ECOGRAPHY

Software notes

megaSDM: integrating dispersal and time-step analyses into species distribution models

Benjamin R. Shipley, Renee Bach, Younje Do, Heather Strathearn, Jenny L. McGuire and Bistra Dilkina

B. R. Shipley (https://orcid.org/0000-0003-0739-309X) ✉ (bshipley6@gatech.edu) and J. L. McGuire, School of Biological Sciences, Georgia Inst. of Technology, Atlanta, GA, USA. JLM also at: School of Earth and Atmospheric Sciences, Georgia Inst. of Technology, Atlanta, GA, USA and Interdisciplinary Graduate Program in Quantitative Biosciences, Georgia Inst. of Technology, Atlanta, GA, USA. – R. Bach, Y. Do and B. Dilkina, School of Computational Science and Engineering, Georgia Inst. of Technology, Atlanta, GA, USA. YD also at: Amazon, Seattle, WA, USA. BD also at: Viterbi School of Engineering, Univ. of Southern California, Los Angeles, CA, USA. – H. Strathearn, Lyles School of Civil Engineering, Purdue Univ., West Lafayette, IN, USA and Dept of Civil and Environmental Engineering, Stanford Univ., Palo Alto, CA, USA.

Ecography
2022: e05450
doi: 10.1111/ecog.05450

Subject Editor: Thiago F. Rangel
Editor-in-Chief:
Jens-Christian C Svenning
Accepted 14 September 2021

Understanding how species ranges shift as climates rapidly change informs us how to effectively conserve vulnerable species. Species distribution models (SDMs) are an important method for examining these range shifts. The tools for performing SDMs are ever improving. Here, we present the `megaSDM` R package. This package facilitates realistic spatiotemporal SDM analyses by incorporating dispersal probabilities, creating time-step maps of range change dynamics and efficiently handling large datasets and computationally intensive environmental subsampling techniques. Provided a list of species and environmental data, `megaSDM` synthesizes GIS processing, subsampling methods, MaxEnt modelling, dispersal rate restrictions and additional statistical tools to create a variety of outputs for each species, time period and climate scenario requested. For each of these, `megaSDM` generates a series of distribution maps and outputs visual representations of statistical data. `megaSDM` offers several advantages over other commonly used SDM tools. First, many of the functions in `megaSDM` natively implement parallelization, enabling the package to handle large amounts of data efficiently without the need for additional coding. `megaSDM` also implements environmental subsampling of occurrences, making the technique broadly available in a way that was not possible before due to computational considerations. Uniquely, `megaSDM` generates maps showing the expansion and contraction of a species range across all considered time periods (time-maps), and constrains both presence/absence and continuous suitability maps of species ranges according to species-specific dispersal constraints. The user can then directly compare non-dispersal and dispersal-limited distribution predictions. This paper discusses the unique features and highlights of `megaSDM`, describes the structure of the package and demonstrates the package's features and the model flow through examples.

Keywords: climate change, dispersal, ecological niche model, habitat, MaxEnt, range, range shift, species richness, time step

NORDIC SOCIETY OIKOS

www.ecography.org

Background

Increasing greenhouse emissions continue to influence global climate. The ability of species to track these changes and persist in suitable habitats may determine if they are able to avoid extinction within the next century. Understanding the dynamics of species' range expansions and contractions in the context of a rapidly shifting climate is of paramount importance. Researchers often use species distribution modelling to predict suitable habitats for species for use in research and management. Species distribution models (SDMs), also known as ecological niche models or habitat suitability models, use the environmental conditions at geo-referenced species observations (hereafter, occurrences or occurrence points) to estimate suitable habitats and provide hypotheses for the spatial distribution of the species (Varela et al. 2014). In addition, SDMs are frequently extrapolated for past or future climates, making predictions about how a species range might change under different environmental conditions and how these ranges interact with the existing network of protected areas and regions of high human impact (Elith et al. 2010).

Researchers have begun applying SDMs over multiple time steps to gauge the stability of transitory habitats (i.e. habitats that are suitable for a species for only a brief amount of time) (Early and Sax 2011, Huang et al. 2020), for many species simultaneously (Lehtomäki et al. 2019), and for species that are constrained by their dispersal ability (Schloss et al. 2012). However, the availability of statistical software aimed at investigating these intricate questions has lagged behind the field. For example, no software tools have yet been developed to investigate transitory range dynamics, and many of the tools currently in use do not natively implement the newest methods for accurate modelling. Recent R packages have been developed that incorporate some components of these advances such as species-specific dispersal rate (e.g. 'MIGCLIM'; Engler et al. 2012) and a variety of options for occurrence and background subsampling (e.g. 'ecospat'; Di Cola et al. 2017). However, no SDM software is yet able to efficiently implement environmental subsampling and integrate dispersal ability to evaluate changes in habitat suitability for many species and climate models at once, nor display the results of such analyses in an easily interpretable manner.

Here, we present the R package `megaSDM`, which applies a new, efficient implementation of environmental subsampling, the generation of distribution maps (using the MaxEnt software by default) showing dispersal-constrained range shifts across multiple time steps, and native parallel processing. This package provides an improvement in the implementation and efficiency of investigations of species range shifts (Fig. 1). It integrates multi-step range movements and species-specific dispersal rate to predict with greater accuracy how species ranges and richness will respond on the landscape to the dynamic pressures of current and future climate change (Fig. 2).

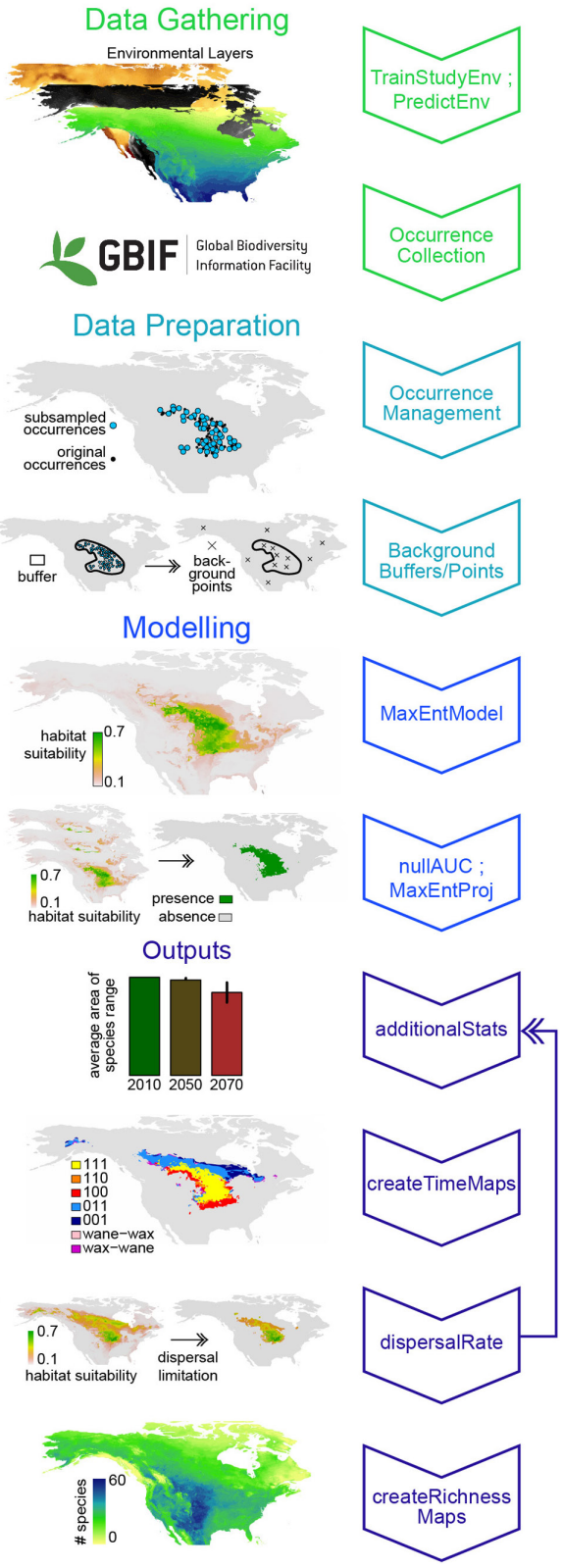

Figure 1. Simplified flowchart of megaSDM, showing each of the package's main functions and example outputs.

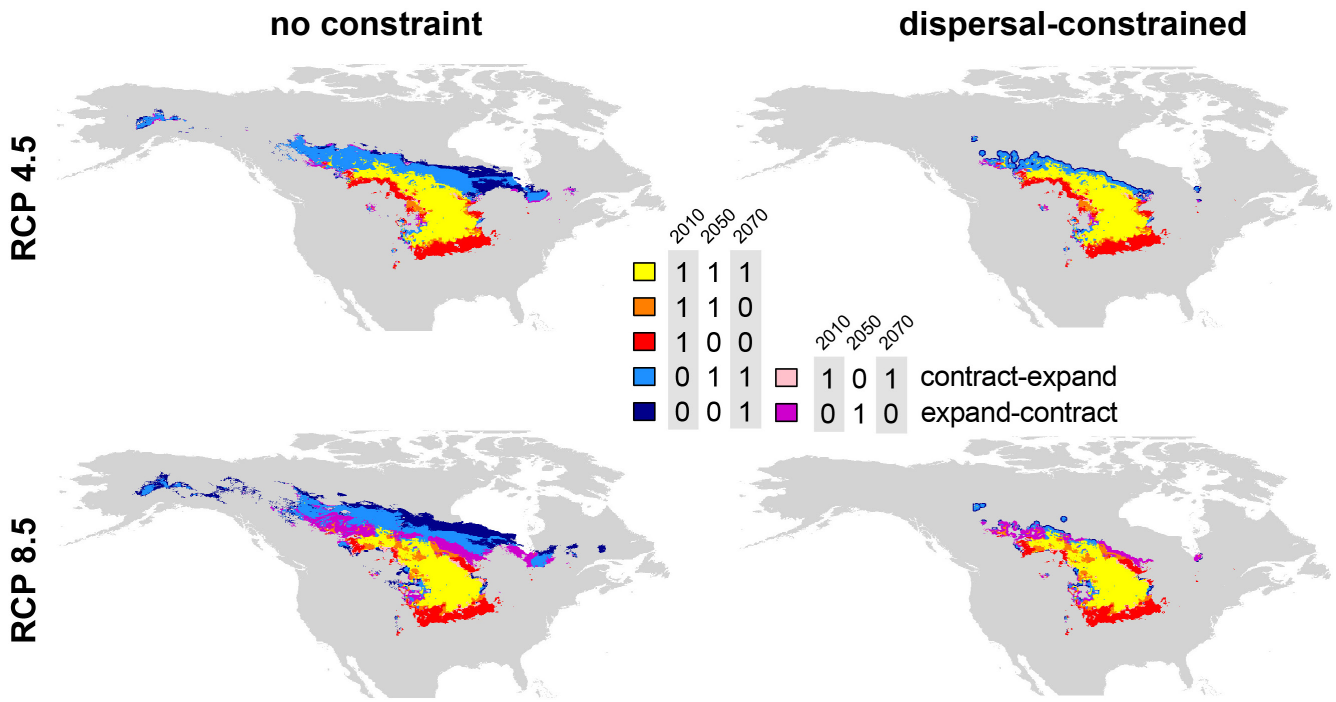

Figure 2. Output 'time maps' from the createTimeMaps() function in `megaSDM`, detailing range shifts for Franklin's ground squirrel *Poliocitellus franklinii* across two different climate scenarios (RCP4.5 and RCP8.5; Riahi et al. 2011, Thomson et al. 2011, respectively) and for three separate times (2010, 2050 and 2070). Blue regions indicate areas of expansion, red regions indicate areas of contraction and purple/pink areas indicate areas of momentary fluctuations among the three times (e.g. expansion from 2010 to 2050 followed by contraction from 2050 to 2070). Yellow areas remain occupied throughout the entire time period. The maps in the right column constrain range expansion to the average dispersal rate of *P. franklinii* (1.23 km year^{-1}; Schloss et al. 2012).

Package highlights

`megaSDM` has many innovative features for modelling species range dynamics, but its most important ability is to efficiently synthesize the SDMs for many species, time periods and climate scenarios, outputting unique maps describing the changes in species ranges in response to environmental changes (Fig. 2). These maps succinctly represent transitory range dynamics, where the range of a species expands briefly before contracting again, or vice versa. Transitory range dynamics such as those can significantly influence the availability and accessibility of habitat for range-contracting species (Huang et al. 2020). Although some packages can effectively display unidirectional range shifts (e.g. 'kuenm', Cobos et al. 2019), no other SDM software generates maps that display these transitory range dynamics.

In generating these range maps, `megaSDM` can also integrate dispersal limitations into SDMs using probability functions. Some R packages incorporate dispersal rate into binary SDMs that show presence and absence (e.g. 'MIGCLIM'; Engler et al. 2012). However, applying dispersal rate probabilities to continuous habitat suitability models has not achieved widespread use, despite the profound effects of varying thresholds on the interpretation of binary species distributions (Norris 2014). `megaSDM` allows for the integration of dispersal ability into continuous climate suitability models, offering a more nuanced take on dispersal limitations (Fig. 3). To do this, `megaSDM` uses a new metric, called 'invadable suitability', which incorporates both continuous habitat suitability and the dispersal ability of a species. Invadable suitability represents the potential for a species to expand its range into new territory given changing conditions. Given a user-provided set of dispersal data in distance per time-step (per year), invadable suitability is calculated by multiplying the habitat suitability (generated in the SDM) by the probability of dispersal as a function of distance. `megaSDM` models both invadable suitability and the standard dispersal-constrained presence/absence species distribution maps (similar to those created by MIGCLIM) over multiple dispersal events (Fig. 2).

`megaSDM` also implements several improvements on strategies for reducing spatial or environmental bias in the SDMs themselves. Several papers have demonstrated that an environmental-subsampling technique mitigates sampling bias and improves model performance (Varela et al. 2014, Castellanos et al. 2019), but this has thus far not been widely implemented, largely due to computational challenges. `megaSDM` is the first to allow multivariate environmentally stratified filtering of occurrence and background points prior to modelling. `megaSDM` can also generate background (pseudo-absence) points in several different ways, including a new technique that spatially weights the background points by increasing the density of the background points within a buffer around the occurrences (Fig. 4). Other R packages

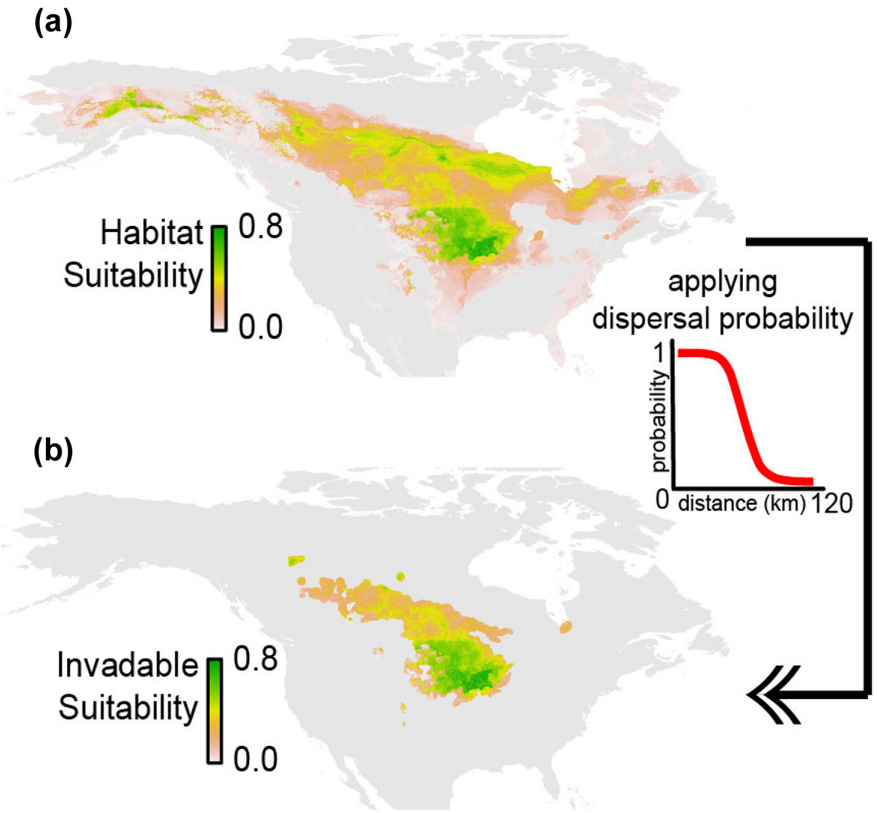

Figure 3. Output maps showing continuous habitat suitability (0–1) for *P. franklinii* for the year 2070 using the RCP 8.5 climate scenario (Riahi et al. 2011) without incorporating dispersal from the 2010 modelled distribution (a) and (b) 'invadable suitability' by multiplying the dispersal probability function (red curve) by the suitability data given an average dispersal rate of 1.23 km year^{-1} since 2010 (Schloss et al. 2012). The red curve approximates the probability of dispersal to at least a given distance using a gamma distribution.

allow some of these strategies (e.g. 'dismo' (Hijmans et al. 2017) can generate random or spatially constrained background points). However, no one package has yet to merge spatial and environmental filtering of background points.

The efficient implementation of these features within `megaSDM` has been achieved through the employment of native, multi-core parallel processing within the individual functions. This integrated parallelization allows users to simultaneously analyse the species of interest in batches, without requiring the users to edit the native functions, change the workflow (e.g. running a single species at a time) or apply multi-core parallelization outside of the function

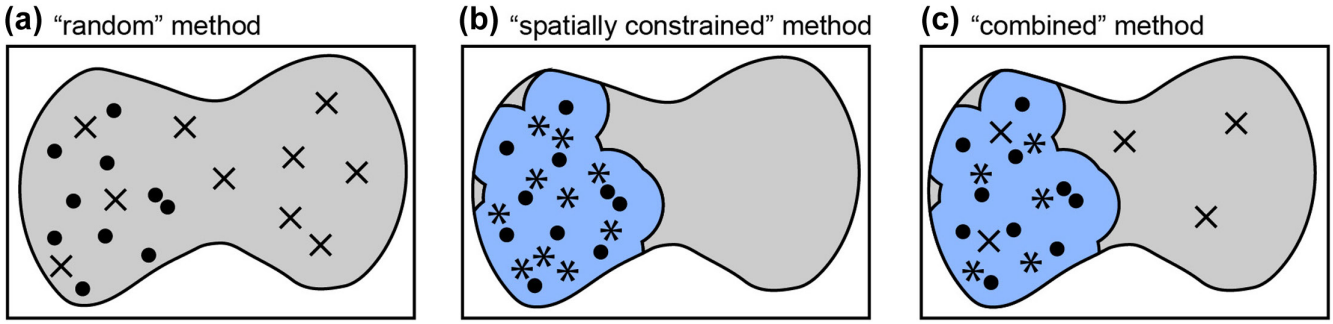

Figure 4. Diagrams detailing the different methods of generating background points available in the BackgroundPoints() function in `megaSDM`. Panel (a) shows the 'random' generation method, where 10 background points (black x's) are sampled randomly throughout the training area (grey polygon) without considering the locations of the occurrence points (black dots). The other commonly used technique for generating background points is the 'spatially constrained' method, in which a buffer is constructed around the occurrence points (b, blue polygon), and the 10 background points are only sampled within the buffer (b, black asterisks). The 'combined' method (c) generates a proportion (given as an argument to the function) of the background points from within the buffer (black asterisks) and the rest from the entire study area (black x's), providing a spatial weighting scheme.

(which runs more slowly than parallelizing within each function). See the Supporting information for details on the relationship between parallel processing and total analysis time for the entire workflow of megaSDM.

Package structure

Structure and configuration of the package

megaSDM comprises a set of independent functions that together perform occurrence and GIS data manipulation, distribution modelling and data analysis (Table 1). These functions are modular, constructed as 'building-block' functions that each contain one aspect of SDM generation and evaluation that can be mixed and matched according to the requirements of the project, but we also provide a cohesive workflow that assists in the entire process of species distribution modelling (Fig. 1, Supporting information; also see the package vignette, Supporting information). megaSDM provides options to manipulate the environmental data, download and filter occurrence points or create background points. Subsequent functions model habitat suitability for all species across all time steps and climate scenarios and provide options to calculate summary statistics for each species distribution, create maps showing the transitory range dynamics of each species and calculate species richness for each combination of times and scenarios. Finally, if the user provides information about the average yearly dispersal rate for the species, megaSDM can apply a dispersal-rate constraint to each output.

Data inputs

At a minimum, megaSDM must be provided with 1) a vector of species names to be analysed and 2) sets of environmental raster layers covering the geographic region of interest for each desired time period, whether those are future or past climate scenarios. However, this package can incorporate many other types of provided data. First, users may provide their own species occurrence data instead of using megaSDM's built-in OccurrenceCollection() function to download species observations. Dispersal rates for each species (in km per year) may also be provided, for the dispersalRate(), createTimeMaps() and createRichnessMaps() functions to incorporate dispersal limitations into the model predictions generated by the MaxEntProj() function. Users may also provide background points, shapefiles for each species outlining a portion of the training map for selecting background points, or even SDM-generated habitat suitability maps created elsewhere for statistical analysis and map generation. The documentation of each function and the example workflow provided as a vignette with the download of the package give instructions on how to format all optional data. Refer to these guides and the documentation provided with the package functions for more detailed description of input data options.

Model flow and example

Most of the functions in megaSDM can be used to conduct single, stand-alone analyses (e.g. performing environmental subsampling of occurrences or evaluating the effects of dispersal limitations on projected suitable habitat). Alternatively, these functions can be linked together easily to create an entire, self-contained workflow (Fig. 1; Supporting information). To demonstrate the flexibility and functionality of megaSDM and to discuss the workflow of the functions, we have applied megaSDM to a set of test data (Fig. 5). These test data consist of a list of 165 native North American mammals and a set of bioclimatic variables, downloaded from WorldClim 1.4 <www.worldclim.org/version1> (Fick and Hijmans 2017, WorldClim 2.0). All examples highlighting a single species use the results for *Poliocitellus franklinii* (Franklin's ground squirrel; Fig. 2, 3). Figure 1 provides a simplified flowchart of the main functions in this package (a more detailed flowchart may be found in the Supporting information).

We have also provided a user example (Supporting information) that is installed with the package as a vignette and found on GitHub (<https://github.com/brshipley/megaSDM/megaSDM_vignette.html>). The user example can be run after setting the working directory at the beginning of the provided script. Using five mammal species and one subspecies that reside in the southeast United States, this example quickly demonstrates many of the features of megaSDM and the inputs necessary to run each function.

Data gathering

GIS environmental layer manipulation TrainStudyEnv() and PredictEnv()

The TrainStudyEnv() and PredictEnv() functions manipulate the input environmental data. These data provide the independent variables used to generate a relationship between species occurrence and the environmental/climatic factors. TrainStudyEnv() re-projects, clips and resamples the current environmental data, resulting in environmental layers with consistent projection, resolution and extent. Similarly, PredictEnv() takes the forecasted/hindcasted environmental rasters and projects, clips and resamples them to the parameters of the current data.

Species occurrences OccurrenceCollection()

The OccurrenceCollection() function acts as a wrapper for the occ_search() function in the rgbif package (Chamberlain et al. 2019), making the function more efficient for a large number of species. OccurrenceCollection() allows the user to directly download species occurrence date from the Global Biodiversity Information Facility (GBIF) <www.gbif.org>. Although we suggest that users carefully vet all occurrence information used in SDM analyses, this step can be useful for preliminary analyses or for educational and training purposes. Users can also directly input a table of species occurrence data for use with the remainder of the functions.

Table 1. A list and description of the stand-alone functions that are contained within the megaSDM package.

Function	Description
TrainStudyEnv	*Project/clip training and study environmental layers:* projects, clips and resamples environmental layers for the training area (i.e. where the model will be trained) and study area (i.e. where the parameters of the model will be applied and habitat suitability will be predicted) of an SDM analysis.
PredictEnv	*Project, clip and store forecasted/hindcasted environmental rasters for SDM prediction:* takes lists of RasterStacks that correspond to future or past time periods of a single climate model (e.g. RCP4.5, CCSM3), ensures that the environmental variables are the same as those that the model will be trained on, and projects, clips and resamples these layers to the characteristics of a given study region.
OccurrenceCollection	*Download and vet GBIF occurrence data:* takes a list of species and collects occurrence data from GBIF (Global Biodiversity Information Facility, <www.gbif.org>). Acts as a wrapper for rgbif::occ_search; however, this function is more efficient for a large number of species. It also checks the taxonomy of the given species list against the GBIF taxonomy, renaming or merging taxa if necessary. Furthermore, this function vets the occurrence data, removing occurrence points that are of insufficient quality for species distribution modelling. For a full list of issues removed by this package, refer to Supporting information. Further vetting may be done by the user. Finally, it provides the number of occurrences found within given training and study areas.
OccurrenceManagement	*Manage and environmentally filter occurrence points:* takes a set of occurrence points (whether downloaded from GBIF or provided), standardizes the column headings for effective use in species distribution modelling, and, if requested, extracts the values of each environmental variable used in the modelling for each occurrence point and environmentally subsamples the data (Varela et al. 2014).
BackgroundBuffers	*Create buffers for spatially constrained background point generation:* takes a list of occurrence point files and generates buffer shapefiles around each set of points. These buffers will be used if spatially constrained background points are required. The radius of the buffer can be defined as a single value for all species or as a distinct value for each species. If no radius values are given, the distances between the occurrence points themselves inform the buffer radius.
BackgroundPoints	*Generate background points for species distribution modelling:* generates a set of species-specific background points using one of several methods. These points can be randomly generated across a given training area, or if environmental data are provided, environmental subsampling (sensu Varela et al. 2014) can be conducted. If a list of buffers around the occurrence points of each species are provided, this function will conduct spatially constrained sampling within the buffer.
VariableEnv	*Use species-specific sets of environmental data for SDMs:* using environmental variables that are specific to each species can help to make more informative species distribution models. This function prepares Maxent inputs for the modelling of each species based upon a unique subset of the environmental variables.
MaxEntModel	*Model species distributions with MaxEnt using parallel processing:* takes occurrence points and background points of many species and models them using the MaxEnt algorithm, parallelizing the process across multiple computer cores.
nullAUC	*Generate null distribution models for AUC comparison:* one way to use AUC values to examine presence-only model predictions is to generate model replicates using randomly generated occurrence data and evaluating their performance using a subset of the real occurrence data. This function generates null models and calculates the test AUC values when applied to the subset of real occurrence data for comparison with the model training on the actual data. This method was developed by Bohl et al. (2019).
MaxEntProj	*Construct ensemble models and project habitat suitability to current, past and future climates:* conducts ensemble modelling on all replicates of the MaxEnt model by calculating the median habitat suitability for each pixel across all replicates. Next, the function generates binary presence/absence maps by applying a given threshold to the data. These processes are repeated for each scenario/time period combination provided.
createTimeMaps	*Create maps describing species range shifts across many time periods:* creates maps describing species range shifts across multiple time periods. These maps detail the step-wise expansions and contractions of the species distribution through those time-steps, allowing for the visualization of both unidirectional range shifts and more complex dynamics (e.g. a range expansion followed by a range contraction).
additionalStats	*Generate other statistics for species range shifts:* generates graphs showing the changes in range size and position between the data the model was trained on and future or past projections of species ranges.
dispersalRate	*Constrain modelled species distributions by dispersal rate:* incorporates the ability of a species to disperse over time into projected habitat suitability models and presence/absence maps. The probability of dispersal per year as a function of distance is modelled using an exponential distribution, and summed together to create a probability of dispersal for the intervals between each provided time step. Dispersal-constrained binary (presence and absence) maps are generated, as well as continuous maps of 'invadable suitability'
createRichnessMaps	*Create regular and dispersal-constrained richness maps from stacked SDMs:* This function stacks binary (presence/absence) species distribution maps to create richness maps for a list of species. If higher taxa are provided, separate richness maps for each higher taxon will be created in addition to the full species richness maps. Given hindcasted/forecasted binary maps, future/past species richness will also be calculated. Finally, provided distribution maps that are constrained by dispersal rate, compares between the dispersal-constrained and regular richness maps.

If the user wishes to use the package to download species occurrence data, they must provide a list of species and the geographic extent from which the occurrence points will be downloaded. The function will then download occurrence points of each species to be analysed from GBIF, the world's largest digital repository of biodiversity information (Telenius 2011). However, GBIF often contains incomplete or inaccurate results for some species (Beck et al. 2013), which may lead to inaccurate distribution models in those regions (Ferro and Flick 2015). This package increases the overall quality of the downloaded occurrence data by filtering out lower-quality data (e.g. duplicate observations, occurrence points with an improper datum conversion, rounded latitude/longitude coordinates). This is an additional improvement over simply using the rgbif::occ_search() function, providing a more rigorous assessment of where each species has been observed. For a complete list of GBIF error codes filtered out by this package, see Supporting information. If users provide their own occurrence dataset, they can bypass the OccurrenceCollection() function.

Data preparation

Environmental subsampling OccurrenceManagement()

Study design, sampling constraints and observer errors invariably lead to biases in occurrence data (Boakes et al. 2010). Biased input data decrease the overall accuracy of SDMs (Phillips et al. 2009, Beck et al. 2013, Varela et al. 2014). Therefore, the biases implicit in the collected or downloaded occurrence data must be accounted for (Phillips et al. 2009). The OccurrenceManagement() function in `megaSDM`, employs a method developed by Varela et al. (2014) and modified by Castellanos et al. (2019) to mitigate environmental and spatial biases within the occurrences by environmentally filtering the occurrence data. First, the environmental values for each occurrence point are divided into a desired number of bins such that the total number of bins possible is (nbins) ^ (number of environmental variables), although in practice many of these bins will be empty (Castellanos et al. 2019). This method can either be implemented on the raw data or conducted using a scaled principal component analysis (PCA) on the climatic values from each occurrence point, which has been demonstrated to perform better than environmentally filtering with unscaled data (Castellanos et al. 2019). A desired number of the PC axes are then designated as a parameter in the function or, if not given, the package will include PC axes until more than 95% of the climatic variance is explained. If categorical environmental variables are supplied, the PCA is not conducted and the data are subsampled using their original values.

Next, a single occurrence point from each n-dimensional bin is extracted for use in subsequent steps, resulting in a subset of occurrence points filtered by environment. This method of filtering allows for the removal of environmentally/climatically clustered or oversampled records while maintaining the total range of environments a species was found in Varela et al. (2014). Furthermore, this method does not require a priori knowledge of sampling effort. Models applying this method significantly outperform those using both random subsampling and geographic/spatial filtering (Fourcade et al. 2014, Varela et al. 2014, Castellanos et al. 2019).

Background sampling BackgroundBuffers(), BackgroundPoints()

There are two well-established strategies for selecting background points for SDMs (Barbet-Massin et al. 2012). The first method (hereafter called the 'random' method) involves randomly selecting background points throughout the entire area of interest (Fig. 4a). Although this method is simple and easy to implement, if the occurrence data are spatially biased

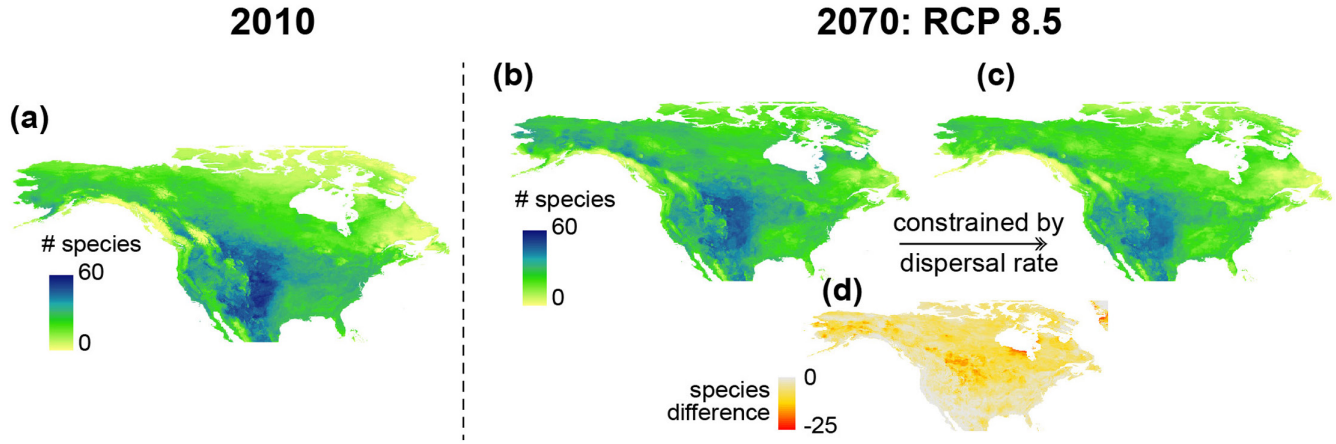

Figure 5. Species richness maps generated from SDMs of 165 North American mammal species for 2010 (a) and the RCP 8.5 climate scenario for 2070 (b–c; Riahi et al. 2011). For (b), dispersal ability is not considered. For (c), range expansions are constrained by average dispersal rates (km year^{-1}), calculated by Schloss et al. (2012), and distance from the species distribution in 2010. Map (d) shows the difference between the dispersal-applied richness map (b) and the map for which dispersal rate is not applied (c). The list of species modelled may be found in the Supporting information.

(e.g. greater densities of occurrence points in more easily accessed areas), the model may overestimate the environmental suitability of those regions (Lobo et al. 2010, Kramer-Schadt et al. 2013). To counteract this spatial bias, a second method (hereafter, 'spatially-constrained') only generates background points that are located within a certain buffer distance around the occurrence points (Fig. 4b), allowing the spatial bias of the set of background points to mirror that of the occurrence points themselves. The 'spatially-constrained' method decreases the effects of spatial bias in occurrence sampling and can increase model accuracy over the first method (Barve et al. 2011, Fourcade et al. 2014). However, when modelling multiple species, distributed across a large study area, this method is susceptible to extreme extrapolation errors and overfitting (Radosavljevic and Anderson 2014).

The BackgroundPoints() function in `megaSDM` allows users to apply either strategy for selecting background points, in addition to a new, 'combined' method that attempts to minimize the error in both methods. The 'combined' method samples a desired (user-defined) proportion of the background points from within a buffer around the occurrences and the rest of the background points from the entire study area (including the buffered area; Fig. 4c). This provides a de facto spatial weighting scheme, operating similarly to more widespread methods of mitigating spatial sampling bias through background point generation (Kramer-Schadt et al. 2013, Senay et al. 2013, Fourcade et al. 2014). Although anecdotal evidence suggests that it may decrease extrapolation errors in models of small species ranges in a large study area, this particular spatial weighting scheme is still experimental, and no rigorous tests have yet been conducted.

For the creation of 'spatially-constrained' or 'combined' background sampling schemes, `megaSDM` can either use buffers input by the users or it can create buffers. The BackgroundBuffers() function generates a buffer around each occurrence point. The radius of these buffers can be manually defined or be proportional to the 95% quantile of the distance to nearest neighbour for each point, therefore excluding outlier points (Fig. 4b–c).

Although the most appropriate method for generating background points is still a matter of discussion (Barbet-Massin et al. 2012, Senay et al. 2013, Liu et al. 2019), background points should be generated with similar biases (or lack thereof) as the occurrence points used in the model (Phillips et al. 2009). Therefore, if environmental filtering was conducted on the occurrence points (removing environmental biases), the background points within and outside the buffers should also be environmentally filtered, creating an even spread across available environmental space while retaining the spatial weighting.

Because each species is likely to have different environmental requirements, using a species-specific subset of environmental layers can additionally increase SDM accuracy and predictive ability (Elith et al. 2006, Austin and Niel 2011). This is often time-consuming to do manually (particularly when modelling many species at once), so the [VariableEnv] function provides a way to automate such a process. In this function, the user designates which environmental variables should be used for each species as an argument.

Modelling

MaxEnt modelling, subsampling and replication MaxEntModel()

After generating environmentally filtered and subsampled occurrence and background points for each species, the MaxEntModel() function in `megaSDM` estimates habitat suitability using the MaxEnt modelling technique, which applies maximum entropy methods and machine learning to produce estimates of habitat suitability and distribution (Phillips et al. 2006). MaxEnt has consistently exhibited high accuracy in a variety of species distribution modelling tasks, regularly outperforming other SDM techniques (Elith et al. 2006, Phillips and Dudík 2008, Feng et al. 2019). Many researchers ensemble the results of several different modelling methods (e.g. generalized linear models, random forests), creating a consensus model. However, each additional method used introduces different types of error and uncertainty into the consensus model (Elith et al. 2010). Therefore, rather than aggregating multiple methods, this package relies solely on MaxEnt. However, the package is able to conduct all relevant analyses with habitat suitability models that were not derived from MaxEnt specifically (e.g. the consensus model outputs from the 'ecospat' (Di Cola et al. 2017) or 'biomod2' (Thuiller et al. 2019) R packages).

Because parameter tuning is essential for accurate species distribution modelling in MaxEnt (Radosavljevic and Anderson 2014), this function allows for the manipulation of several MaxEnt parameters including regularization (penalizing complex models) and which features should be used to construct the models. To generate SDMs that are statistically rigorous, `megaSDM` allows replication with subsequent ensembling. During the modelling, the MaxEnt program can hold back a random subset of the occurrence data to be used for model evaluation and validation. This replication can be conducted multiple times for each set of validation occurrence data. Alternatively, spatial cross-validation can be conducted by a priori defining a set of validation points as an argument to the MaxEntModel() function.

Outputs

Continuous and binary distribution mapping nullAUC(), MaxEntProj()

After all replicates of the species have been modelled, `megaSDM` allows for a few strategies for evaluating each model replicate. The validation AUC values for each replicate are calculated in MaxEntModel(). A high validation AUC value generally indicates that the model is able to discern background records from true occurrences, and AUC values are commonly used in validating SDMs (Marmion et al. 2009). However, absolute comparisons between the AUC values of presence-background models such as MaxEnt may be untenable, because AUC values are highly influenced by factors

such as the geographical extent of the model and the proportion of presences to background points (Lobo et al. 2008, Jiménez-Valverde 2012). They instead must be compared to null models, where multiple replicates of occurrence points are randomly placed throughout the training area and evaluated on either the same set of cross-validation folds used for the MaxEntModel() function (Bohl et al. 2019) or, if 'testsamples' is not given as an argument, to a random subset of the null data (Raes and ter Steege 2007). It can then be determined whether a given model has a higher validation AUC than some percentage (i.e. 95%) of the validation AUCs calculated for the null models. This method can be conducted easily in `megaSDM`, using the nullAUC() function. Once these species-specific thresholds are defined (if requested), the MaxEntProj() function removes model replicates that contain a validation AUC value lower than a desired threshold.

After model evaluation, MaxEntProj() projects all models onto environmental rasters of all time periods and climate scenarios, and ensembles all replicate maps by taking the median value of each pixel. A median consensus model is always more accurate than at least 50% of the replicates, and median ensembling reduces the effect of outliers (Araújo and New 2007).

`megaSDM`'s MaxEntProj() function can create individual species maps comprised of continuous habitat suitability values or of binary habitat suitability maps, indicating that the species is predicted to be present or absent from each raster pixel for each provided time period and climate scenario. To create the binary maps, `megaSDM` applies a threshold to the continuous habitat suitability data given, creating a binary distribution map of locations where a species is anticipated to be either present (suitability $\geq$ threshold) or absent (suitability $<$ threshold). Although the choice of threshold can dramatically affect the accuracy of the model (Norris 2014), binary distribution models are often necessary for examining dispersal rate and conducting areal statistics. `megaSDM` can implement several commonly used threshold values. However, the default threshold is the 'maximum test sensitivity and specificity' logistic threshold, which attempts to maximize both specificity and sensitivity of the receiver operating curve generated by MaxEnt. This threshold is particularly effective at generating binary maps for presence-only data, and models applying this technique consistently outperformed models using other provided thresholds (Liu et al. 2015).

Time maps createTimeMaps()

If the user includes data for multiple time steps, `megaSDM` can generate maps for each species and climate scenario, detailing the step-wise expansions and contractions of the species distribution through those time-steps (Fig. 2). Studies using SDMs to predict future species ranges predominantly assume that the range shifts associated with future climate dynamics will be unidirectional (Bennett et al. 2019, He et al. 2019). However, unidirectional range shifts are not always observed because of non-uniform changes in atmospheric and ocean circulation, oscillations in radiative forcing (MacMartin et al. 2013) and spatial heterogeneity (Walther 2010, Terray 2012). These non-linear changes in climate result in transitory fluctuations in species ranges (e.g. intermittent expansions during a steady period of contraction) (Early and Sax 2011). Secondary range dynamics like this influence the habitat area that is functionally accessible to a species, as opposed to areas that have suitable habitat but are not accessible (i.e. regions that are not contiguous with the current species range) (Early and Sax 2011). Therefore, any examination of the actual ability of species to respond to long-term climatic changes must include these range shifts (Brown and Yoder 2015). The createTimeMaps() function combines the set of binary distribution maps generated across all time steps into a single raster file displaying the regions vacated and migrated to during each time period (Fig. 2). The value of each raster pixel is given as code describing presence (1) and absence (0) for each time step. For example, a code of '101' would suggest that a species vacated and then returned to that location (called 'wane-wax' in the PDF map files generated by this function). Individual raster maps for each time period are provided for the user to conduct additional analyses and desired (for example, generating more detailed spatial statistics on changes in range size or range centroid, or evaluating the coincidence and range overlaps of multiple species through time).

Dispersal rate dispersalRate()

If dispersal rate data are provided and added as a parameter in the additionalStats() and createTimeMaps() functions, those functions will repeat their analyses, applying a dispersal constraint on the projected range expansions from the binary presence–absence model generated for the first time-step. Many recent studies using SDMs to predict range shifts assume that species can colonize any suitable habitat, regardless of distance from the original range (Ureta et al. 2018). However, this assumption does not reflect how species ranges actually track climatic changes along the landscape (Schloss et al. 2012, Holloway et al. 2016). Incorporating species-specific dispersal rates into SDMs can provide a more accurate estimate of the habitats available for a species in the context of dynamic climates and habitat change (Barve et al. 2011, Uribe-Rivera et al. 2017).

The dispersalRate() function uses dispersal rate data provided by the user and applies exponential distributions to model the probability of dispersal as a function of distance per unit time. Although each species is likely to have a unique probability distribution, exponential distributions are commonly used to model dispersal across taxa, including the possibility of rare, distant dispersal events (Sutherland et al. 2000, Truvé and Lemel 2003, Nathan et al. 2012, Aparicio et al. 2018). This dispersal probability is then applied to the continuous habitat suitability maps, based on the distance away from the modelled binary species distribution at the first time-step. Suitable areas are therefore constrained by the ability of the species to expand its range, and creating an estimate of 'invadable suitability', incorporating both habitat suitability and the constraints of dispersal (Fig. 3). This new metric provides

information on the areas of suitable habitat that would be attainable by a range-shifting species without assisted migration. In addition, the dispersalRate() function multiplies the dispersal probability by the binary distribution maps, creating dispersal-constrained presence/absence distributions.

Additional statistics additionalStats()

When the binary distribution maps for each time period, climate scenario and species have been created (with dispersal constraints applied, if necessary), the additionalStats() function calculates summary statistics. First, the overall area of the binary distribution (in units of raster pixels) is calculated, and the function creates bar graphs of the binary range area across each provided time period and climate scenario. The percent change in range area is also calculated and displayed in a bar graph. Finally, additionalStats() calculates the centroid of the species range (the mean latitude and longitude of the occupied pixels) for each time and scenario. If dispersal rate has been applied, the additionalStats() function also compares the range sizes with and without dispersal constraints.

Species richness createRichnessMaps()

The final step of the workflow shown in the vignette is the createRichnessMaps() function, which uses generated (or user-provided) binary distribution maps to analyse changes in species richness through time. Understanding the spatial and temporal patterns and trends of species richness is a fundamental question in a wide range of scientific and public policy disciplines (Blackburn and Gaston 1996, Murphy and Romanuk 2014). Recently, researchers in these fields have begun to apply species distribution modelling to examine spatial trends in alpha (local) species richness (e.g. the R package SSDM, Schmitt et al. 2017). Species richness estimates are generated by stacking and summing the thresholded binary distribution models from many individual species (Ferrier and Guisan 2006, but see Calabrese et al. 2014, Scherrer et al. 2018, Del Toro et al. 2019). However, the effect of dispersal constraints on the temporal dynamics of species richness is still under-studied (Schloss et al. 2012). createRichnessMaps() sums the generated binary distribution models of all species for each time period and climate scenario, creating richness maps showing the responses of species richness to temporal climate changes (Fig. 5a–b). In addition, if the dispersal rate has been analysed for each species, this function generates dispersal-constrained richness maps (Fig. 5c) and calculates the differences between total dispersal and dispersal-constrained scenarios across all times and climate scenarios (Fig. 5d).

Example results

In our example, using 165 North American mammals, we ran through this entire workflow using the extent of North America as the training and the study areas. We environmentally filtered the occurrence and background points, generating 5000 background points using the 'combined' spatial weighting scheme. All other settings were set to the provided defaults (see the documentation of each function). Dispersal rate data were gathered from Schloss et al. (2012). Supporting our expectations and the results of Schloss et al. (2012), we found a predicted decline in overall species richness from 2010 to 2070 across North America, and a small but visible shift northward (Fig. 5a–b). Many species were unable to colonize all of the available suitable habitat in 2070, leading to marked discrepancies between the dispersal-constrained and the regular richness map (Fig. 5b–c). Dispersal limitations had the largest effect on richness in the north-western Great Plains and the lower Canadian Shield, concordant with the results of Schloss et al. (2012) (Fig. 5d).

Package installation and availability

This package, vignette and all related information are free and open-source under the MIT License and are available for download on GitHub (<https://github.com/brshipley/megaSDM/>). Further instructions on how to use `megaSDM`'s functions can be found by working through the vignette provided with the function (megaSDM_vignette, also found as an html file on GitHub). For details about the package dependencies of `megaSDM` and the citations of the package versions applied, see Supporting information and the description page for the package.

To cite megaSDM or acknowledge its use, cite this Software note as follows, substituting the version of the application that you used for 'version 1.0':

Shipley, B. R. et al. 2022. megaSDM: integrating dispersal and time-step analyses into species distribution models. – Ecography 2022: e05450 (ver. 1.0).

Acknowledgements – We thank Edward B. Davis for programming assistance. We also thank Manu Regalado, Julia Schap, Olivia Williams, Daniel Lauer and Valerie Washington for acting as beta testers. The BEE-Inspired REU Program and Jennifer Leavy provided assistance for several key undergraduate students, as did Georgia Tech's PURA Program.

Funding – Funding for HS was provided by the US Department of Agriculture, National Institute of Food and Agriculture grant no. 2016-67032-24986. JLM was partly funded by National Science Foundation Grants DEB-1655898 and SGP-1945013. BD was funded by National Science Foundation Award 1763108: Preserving Biodiversity via Robust Optimization. BD was also funded by National Science Foundation Award 009103: CHN2-S: Species conservation and collaborative governance in an era of global change.

Conflicts of interest – The authors declare no conflict of interest.

Author contributions

Benjamin R. Shipley: Methodology (supporting); Software (equal); Validation (lead); Visualization (equal); Writing – original draft (lead); Writing – review and editing (supporting). **Renee Bach**: Methodology (equal); Software (equal); Validation (supporting); Visualization (equal); Writing

– review and editing (supporting). **Younje Do**: Investigation (equal); Methodology (equal); Software (equal); Writing – review and editing (supporting). **Heather Strathearn**: Data curation (equal); Validation (equal); Visualization (equal); Writing – review and editing (supporting). **Jenny L. McGuire**: Conceptualization (lead); Funding acquisition (equal); Methodology (lead); Project administration (lead); Supervision (lead); Writing – review and editing (lead). **Bistra Dilkina**: Conceptualization (lead); Methodology (lead); Project administration (equal); Software (lead); Supervision (equal); Writing – review and editing (supporting).

Transparent Peer Review

The peer review history for this article is available at <https://publons.com/publon/10.1111/ecog.05450>.

Data availability statement

Data are available from the Github Digital Repository: <http://github.com/brshipley/megaSDM> (Shipley et al. 2021).

References

Aparicio, E. et al. 2018. Movements and dispersal of brown trout (*Salmo trutta* Linnaeus, 1758) in Mediterranean streams: influence of habitat and biotic factors. – PeerJ 6: e5730.

Araújo, M. B. and New, M. 2007. Ensemble forecasting of species distributions. – Trends Ecol. Evol. 22: 42–47.

Austin, M. P. and Niel, K. P. V. 2011. Improving species distribution models for climate change studies: variable selection and scale. – J. Biogeogr. 38: 1–8.

Barbet-Massin, M. et al. 2012. Selecting pseudo-absences for species distribution models: how, where and how many? – Methods Ecol. Evol. 3: 327–338.

Barve, N. et al. 2011. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. – Ecol. Model. 222: 1810–1819.

Beck, J. et al. 2013. Online solutions and the 'Wallacean shortfall': what does GBIF contribute to our knowledge of species' ranges? – Divers. Distrib. 19: 1043–1050.

Bennett, M. et al. 2019. Shifts in habitat suitability and the conservation status of the endangered Andean cat *Leopardus jacobita* under climate change scenarios. – Oryx 53: 356–367.

Blackburn, T. M. and Gaston, K. J. 1996. Spatial patterns in the species richness of birds in the new world. – Ecography 19: 369–376.

Boakes, E. H. et al. 2010. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. – PLoS Biol. 8: e1000385.

Bohl, C. L. et al. 2019. A new null model approach to quantify performance and significance for ecological niche models of species distributions. – J. Biogeogr. 46: 1101–1111.

Brown, J. L. and Yoder, A. D. 2015. Shifting ranges and conservation challenges for lemurs in the face of climate change. – Ecol. Evol. 5: 1131–1142.

Calabrese, J. M. et al. 2014. Stacking species distribution models and adjusting bias by linking them to macroecological models. – Global Ecol. Biogeogr. 23: 99–112.

Castellanos, A. A. et al. 2019. Environmental filtering improves ecological niche models across multiple scales. – Methods Ecol. Evol. 10: 481–492.

Chamberlain, S. et al. 2019. rgbif: interface to the Global Biodiversity Information Facility API. – R package ver. 1.3.0, <https://CRAN.R-project.org/package=rgbif>.

Cobos, M. E. et al. 2019. kuenm: an R package for detailed development of ecological niche models using Maxent. – PeerJ 7: e6281.

Del Toro, I. et al. 2019. Are stacked species distribution models accurate at predicting multiple levels of diversity along a rainfall gradient? – Austral Ecol. 44: 105–113.

Di Cola, V. et al. 2017. ecospat: an R package to support spatial analyses and modeling of species niches and distributions. – Ecography 40: 774–787.

Early, R. and Sax, D. F. 2011. Analysis of climate paths reveals potential limitations on species range shifts. – Ecol. Lett. 14: 1125–1133.

Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. – Ecography 29: 129–151.

Elith, J. et al. 2010. The art of modelling range-shifting species. – Methods Ecol. Evol. 1: 330–342.

Engler, R. et al. 2012. The MIGCLIM R package – seamless integration of dispersal constraints into projections of species distribution models. – Ecography 35: 872–878.

Feng, X. et al. 2019. Collinearity in ecological niche modeling: confusions and challenges. – Ecol. Evol. 9: 10365–10376.

Ferrier, S. and Guisan, A. 2006. Spatial modelling of biodiversity at the community level. – J. Appl. Ecol. 43: 393–404.

Ferro, M. L. and Flick, A. J. 2015. 'Collection bias' and the importance of natural history collections in species habitat modeling: a case study using *Thoracophorus costalis* Erichson (Coleoptera: Staphylinidae: Osoriinae), with a Critique of GBIF.org. – Coleopts. Bull. 69: 415–425.

Fick, S. E. and Hijmans, R. J. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. – Int. J. Climatol. 37: 4302–4315.

Fourcade, Y. et al. 2014. Mapping species distributions with MAXENT using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. – PLoS One 9: e97122.

GBIF.org 2019. GBIF home page. – <www.gbif.org>.

He, X. et al. 2019. Upward elevation and northwest range shifts for alpine *Meconopsis* species in the Himalaya–Hengduan Mountains region. – Ecol. Evol. 9: 4055–4064.

Hijmans, R. J. et al. 2017. dismo: species distribution modeling. – R package ver. 1.1-4. <https://CRAN.R-project.org/package=dismo>.

Holloway, P. et al. 2016. Incorporating movement in species distribution models: how do simulations of dispersal affect the accuracy and uncertainty of projections? – Int. J. Geogr. Inf. Sci. 30: 2050–2074.

Huang, J.-L. et al. 2020 Importance of spatio-temporal connectivity to maintain species experiencing range shifts. – Ecography 43: 1–13.

Jiménez-Valverde, A. 2012. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. – Global Ecol. Biogeogr. 21: 498–507.

Kramer-Schadt, S. et al. 2013. The importance of correcting for sampling bias in MaxEnt species distribution models. – Divers. Distrib. 19: 1366–1379.

Lehtomäki, J. et al. 2019. Spatial conservation prioritization for the East Asian islands: a balanced representation of multitaxon biogeography in a protected area network. – Divers. Distrib. 25: 414–429.
Liu, C. et al. 2015. On the selection of thresholds for predicting species occurrence with presence-only data. – Ecol. Evol. 6: 337–348.
Liu, C. et al. 2019. The effect of sample size on the accuracy of species distribution models: considering both presences and pseudo-absences or background sites. – Ecography 42: 535–548.
Lobo, J. M. et al. 2008. AUC: a misleading measure of the performance of predictive distribution models. – Global Ecol. Biogeogr. 17: 145–151.
Lobo, J. M. et al. 2010. The uncertain nature of absences and their importance in species distribution modelling. – Ecography 33: 103–114.
MacMartin, D. G. et al. 2013. Management of trade-offs in geoengineering through optimal choice of non-uniform radiative forcing. – Nat. Clim. Change 3: 365–368.
Marmion, M. et al. 2009. Evaluation of consensus methods in predictive species distribution modelling. – Divers. Distrib. 15: 59–69.
Murphy, G. E. P. and Romanuk, T. N. 2014. A meta-analysis of declines in local species richness from human disturbances. – Ecol. Evol. 4: 91–103.
Nathan, R. et al. 2012. Dispersal kernels: review – In: Clobert, J. et al. (eds), Dispersal ecology and evolution. Oxford Univ. Press, pp. 187–202.
Norris, D. 2014. Model thresholds are more important than presence location type: understanding the distribution of lowland tapir *Tapirus terrestris* in a continuous Atlantic forest of southeast Brazil. – Trop. Conserv. Sci. 7: 529–547.
Phillips, S. J. and Dudík, M. 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. – Ecography 31: 161–175.
Phillips, S. J. et al. 2006. Maximum entropy modeling of species geographic distributions. – Ecol. Model. 190: 231–259.
Phillips, S. J. et al. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. – Ecol. Appl. 19: 181–197.
Radosavljevic, A. and Anderson, R. P. 2014. Making better Maxent models of species distributions: complexity, overfitting and evaluation. – J. Biogeogr. 41: 629–643.
Raes, N. and ter Steege, H. 2007. A null-model for significance testing of presence-only species distribution models. – Ecography 30: 727–736.
Riahi, K. et al. 2011. RCP 8.5 – a scenario of comparatively high greenhouse gas emissions. – Clim. Change 109: 33.
Scherrer, D. et al. 2018. How to best threshold and validate stacked species assemblages? Community optimisation might hold the answer. – Methods Ecol. Evol. 9: 2155–2166.
Schloss, C. A. et al. 2012. Dispersal will limit ability of mammals to track climate change in the Western Hemisphere. – Proc. Natl Acad. Sci. USA 109: 8606–8611.
Schmitt, S. et al. 2017. SSDM: an r package to predict distribution of species richness and composition based on stacked species distribution models. – Methods Ecol. Evol. 8: 1795–1803.
Senay, S. D. et al. 2013. Novel three-step pseudo-absence selection technique for improved species distribution modelling. – PLoS One 8: e71218.
Shipley, B. R. et al. 2021. Data from: megaSDM: integrating dispersal and time-step analyses into species distribution models. – Github Digital Repository, <http://github.com/brshipley/megaSDM>.
Sutherland, G. et al. 2000. Scaling of natal dispersal distances in terrestrial birds and mammals. – Conserv. Ecol. 4: 16–51.
Telenius, A. 2011. Biodiversity information goes public: GBIF at your service. – Nord. J. Bot. 29: 378–381.
Terray, L. 2012. Evidence for multiple drivers of North Atlantic multi-decadal climate variability. – Geophys. Res. Lett. 39: L19712.
Thomson, A. M. et al. 2011. RCP4.5: a pathway for stabilization of radiative forcing by 2100. – Clim. Change 109: 77.
Thuiller, W., et al. 2019. biomod2: ensemble platform for species distribution modeling. – R package ver. 3.3-7.1. <https://CRAN.R-project.org/package=biomod2>.
Truvé, J. and Lemel, J. 2003. Timing and distance of natal dispersal for wild boar *Sus scrofa* in Sweden. – Wildl. Biol. 9: 51–57.
Ureta, C. et al. 2018. A first approach to evaluate the vulnerability of islands' vertebrates to climate change in Mexico. – Atmósfera 31: 221–254.
Uribe-Rivera, D. E. et al. 2017. Dispersal and extrapolation on the accuracy of temporal predictions from distribution models for the Darwin's frog. – Ecol. Appl. 27: 1633–1645.
Varela, S. et al. 2014. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. – Ecography 37: 1084–1091.
Walther, G.-R. 2010. Community and ecosystem responses to recent climate change. – Phil. Trans. R. Soc. B 365: 2019–2024.