeničić, J., Pärt, T., 2022. The future distribution of wetland birds breeding in Europe validated against observed changes in distribution. Environ. Res. Lett. 17, 024025 https://doi.org/10.1088/1748-9326/ac4ebe.

Stan Development Team, 2022. RStan: the R interface to Stan.
Talluto, M.V., Boulangeat, I., Ameztegui, A., Aubin, I., Berteaux, D., Butler, A., Doyon, F., Drever, C.R., Fortin, M.-J., Franceschini, T., Liénard, J., McKenney, D., Solarik, K.A., Strigul, N., Thuiller, W., Gravel, D., 2016. Cross-scale integration of knowledge for predicting species ranges: a metamodelling framework. Glob. Ecol. Biogeogr. 25, 238–249. https://doi.org/10.1111/geb.12395.
Thuiller, W., Lafourcade, B., Engler, R., Araújo, M.B., 2009. BIOMOD - a platform for ensemble forecasting of species distributions. Ecography 32, 369–373. https://doi.org/10.1111/j.1600-0587.2008.05742.x.
Thuiller, W., Guéguen, M., Renaud, J., Karger, D.N., Zimmermann, N.E., 2019. Uncertainty in ensembles of global biodiversity scenarios. Nat. Commun. 10, 1446. https://doi.org/10.1038/s41467-019-09519-w.
Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., 2021. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. Ecol. Monogr. 0, 1–27. https://doi.org/10.1002/ecm.1486.
Warton, D.I., Shepherd, L.C., 2010. Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. Ann. Appl. Stat. 4, 1383–1402. https://doi.org/10.1214/10-AOAS331.
Warton, D.I., Renner, I.W., Ramp, D., 2013. Model-based control of observer bias for the analysis of presence-only data in ecology. PLoS One 8, e79168. https://doi.org/10.1371/journal.pone.0079168.
Weber, M.M., Stevens, R.D., Diniz-Filho, J.A.F., Grelle, C.E.V., 2017. Is there a correlation between abundance and environmental suitability derived from ecological niche modelling? A meta-analysis. Ecography 40, 817–828. https://doi.org/10.1111/ecog.02125.
Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis, UseR!, Use R! Springer International Publishing, New York, NY. https://doi.org/10.1007/978-3-319-24277-4.
Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., Farooq, H., Herdean, A., Ariza, M., Scharn, R., Svantesson, S., Wengström, N., Zizka, V., Antonelli, A., 2019. CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases. Methods Ecol. Evol. 10, 744–751. https://doi.org/10.1111/2041-210X.13152.
Zurell, D., Thuiller, W., Pagel, J., Cabral, J.S., Münkemüller, T., Gravel, D., Dullinger, S., Normand, S., Schiffers, K.H., Moore, K.A., Zimmermann, N.E., 2016. Benchmarking novel approaches for modelling species range dynamics. Glob. Chang. Biol. 22, 2651–2664. https://doi.org/10.1111/gcb.13251.
Zurell, D., König, C., Malchow, A., Kapitza, S., Bocedi, G., Travis, J., Fandos, G., 2022. Spatially explicit models for decision-making in animal conservation and restoration. Ecography 2022. https://doi.org/10.1111/ecog.05787 ecog.05787.