An integrated species distribution modelling framework for heterogeneous biodiversity data

Martin Jung

PII: S1574-9541(23)00156-5
DOI: https://doi.org/10.1016/j.ecoinf.2023.102127
Reference: ECOINF 102127

To appear in: Ecological Informatics

Received date: 21 February 2023
Revised date: 10 April 2023
Accepted date: 12 May 2023

Please cite this article as: M. Jung, An integrated species distribution modelling framework for heterogeneous biodiversity data, Ecological Informatics (2023), https://doi.org/10.1016/j.ecoinf.2023.102127

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2023 Published by Elsevier B.V.
An integrated species distribution modelling framework for heterogeneous biodiversity data

Martin Jung¹

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

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.

keywords

Species distribution model, data integration, environmental niche, offset, point-process-model, Bayesian, R-package
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 (A. Lee-Yaw et al., 2022; Evans et al., 2016; Hannemann et al., 2016; 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, 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.

**Modelling framework:**

**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 data sets 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 \( \lambda \) integrated over of an area \( A \), e.g. \( N(y) \approx \text{Poisson} \left( \int_A \lambda(s) ds \right) \). The intensity \( \lambda \) can be estimated as \( \log(\lambda_s) = \beta_0 + \beta_k x_s + \varepsilon_s \) based on thinned observations \( s \), \( \beta \) being the 1 to \( k \) coefficients in the model including an intercept \( (\beta_0) \), \( x \) being the covariate values in given area and \( \varepsilon \) 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 (Figure 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 tidyverseobject structure (Wickham, 2016), allowing to retain data and functions contained within each object to facilitate reuse through other functions (Figure 1, SI Figure 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 Figure 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 known areas of occurrence, can also be added to the same object (SI Figure 1). Finally, after specifying an engine and training the model, the resulting fit can then be interrogated and validated (Figure 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.

Figure 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 Figure 1. Icons are created by the authors or are under public domain (CC-0).

**Integration**

The *ibis.iSDM* package supports all types of integration outlined by Fletcher et al. (2019), some even in multiple different ways (Figure 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 without and with weights (Figure 2a-b), the latter can for example give higher weight to potentially fewer, but more accurate records (Figure 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 often 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 its direction using for example monotonicity constraints (e.g. specifying that a certain variable have to be positive, Figure 2g). Many priors can be particularly useful to avoid nonsensical response functions (Hofner et al., 2011), for example when owing to differences in grain a known forest-associated species does show the intended directional response towards this variable.

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 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 (Figure 2h).
Figure 2: The suitable habitat estimated with a SDM can vary depending on how different datasets are integrated as shown for the European ground squirrel (*Spermophilus 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.

**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 derivate, as done in the popular maxent/maxnet modelling approach (Merow et al., 2013; Phillips et al., 2017). Functionalities to create such derivate are readily available when adding covariates to a distribution model (see SI Figure 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/).

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 (Figure 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 Figure 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.

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; Soutan 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 (Figure 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.
Figure 3: Future projections of suitable habitat for a virtual species up to the year 2095, with each scenario being run without or with 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.

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 Figure 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 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 (Figure 3a), as well as more sophisticated simulators based on cellular automata such as the popular MIGCIM (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).

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 (Figure 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 (Figure 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 exclusively 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 (Figure 4b).
Figure 4: Single Poisson process model (PPM) of a virtual Scandinavian species using Bayesian regularized regression. (a) Shows the predicted $\lambda$ 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 $\lambda$) 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.

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 (Figure 4b). A number of other threshold methods are available, for example by maximizing validation statistics such as the Area under the Curve (AUC) 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.
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-absence 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* (Figure 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 = 0.976), 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.

![Figure 5](image)

Figure 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.

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 (Deser 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 shortcomings, 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 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., 2020).
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.

Acknowledgements and funding:

This project was funded by a European Union tender grant (BIOCLIMA). I thank the members of the IIASA Biodiversity, Ecology and Conservation group for testing out the package and providing first feedback and bug reports.

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/).

Conflict of interest:

None

References:


Pebesma, E., 2022. stars: Spatiotemporal Arrays, Raster and Vector Data Cubes.


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? Frontiers of Biogeography 12, 80. https://doi.org/10.21425/F5FBG43130


breeding in Europe validated against observed changes in distribution. Environmental Research Letters 17, 024025. https://doi.org/10.1088/1748-9326/ac4ebe


• Most biodiversity data are biased and of heterogeneous origin and types.
• A species distribution modelling framework is presented in the form of an R-package.
• The ibis.iSDM package is highly modular and enables linear and non-linear inference and projections of species distributions.
• Various forms of integration are supported, and their use demonstrated through examples.
• Additional functionalities allow the creation of unconstrained and constrained spatial-temporal scenarios.
• Advances and challenges in model-based integration are highlighted.