Supplement of

A protocol for a biodiversity model intercomparison project: BMIP 1.0

Section S1: Protocol of different modeling approaches used in BMIP 1.

Species Distribution Model (SDM)

This modeling approach correlates species presences and absences (or pseudo-absences) with environmental predictor variables to estimate species suitability over space and time (Franklin, 2010). The species presence data were sourced from the Global Biodiversity Information Facility (www.gbif.org), combined with location presence data from the prepared regional species abundance data. SDM performance and suitability patterns are sensitive to the approach and size of the training area (accessible or calibration areas) (Rojas-Soto et al., 2024). Consequently, a training area must be delimited based on species presence (e.g., minimum convex hull plus a buffer around it or based on biogeographical entities). It is advisable to train models using a calibration approach at a yearly resolution, ensuring that the year of occurrence matches the year of the environmental data. Occurrence data often suffer from geographical and temporal sampling biases (Beck et al., 2014; Boakes et al., 2010), and geographical or temporal thinning can be applied to reduce these biases. Geographical sampling bias correction can be performed based on a geographical or environmental thinning approach (Aiello-Lammens et al., 2015; Varela et al., 2014). Temporal bias correction could be based on temporal thinning (Dobson et al., 2023). In the case of accounting for a database solely based on presences (i.e., without absences), pseudo-absences and background points must be sampled (Barbet-Massin et al., 2010). To temporally distribute pseudo-absences and background points, they should be sampled across periods (i.e., days, years, decades) corresponding to species presence records (Leão et al., *manuscript under review*). Before model calibration, multicollinearity among predictors must be evaluated and addressed (e.g., correlation matrix, linear dimensionality reduction, or univariate variable importance) (Dormann et al., 2013). For model calibration, different modeling algorithms (e.g., Random Forest, Maximum Entropy, Generalized Linear Models) must be used to capture the full range of model variability (Thuiller et al., 2019). Hyperparameter tuning is used to select the optimal values for each algorithm's

hyperparameters, i.e., the values that deliver the highest model performance. Model transferability is evaluated using a structured data partitioning approach. For example, this could be achieved using an environmental, geographic, or temporal approach (Roberts et al., 2017). At least two different model metrics must be used: one threshold-dependent and one threshold-independent. Finally, to reduce model uncertainty, ensemble modeling approaches can be used (e.g., by averaging model suitability maps generated by different algorithms). Such an ensemble must be based on the model sets that meet a minimum performance (Rose et al., 2023). Since SDMs can predict suitable areas that are far from the actual species distribution (i.e., model overprediction), the output of an SDM must be constrained to regions occupied by the species. These constraints can be applied using different approaches (see Mendes et al., 2020; Velazco et al., 2020).

Abundance-based species distribution models (ASDM)

This modeling approach correlates species abundance or density data with environmental variables to estimate abundance (or density) maps (de Oliveira Junior and Velazco, 2025; Waldock et al., 2022). Abundance of individuals may be ecologically more relevant than presence-only models, as species occurrence or habitat suitability does not always correlate linearly with abundance; however, datasets are often better suited to modeling species occurrence. Similar to SDMs, the size of the training area can affect these models because it geographically constrains the location where a species was not detected (De Oliveira Junior and Velazco, *manuscript under review*). Therefore, for an abundance data survey in geographically extensive areas, a training area must be delimited. Such delimitation can be based on the locations where a species was detected (density of count data > 0) (De Oliveira Junior and Velazco, *manuscript under review*). The data must contain absence data, which can be balanced to enhance the model's performance and prevent underfitting of the abundance data. Models are constructed based on a time-calibrated approach, matching species abundance to the year of the environmental data. Multicollinearity must be addressed before model calibration (e.g., correlation matrix, linear dimensionality reduction, or univariate variable importance). For model calibration, various algorithms should be used (e.g., Random Forest, Gradient Boosting Machines, Generalized Linear Models, Deep Neural Networks, Convolutional Neural Networks), and the highest performing model should be chosen. Hyperparameter tuning is used to select the optimal hyperparameters for each algorithm, i.e., the values that deliver the highest model performance. Spatially,

environmentally, or time-structured data partitioning should be used to evaluate model transferability (Santini et al., 2021). Model performance metrics must be based on the combination of Accuracy (e.g., Mean Absolute Error, Root Mean Squared Error), Discrimination (e.g., Pearson's correlation), and Precision (e.g., standard deviations) (Waldock et al., 2022) (Table 2 and Apendix B). The algorithm with the highest performance is used to deliver spatial abundance predictions.

Species exposure models

This approach quantifies species' exposure to unprecedented climate conditions by comparing recent extremes to historical baselines inferred from species' geographic distributions (Serra-Diaz et al., 2026). Exposure is evaluated across hot, cold, wet, and dry extremes, considering both annual and seasonal values to capture relevant variability in different periods. Extremes are defined as high- or low-quantiles of climate variables, following established methods, rather than absolute thresholds, allowing consistency across taxa and regions. Climate history is estimated in two steps. First, the 99% quantile (or 1% for cold/dry extremes) of the climate variable across years is calculated for each spatial unit to characterize local extreme conditions while limiting sensitivity to outliers. Second, for each species, a "niche quantile" is calculated across all occupied units, defining an upper or lower bound of historically experienced conditions while reducing sensitivity to spatial uncertainty or marginal populations. Exposure occurs when the contemporary climate exceeds these species-specific limits (Serra-Diaz et al., 2026). Species distributions are compiled from range maps or habitat-based estimates. These data provide standardized representations of geographic ranges but may include spatial uncertainty and do not capture temporal shifts. Distributions are treated as static over the study period, and derived range metrics are used only for relative comparisons, not to infer absolute risk thresholds. Exposure is updated iteratively as new data becomes available (e.g., the estimate for a given year incorporates all prior years in the record). This approach identifies conditions that are unprecedented relative to past experience without making assumptions about persistence or demographic consequences. Exposure is summarized using thresholds of geographic extent within species' ranges. Exposure is considered non-trivial when ≥1% of a range exceeds the climate limit, and high exposure when ≥25% of a range is affected. These thresholds balance interpretability and comparability with prior work while avoiding overemphasis on minimal spatial exposure. This correlative framework reflects the combined influence of processes

shaping species' realized climatic niches, including physiological limits, biotic interactions, and dispersal constraints. However, exposure estimates are sensitive to the spatial resolution of climate data, often overlooking local climate heterogeneity and climate refugia (Colwell, 2021). Exposure estimates should therefore be interpreted as a tool for prioritizing monitoring and further investigation rather than as direct evidence of biological impact.

Dispersal models (MigClim)

MigClim is a hybrid dispersal model that uses a cellular automaton to simulate dispersal within a landscape using a limited number of input parameters (Engler et al., 2012; Engler and Guisan, 2009). MigClim is coupled to SDMs by simulating dispersal either on binary maps or probabilistic suitability maps predicted by SDMs. Required input parameters describe the dispersal kernel, the frequency of long-distance dispersal, dispersal barriers, and propagule production (establishment) potential and maturity age (Engler et al., 2012), while the minimal set of parameters includes the dispersal rate, dispersal distance, and kernel. These parameters are defined based on estimates from the literature and run with as many relevant parameters as are available in the literature. MigClim simulation is run with an annual temporal resolution. Model outputs a presence/absence map of colonized cells per replicate simulation. Maps of predicted colonization probability are derived by averaging across replicates.

Metabolically explicit metapopulation range model (metaRange)

The metaRange framework creates a spatially and temporally explicit meta-population model tailored for a target species (Fallert et al., 2025). Based on the available inputs, a spatial resolution of 1 km^2 (SDM input and environmental data input) is used, with each grid cell representing one population. The temporal resolution of the simulation is annual. Trait values (body mass, mean dispersal distance, demographic rates, and carrying capacity) should be either estimated from abundance data (demographic and dispersal traits, such as mean dispersal distance, demographic rates, and carrying capacity) or obtained through a literature search (all traits). The carrying capacity is estimated using the maximum measured abundance, which is then scaled up to the grid cell size and multiplied by observation-dependent factors to estimate the actual abundance in a grid. Also from the literature, reproductive and mortality rates can be calibrated using the mean values of reported population transition matrices (e.g., matrices from the COMPADRE and

COMADRE trait databases; Salguero-Gómez et al., 2015, 2016). If the focus species has not been investigated, closely related species (i.e., from the same genus and biogeographical region) might be used to infer trait values. If no reproductive rate is available, metabolic scaling based on the metabolic theory of ecology should be applied (Brown et al., 2004). This method uses known species to calibrate the reproduction-specific metabolic constant. The mean and max dispersal distances can be estimated from the abundance points and other information (e.g., routes that were surveyed in the Breeding Bird Survey), using ABC plus pattern-oriented modeling to derive credible ranges of demographic and dispersal parameters from the time series of occurrence data (Malchow et al., 2024), or using a general Bayesian calibration (i.e., using a likelihood function) to derive a likelihood from the distribution data. Each habitable cell is initialized in a 10 km radius around the occurrence points with an abundance value based on the maximal carrying capacity, relative to the calculated suitability of the specific habitat. To calculate the habitat suitability, first, an estimation of the species' environmental niche is performed by extracting the environmental values at known species points. A density estimation is calculated, from which the mode (representing the optimal niche value) and the 1% and 99% quantiles (representing the minimum and maximum values, respectively) are used as the defining parameters of a beta distribution. This distribution represents the species niche and is subsequently applied to the landscape values to generate suitability maps. Only environmental variables known from the literature to impact the species mechanistically should be used. These variables are used to calculate habitat suitability by matching the species niche with the local grid cell environmental values, followed by normalization against the maximum value to rescale all outputs between 0 and 1. The suitability values modulate the reproduction rate and the carrying capacity by multiplying them with the normalized suitability. Reproduction and individual survival are simulated by a time-discrete version of the Beverton-Holt equation, which uses the suitability-modulated demographic parameters. To simulate dispersal, the mean dispersal distance is used as the single parameter for calculating a negative exponential dispersal kernel for simplicity. This kernel is adjusted based on the surroundings of each population, such that the probability of dispersal within the kernel is a product of the negative exponential function and the relative suitability of the habitat in the kernel. The incoming individuals from dispersal at the end of the time step are used as the lambda of a Poisson distribution to set the recruited abundance. This recruited abundance is then added to the surviving adult and juvenile individuals. This represents the typical stochastic effects that are expected through

environmental noise and survival probabilities. Time series maps are generated for the abundance, reproduction rate, and carrying capacity of each population.

Dynamic occupancy models (DOM)

DOMs explicitly model changes in the occupancy state of sites across time, as a function of local extinction and colonization processes (MacKenzie et al., 2003; Royle and Kéry, 2007). They use binary detection/non-detection data and can account for imperfect detection provided multiple detection/non-detection records are available at a sufficient subset of sites, to estimate species detectability. To calibrate DOMs to the available datasets, data in different formats (e.g., counts) may be processed to obtain detection/non-detection records at each sampling occasion, resulting in a series of detection histories for each sampling site. The three ecological components of the model (initial occupancy, colonization, extinction/persistence) are parameterized as a function of environmental covariates, considering quadratic relationships and covariate interactions where relevant. Temporal and spatial effects can also be considered (Diana et al., 2023; Rushing et al., 2019). Before model calibration, continuous environmental covariates are standardized and assessed for multicollinearity to avoid including heavily correlated predictors in the same model. Where data about sampling effort, or other covariates potentially relevant for detectability, are available, they are included as predictors of detection probability. A set of plausible models for the different processes is defined and fitted to the data. Datasets can be split into training and validation sets, and the best-performing model selected based on predictive performance on the left-out (validation) data, based on metrics such as likelihood/deviance (Gelman et al., 2003) and AUC (Area Under the Curve of a Receiver Operating Characteristic) (Zipkin et al., 2012). Regularization techniques can also be considered. Once a final model is identified, it is recalibrated with the whole dataset. Models are fitted in a Bayesian framework using dedicated software, such as Stan or JAGS.

Spatially and temporally explicit population simulations (STEPS)

STEPS is a platform for spatially and temporally explicit matrix modeling of individual species populations (Visintin et al. 2020). STEPS brings together spatial models of environmental change with matrix models of population change in a highly modularized environment that allows inclusion (or exclusion) of any aspects of environmental change that

influence habitat suitability, carrying capacity, and population vital rates (fecundity, survival, dispersal). Its modular design allows users to tailor model complexity by selecting among alternative formulations for environmental change that influence demographic processes. As spatial input, the model requires a habitat map at a minimum and can additionally incorporate layers that modify in time and space habitat quality, dispersal, or spatial variation in demographic rates. Users also provide species-specific demographic parameters, density-dependence, and dispersal functions. STEPS implementation follows a hybrid approach in which outputs from occurrence-based SDMs provide annual habitat suitability surfaces for each species. Land use may influence habitat suitability and therefore carrying capacity, survival, dispersal, and fecundity. Land-use change maps can be input directly to STEPS to modify carrying capacity and demographic parameters, or via the habitat suitability surface. Likewise, climate metrics can be used to directly alter demographic parameters across space and time or within the habitat suitability surface. Population dynamics are represented with stage-structured populations, while dispersal is modeled using either a cellular automata model or a simple dispersal kernel, depending on available information and the spatial scale of the particular case study. Spatial resolution follows the finest grid used by the SDMs for dispersal processes and to model environmental change. In contrast, population dynamics are simulated using a cell-based approach with the cell size scaled to model populations at the finest scale without including inflated demographic stochasticity. Simulations are run on an annual timestep. Initial parameterization for stage structure, demographic rates, dispersal distances, and average densities is drawn from available literature for each species. Bayesian model calibration (sensu Malchow et al., 2024) may be explored; however, landscape and population simulations are run by default using the time step immediately preceding the first year of prediction to define the starting population and the spatial variation in habitat suitability. Outputs include a time series raster stack of mean projected (simulated) annual abundance, with another stack indicating variance (across simulations). Overall population and carrying capacity change trajectories and maps of continuous population occupancy can also be provided.

Age-structured dynamic range model (DRM)

DRM (Da Cunha Godoy et al., 2026; Fredston et al., 2025) models the spatio-temporal distribution of a population (density or biomass) based on biological processes (reproduction, survival, and dispersal). Structural decisions regarding factors such as the

number of assumed age classes and external sources of mortality are made based on existing literature before model calibration. The model is calibrated to observations of biomass or abundance across space or time, assuming a zero-augmented Gamma distribution for the response variable (Ye et al., 2021). The model structure estimates demographic processes (such as recruitment or survival rates) as functions of environmental conditions (such as temperature or precipitation) to explain species distributions and changes through time. The specific covariates and functional forms are selected using cross-validation and sparsity-inducing priors (Piironen and Vehtari, 2017). The number of estimated parameters is moderate, and because there are very few latent variables, this model can be conceptualized as an SDM with a biologically mechanistic link function. Given its structure, the model easily makes out-of-sample predictions. The model is implemented in a Bayesian framework where samples from the posterior distributions conditioned on the observed data are obtained using Stan. The remaining parameters are assigned weakly informative priors to facilitate algorithm convergence.

RangeShifter

RangeShifter is a spatially explicit, individual-based eco-evolutionary modeling platform (Bocedi et al. 2014, Bocedi et al. 2021) that simulates species-specific population dynamics, dispersal, and, optionally, genetic processes within a given landscape. Its modular design allows users to tailor model complexity by selecting among alternative formulations for demographic processes, density dependence, dispersal behavior, and genetics. As spatial input, the model requires a habitat map at a minimum and can additionally incorporate layers describing habitat quality, dispersal costs, or spatial variation in demographic rates. Users also provide species-specific demographic parameters, density-dependence functions, and (if relevant) genetic architecture. RangeShifter v3.0 via the R interface RangeShiftR should be used (Malchow et al., 2021). The BMIP implementation follows a hybrid approach in which outputs from occurrence-based SDMs provide annual habitat suitability maps for each species.

Population dynamics are represented with stage-structured populations, while dispersal is modeled using a correlated random walk that accommodates individual-level settlement decisions. Spatial resolution follows the finest grid used by the SDMs for dispersal processes. In contrast, population dynamics are simulated using either a cell-based or a patch-based approach, depending on expected mean population densities. For species with low

expected densities (e.g., species with large territories), neighboring cells are aggregated into patches, with minimum patch sizes selected to limit excessive demographic stochasticity. Initial parameterization for stage structure, demographic rates, dispersal distances, and average densities are drawn from the available literature. Model calibration is carried out within a Bayesian framework (Malchow et al., 2024), using the initial parameter values to define informative priors and a likelihood comparing simulated abundance or occupancy to observed time series for each species.

Ecophysiological models

Ecophysiological models encompass a broad class of mechanistic niche models based on the principles of biophysical ecology (Briscoe et al., 2023). They capture the balances of heat, water, and other aspects of energy and mass exchange between organisms and their microclimatic environment and translate these into metrics of performance or fundamental constraints on fitness, such as survival, development, growth, and reproduction, or more indirect measures of fitness (e.g., potential activity time) (Kearney and Porter, 2009). The general workflow is to use a microclimate model to scale environmental conditions to what organisms experience, followed by a biophysical model to balance the exchange of heat, water, or mass between organisms and their environment (Buckley et al., 2023a). The estimated physiological condition of organisms can be used as input for a statistical model or compared to physiological thresholds to estimate suitability. Other ecophysiological models use physiological estimates to inform additional models (e.g., demographic and evolutionary) that explicitly describe how organisms respond to their environments. The models require environmental data at the spatial and temporal scales relevant to organismal processes. Daily or hourly environmental data (e.g., air and soil temperature, solar radiation, wind speed) are often used as input and further temporally refined with models of diurnal variation (Meyer et al., 2023). Necessary parameters describing the environment include surface reflectivity and roughness. Trait data needed to parameterize the process models span morphology (e.g., body mass, shape, integument properties), physiology (e.g., thermal tolerances, preferred body temperatures, metabolic rates, water loss parameters), and behavior (microhabitat use, activity, phenology) (Kearney et al., 2021b). Several R packages implement microclimate models, some of which include comprehensive tools for accessing and processing environmental data (Microclima, Microclimc, NicheMapR) (Kearney and Porter, 2017; Maclean et al., 2019; Maclean and Klinges, 2021). The TrenchR package focuses on

transparent and modular microclimate and biophysical models (Buckley et al., 2023b), whereas the NicheMapR package includes more sophisticated and comprehensive microclimate and biophysical models (Kearney et al., 2021a; Kearney and Porter, 2020). Typical models used in plants, such as those applied in the TTR.PGM package, integrate water-energy dynamics to derive plant growth as a proxy for species suitability (Higgins et al., 2025). Most ecophysiological models are developed for particular organisms and locations based on knowledge of their physiological constraints. NicheMapR also includes Dynamic Energy Budget models (DEB) that provide a general approach to modeling energetic constraints (Kearney and Porter, 2020). Model outputs range from temporally aggregate estimates of physiological conditions to estimates of performance, fitness, or demography (Briscoe et al., 2023).

Mean simple model (MSM)

The mean simple model is used as a null benchmark (Harris et al., 2018). Abundance in each cell is predicted as the mean abundance over the past 10 years of the validation dataset, representing a stationary expectation where the mean, variance, and autocorrelation of the time series show no systematic change (i.e., no trend) (Gonzalez et al., 2023). This provides a simple baseline against which the skill of the more complex models can be evaluated and is used here to define the predictability limit and associated forecast horizon (Petchey et al., 2015).

Temporal Autocorrelation Model (TAM)

An alternative null is that historical trends continue. The TAM serves as a dynamic baseline benchmark that tests the null hypothesis that past local temporal trends will persist over time. Unlike static benchmarks that assume a constant stationary state, TAM extrapolates the historical linear trend independently for each spatial cell. For a given grid cell i, a linear regression model is fitted using historical time-series data to capture directional trajectories in species abundance or occurrence. This cell-specific relationship is then projected forward into future time steps. By incorporating localized linear drift, TAM accounts for persistent temporal shifts without relying on environmental predictor variables, providing an essential baseline to evaluate whether models offer superior predictive skill over simple extrapolation of past temporal trends.

Spatial Autocorrelation Model (SAM)

The SAM serves as a spatial null benchmark that assumes nearby location values are the primary driver of local species abundance or occurrence. SAM fits a localized spatial autocorrelation function to the observed data across the study grid. By estimating the spatial covariance structure from adjacent grid cells, the model generates predictions based solely on the geographic proximity and spatial clustering of historical species records. SAM evaluates how much performance metric can be explained purely through spatial contagion and geographic proximity.

Section S2: Equations of performance metrics listed in Table 2

Metrics for binary presence-absence output

Many performance metrics depend on a confusion matrix calculation. This is used to organize and visualize an algorithm's errors in classification problems. In binary classification tasks (e.g., SDMs), the rows typically represent the actual class or positive and negative detections of a class, while the columns represent the predicted values. For binary presence-absence models, where the objective is to classify each point as either present or absent, the matrix takes the following format:

	Predicted Presence (+)	Predicted Absence (-)
Actual Presence (+)	True Positives (TP)	False Negatives (FN)
Actual Absence (-)	False Positives (FP)	True Negatives (TN)

In that matrix, FP and FN are wrongly classified presences and absences; therefore, the goal of any algorithm is to maximize the counts of TP and TN, which are the correctly classified presences and absences, respectively. From this, it is possible to derive many of the most used metrics in binary prediction.

Overall accuracy

One of the simplest metrics is the Overall Accuracy, which expresses the proportion of correct predictions to the total number of predictions. Therefore, it can be written as

$$Overall\ Accuracy\ = \frac{TP + TN}{TP + FP + TN + FN}$$

346 *Sensitivity*

347 The Sensitivity (also True Positive Rate or Recall) expresses the proportion of
348 correctly classified presences (TP) to the total number of actual presences in the observations
349 (TP+FN), which is defined as

350 $$Sensitivity = \frac{TP}{TP + FN}$$

351

352 *Specificity*

353 Specificity (also True Negative Rate) expresses the proportion of correctly classified
354 absences (TN) to the total number of actual absences (TN+FP), written as

355 $$Specificity = \frac{TN}{TN + FP}$$

356

357 *True Skill Statistics (TSS)*

358 This metric combines Sensitivity and Specificity into a single metric, which consists
359 of a sum of the two metrics minus one, relocating the output into the interval [-1,1]:

360 $$TSS = Sensitivity + Specificity - 1$$

361

362 *Positive Predictive Value (PPV)*

363 PPV (also Precision) measures the proportion of correctly predicted presences (TP) to
364 the number of presences predicted (TP+FP):

365 $$PPV = \frac{TP}{TP + FP}$$

366

Negative Predictive Value (NPV)

NPV measures the proportion of correctly predicted absences (TN) to the number of predicted absences (TN+FN):

$$NPV = \frac{TN}{TN + FN}$$

F1-Score and Sørensen's similarity index

It is possible to combine PPV and Sensitivity into the F1-Score, which is the harmonic mean of them. Therefore, it is defined as:

$$F1 = \frac{2}{\frac{1}{PPV} + \frac{1}{Sensitivity}} = \frac{2 \cdot PPV \cdot Sensitivity}{PPV + Sensitivity} = \frac{2TP}{2TP + FP + FN}$$

Note that for binary data, F1-Score is equivalent to Sørensen's similarity index (or Dice-Sørensen coefficient). In the context of binary presence-absence models, the index is defined as

$$Sorensen = \frac{2|X \cap Y|}{|X| + |Y|}$$

Where: $|X|$ is the number of elements in the set X, which contains all values predicted as presences, thus $|X| = TP + FP$; $|Y|$ is the number of elements in the set Y, which contains all actual presences in the sample, thus $|Y| = TP + FN$; and $|X \cap Y|$ is the number of elements in $X \cap Y$, thus $|X \cap Y| = TP$. Therefore, Sørensen's index can be defined as:

$$Sorensen = \frac{2TP}{(TP + FP) + (TP + FN)} = \frac{2TP}{2TP + FP + FN}$$

which is the same as the F1-Score described above.

Cohen's Kappa

Cohen's Kappa (also Coefficient of Agreement for Nominal Scales) is a classic measurement of agreement between two classifiers. It is defined as

$$\kappa = \frac{p_o - p_c}{1 - p_c} = \frac{f_o - f_c}{N - f_c}$$

where N is the sample population size, f_o is the frequency of agreement between the classifiers, $p_o = \frac{f_o}{N}$ is the proportion of agreement, f_c is the expected agreement frequency, and $p_c = \frac{f_c}{N}$ is the expected proportion of agreement. In the context of binary presence-absence models, "agreement" means TP and TN, thus $f_o = TP + TN$. The expected frequencies of TP f_c^{TP} and TN f_c^{TN} are calculated as

$$f_c^{TP} = \frac{(TP + FP) \cdot (TP + FN)}{N}$$

$$f_c^{TN} = \frac{(TN + FN) \cdot (TN + FP)}{N}$$

therefore

$$f_c = \frac{(TP + FP) \cdot (TP + FN) + (TN + FN) \cdot (TN + FP)}{N}$$

Metrics for probability of occurrence outputs

Area Under Curve (AUC)

AUC represents the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve is constructed with paired values of True Positive Rate (TPR, Sensitivity) and False Positive Rate (FPR, $1 - Specificity$) for each unique predicted probability as threshold values, with TPR on the y axis and FPR on the x axis. As both TPR and FPR are bounded to the $[0,1]$ interval, the maximum AUC possible is 1, while $AUC = 0.5$ represents a perfectly diagonal ROC curve, and therefore a completely random prediction.

411 Grouping and ranking predicted environmental suitability or presence probabilities for
412 known presence and absence points, it's also possible to calculate the AUC through the
413 Wilcoxon–Mann–Whitney U_1 statistics, using:

414 $$U_1 = R - \frac{n_p(n_p + 1)}{2}$$

415 $$AUC = \frac{U_1}{n_p \cdot n_a}$$

416 Where R is the sum of the ranks of the predicted values for the presences, n_p is the
417 number of presences, and n_a is the number of absences.

418

419 *Log-score*

420 Let $p_i \in [0, 1]$ denote the probability of occurrence for a focal species at the data point
421 i (where i represents a point in time and space). Denote $y_i \in \{0, 1\}$ (with one denoting
422 occurrence) the "true" occurrence at i. The log score for data point i is given by:

423 $$S(p_i, y_i) = -I(y_i = 0)\ log\ (1 - p_i) - I(y_i = 1)\ log\ (p_i),$$

424 where $I(A)$ is an indicator function, returning 1 if condition A is true and 0 otherwise. Thus,
425 given a set n observations and corresponding probabilities of occurrence, the mean log-score
426 is given by

427 $$Log - score = \frac{1}{n} \sum_{i=1}^{n} S(p_i, y_i).$$

428

429 *Brier score (BS)*

430 The BS is defined similarly to the log-score:

431 $$S(p_i, y_i) = (p_i - y_i)^2,$$

432 which yields:

$$BS = \frac{1}{n}\sum_{i=1}^{n} S(p_i, y_i)$$

Brier Skill Score (BSS)

Measures relative improvement over a reference/baseline model (such as climate prevalence or random guessing):

$$BSS = 1 - \frac{BS_{model}}{BS_{reference}}$$

Where 1 is perfect skill, 0 means no better than baseline, and <0 means worse than baseline.

Metrics for abundance outputs

Root Mean Square Error (RMSE)

For the abundance outputs, the error for the i-th observation is simply $e_i = y_i - \hat{y}_i$, where y_i is the observed abundance and $\hat{y}_i$ is the predicted abundance at a i given point. Using the error, it's possible to compute a number of model performance metrics. For a model with sample size n, the Root Mean Squared Error, defined as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Note that $\frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2$ corresponds to the Mean Squared Error (MSE).

Slope and intercept

Using observed abundance y as predictor and predicted abundance $\hat{y}$ as response variable, it is possible to fit a linear model with formula.

$$y = \beta_0 + \beta_1 \hat{y}$$

where the intercept β_0 and the slope β_1 are estimated parameters. Note that when $\beta_0 = 0$ and $\beta_1 = 1$, the equation produces a diagonal line indicating perfect correspondence between predicted and observed abundance.

Percent Bias

Let y_i and $\hat{y}_i$ denote the observed and predicted abundance, respectively. The Percent Bias is given by

$$Percent\ Bias = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)}{\sum_{i=1}^{n} y_i}.$$

Coverage probability

The Coverage Probability of a confidence interval (CI) procedure is the probability that the random interval $[\hat{L}(X), \hat{U}(X)]$, constructed from sample data X, contains the true population parameter θ.

For a population parameter $\theta \in \Theta$, where $\hat{L}(X)$ and $\hat{U}(X)$ are the lower and upper bounds of the confidence interval calculated from sample X, the theoretical coverage probability is defined as:

$$P_\theta(\hat{L}(X) \leq \theta \leq \hat{U}(X)) = 1 - \alpha$$

Where θ: True, unknown population parameter; $\hat{L}(X)$, $\hat{U}(X)$: Random variables representing the lower and upper confidence limits. $1 - \alpha$: Nominal (target) confidence level (e.g., 0.95 for a 95% confidence interval).

P-dispersion

Let σ_y be the standard deviation of observed abundance values y, and $\sigma_{\hat{y}}$ the standard deviation of predicted abundance values $\hat{y}$; the $P_{dispersion}$ metric can be calculated as

$$P_{dispersion} = \frac{\sigma_{\hat{y}}}{\sigma_y}$$

Continuous Ranked Probability Score (CRPS)

Denote by F_i the cumulative density function of the predictive distribution for the targeted abundance y_i. The CRPS is given by:

$$S(F_i, y_i) = -\int_{-\infty}^{\infty} (F(x) - I(y \leq x))^2 dx.$$

If the predictive distribution is Normal with mean $\hat{y}_i$ and standard deviation σ_i, the formula above simplifies to:

$$S(F_i, y_i) = \sigma_i \left[\frac{1}{\sqrt{\pi}} - 2\varphi\left(\frac{y_i - \hat{y}_i}{\sigma_i}\right) - \frac{y_i - \hat{y}_i}{\sigma_i}\left(2\Phi\left(\frac{y_i - \hat{y}_i}{\sigma_i}\right) - 1\right)\right],$$

where φ and Φ denote the probability and cumulative density functions of a standard Normal distribution. For computing the CRPS based on samples from the predictive distribution, we refer the reader to (Jordan et al., 2019).

Finally, the CRPS is given by:

$$CRPS = \frac{1}{n}\sum_{i=1}^{n} S(F_i, y_i).$$

Skill scores

Given a score function, to compute the associated "skill," one needs both a "reference" and an "optimal" score. Let us consider the CRPS in this illustrative example. Suppose $CRPS(k)$ is the CRPS calculated based on a model which we will call model k. We want to compute the CRPS skill for that model, termed $CRPS_{Skill}(k)$. Define $CRPS_{Ref}$ as the

CRPS from a reference model, and $CRPS_{Opt}$ as an optimal value for the score function. Then, the $CRPS_{Skill}(k)$ is defined as:

$$CRPS_{Skill}(k) = \frac{CRPS(k) - CRPS_{Ref}}{CRPS_{opt} - CRPS_{Ref}}.$$

The skill scores corresponding to the other scoring rules previously defined are calculated analogously.

Moving window forecast skill

Moving windows are widely used with time series. It is possible to use the technique to evaluate the skill of a model to predict species abundance over time. One method is to compute Mean Squared Error (MSE) for a time window, use it as a baseline, and compare it to the MSE of subsequent windows.

For example, consider an observed abundance time series $\{y_1, y_2, y_3, \ldots, y_n\}$ of length n, and the predicted values for the series $\{\hat{y}_1, \hat{y}_2, \hat{y}_3, \ldots, \hat{y}_n\}$, where indexes mean the position t on time. It is possible to divide the series into several sequential and overlapping windows of size w starting on each t, and compute and compare their MSE. For example, considering $w = 3$, the first window of the series would be defined as $\{y_1, y_2, y_3\}$ and $\{\hat{y}_1, \hat{y}_2, \hat{y}_3\}$, and the second as $\{y_2, y_3, y_4\}$ and $\{\hat{y}_2, \hat{y}_3, \hat{y}_4\}$. Taking the first window as baseline, we can compute the MSE_1 with

$$MSE_1 = \frac{(\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 + (\hat{y}_3 - y_3)^2}{w}$$

and the error for the next window MSE_2 with

$$MSE_2 = \frac{(\hat{y}_2 - y_2)^2 + (\hat{y}_3 - y_3)^2 + (\hat{y}_4 - y_4)^2}{w}$$

The ratio $\frac{MSE_2}{MSE_1} - 1$ therefore represents the relative change in the estimator error through time. For example, if $\frac{MSE_2}{MSE_1} - 1 = 0.5$, there was an increase of 50% in error; conversely, if $\frac{MSE_2}{MSE_1} - 1 = -0.5$, there was a decrease of 50% in error.

Metrics for abundance and probability of occurrence outputs

Mean Absolute Error (MAE)

The MAE is a simple metric defined as

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i|$$

Where $\hat{y}_i$ and y_i are the predicted and observed abundance in the i-th observation, and n is the sample size.

R^2 and pseudo-R^2

The coefficient of determination (R^2) measures the proportion of the variance in the dependent variable that is explained by the predictors (independent variables) of a model. Consider y as the observed abundance values, $\hat{y}$ as the predicted abundance values, and n as the sample size. The mean observed abundance is defined as

$$\underline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

The sum of the squares of the residuals SS_{res} and the total sum of the squares SS_{tot} are defined as

$$SS_{res} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \underline{y})^2$$

and the determination coefficient R^2 is defined as

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n} (\widehat{y_i} - y_i)^2}{\sum_{i=1}^{n} (y_i - \underline{y})^2}$$

Correlation

Correlation is the measure of association between two variables, representing how they change together. In the context of distribution models, one of the most used correlation metrics is the parametric Pearson's coefficient r, which measures linear correlation in data and can be defined as

$$r = \frac{\sum_{i=1}^{n} \left(y_i - \underline{y}\right)\left(f_i - \underline{f}\right)}{\sqrt{\sum_{i=1}^{n} \left(y_i - \underline{y}\right)^2 \sum_{i=1}^{n} \left(f_i - \underline{f}\right)^2}}$$

Where y_i and f_i are the observed and predicted abundance, respectively, at the i-th point, and $\underline{y}$ and $\underline{f}$ are the mean observed and predicted abundance, respectively.

When it is not possible to assume that the data are normally distributed, often the non-parametric Spearman's rank correlation is used to compute the non-linear correlation between data. To do that, y and f are ranked, i.e., for each i-th pair (y_i, f_i), there is a pair of their ranks $(R[y_i], R[f_i])$, and the difference d between the ranks is calculated as $d_i = R[y_i] - R[f_i]$. When there are no ties in y nor in f ranks, Spearman's rank correlation ρ for a sample of n observations is defined as

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d^2}{n(n^2 - 1)}$$

When there are ties, which are often the case in ecological data, the common procedure is to resolve ties by assigning them average ranks and calculate the Pearson correlation coefficient r between them. Note that $r = 1$ or $\rho = 1$ both represent a perfect

positive correlation, $r = -1$ or $\rho = -1$ represents a perfect negative correlation, and $r = 0$ or $\rho = 0$ represent no correlation.

Harrell's Concordance Index (C-index)

For continuous data (e.g., continuous predictions vs. continuous observed values), C-Index measures the proportion of all evaluable pairs of observations in which the relative order of predicted values agrees with the relative order of actual observed values.

For observations i and j, define:

$$C = \frac{\sum_{i,j} I(pair\ is\ comparable)\ I(pair\ is\ concordant)}{\sum_{i,j} I(pair\ si\ comparable)}$$

For uncensored continuous data, where every pair is comparable, this simplifies to:

$$C = \frac{\sum_{i<j} I[(x_i - x_j)(y_i - y_j) > 0] + 1/2 \sum_{i<j} I[(x_i - x_j)(y_i - y_j) = 0]}{n(n-1)/2}$$

Where x_i is the predicted value (e.g., species abundance) for sample i; y_i is the observed continuous outcome for sample i; $I(\cdot)$ is the indicator function, equal to 1 when the condition is true and 0 otherwise; n = number of observations; and $n(n-1)/2$ = total number of unique pairs.

References

Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., and Anderson, R. P.: spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models, Ecography, 38, 541–545, https://doi.org/10.1111/ecog.01132, 2015.

Barbet-Massin, M., Thuiller, W., and Jiguet, F.: How much do we overestimate future local extinction rates when restricting the range of occurrence data in climate suitability models?, Ecography, 33, 878–886, https://doi.org/10.1111/j.1600-0587.2010.06181.x, 2010.

Beck, J., Böller, M., Erhardt, A., and Schwanghart, W.: Spatial bias in the GBIF database and its effect on modeling species' geographic distributions, Ecological Informatics, 19, 10–15, https://doi.org/10.1016/j.ecoinf.2013.11.002, 2014.

Boakes, E. H., McGowan, P. J. K., Fuller, R. A., Chang-qing, D., Clark, N. E., O'Connor, K., and Mace, G. M.: Distorted views of biodiversity: Spatial and temporal bias in species occurrence data, PLoS Biol, 8, e1000385, https://doi.org/10.1371/journal.pbio.1000385, 2010.

Briscoe, N. J., Morris, S. D., Mathewson, P. D., Buckley, L. B., Jusup, M., Levy, O., Maclean, I. M. D., Pincebourde, S., Riddell, E. A., Roberts, J. A., Schouten, R., Sears, M. W., and Kearney, M. R.: Mechanistic forecasts of species responses to climate change: The promise of biophysical ecology, Global Change Biology, 29, 1451–1470, https://doi.org/10.1111/gcb.16557, 2023.

Brown, J. H., Gillooly, J. F., Allen, A. P., Savage, V. M., and West, G. B.: Toward a metabolic theory of ecology, Ecology, 85, 1771–1789, https://doi.org/10.1890/03-9000, 2004.

Buckley, L. B., Carrington, E., Dillon, M. E., García-Robledo, C., Roberts, S. B., Wegrzyn, J. L., and Urban, M. C.: Characterizing biological responses to climate variability and extremes to improve biodiversity projections, PLOS Climate, 2, e0000226, 2023a.

Buckley, L. B., Briones Ortiz, B. A., Caruso, I., John, A., Levy, O., Meyer, A. V., Riddell, E. A., Sakairi, Y., and Simonis, J. L.: TrenchR: An R package for modular and accessible microclimate and biophysical ecology, PLOS Climate, 2, e0000139, 2023b.

Colwell, R. K.: Spatial scale and the synchrony of ecological disruption, Nature, 599, E8–E10, https://doi.org/10.1038/s41586-021-03759-x, 2021.

Da Cunha Godoy, L., Fredston, A., Bandara, R. M. W. J., Morales, M., and Pinsky, M.: drmr: A Bayesian approach to Dynamic Range Models in R, https://doi.org/10.32942/X2VT0N, 14 April 2026.

De Oliveira Junior, A. C. and Velazco, S. J. E.: Are deep neural networks worth the effort for abundance-based species distribution models?, Manuscript under review, n.d.

Diana, A., Dennis, E. B., Matechou, E., and Morgan, B. J. T.: Fast Bayesian Inference for Large Occupancy Datasets, Biometrics, 79, 2503–2515, https://doi.org/10.1111/biom.13816, 2023.

Dobson, R., Challinor, A. J., Cheke, R. A., Jennings, S., Willis, S. G., and Dallimer, M.: DYNAMICSDM : An R package for species geographical distribution and abundance modelling at high spatiotemporal resolution, Methods Ecol Evol, 14, 1190–1199, https://doi.org/10.1111/2041-210X.14101, 2023.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., and Lautenbach, S.: Collinearity: a review of methods to deal with it and a simulation study evaluating their performance, Ecography, 36, 27–46, https://doi.org/10.1111/j.1600-0587.2012.07348.x, 2013.

Engler, R. and Guisan, A.: MigClim: Predicting plant distribution and dispersal in a changing climate, Diversity and Distributions, 15, 590–601, https://doi.org/10.1111/j.1472-4642.2009.00566.x, 2009.

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

Franklin, J.: Mapping species distributions: spatial Inference and prediction, Cambridge University Press, United States of America, 320 pp., 2010.

Fredston, A., Ovando, D., Da Cunha Godoy, L., Kong, J., Muffley, B., Thorson, J., and Pinsky, M.: Dynamic range models improve the near-term forecast for a marine species on the move, https://doi.org/10.32942/X24D00, 28 March 2025.

637 Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B.: Bayesian Data Analysis, 0 ed.,
638 Chapman and Hall/CRC, https://doi.org/10.1201/9780429258480, 2003.

639 Gonzalez, A., Chase, J. M., and O'Connor, M. I.: A framework for the detection and
640 attribution of biodiversity change, Phil. Trans. R. Soc. B, 378, 20220182,
641 https://doi.org/10.1098/rstb.2022.0182, 2023.

642 Harris, D. J., Taylor, S. D., and White, E. P.: Forecasting biodiversity in breeding birds using
643 best practices, PeerJ, 6, e4278, https://doi.org/10.7717/peerj.4278, 2018.

644 Higgins, S. I., Burkhardt, A., Conradi, T., and Muhoko, E.: TTR.PGM: An R package for
645 modelling the distributions and dynamics of plants using the Thornley transport resistance
646 plant growth model, Methods in Ecology and Evolution, 16, 1599–1608,
647 https://doi.org/10.1111/2041-210X.70071, 2025.

648 Jordan, A., Krüger, F., and Lerch, S.: Evaluating Probabilistic Forecasts with **scoringRules**,
649 J. Stat. Soft., 90, https://doi.org/10.18637/jss.v090.i12, 2019.

650 Kearney, M. R. and Porter, W. P.: Mechanistic niche modelling: combining physiological and
651 spatial data to predict species' ranges, Ecology Letters, 12, 334–350,
652 https://doi.org/10.1111/j.1461-0248.2008.01277.x, 2009.

653 Kearney, M. R. and Porter, W. P.: NicheMapR–an R package for biophysical modelling: the
654 microclimate model, Ecography, 40, 664–674, 2017.

655 Kearney, M. R. and Porter, W. P.: NicheMapR–an R package for biophysical modelling: the
656 ectotherm and Dynamic Energy Budget models, Ecography, 43, 85–96, 2020.

657 Kearney, M. R., Briscoe, N. J., Mathewson, P. D., and Porter, W. P.: NicheMapR–an R
658 package for biophysical modelling: the endotherm model, Ecography, 44, 1595–1605, 2021a.

659 Kearney, M. R., Jusup, M., McGeoch, M. A., Kooijman, S. A., and Chown, S. L.: Where do
660 functional traits come from? The role of theory and models, Functional Ecology, 35, 1385–
661 1396, 2021b.

662 Leão, C. L. F., Rose, M. B., Lima, M. G. M., and Velazco, S. J. E.: Choices or time? How
663 algorithms, ensembles, and temporal extent of data shape species distribution models,
664 Manuscript under review, n.d.

MacKenzie, D. I., Nichols, J. D., Hines, J. E., Knutson, M. G., and Franklin, A. B.: Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly, Ecology, 84, 2200–2207, https://doi.org/10.1890/02-3090, 2003.

Maclean, I. M. and Klinges, D. H.: Microclimc: A mechanistic model of above, below and within-canopy microclimate, Ecological Modelling, 451, 109567, 2021.

Maclean, I. M., Mosedale, J. R., and Bennie, J. J.: Microclima: An r package for modelling meso-and microclimate, Methods in Ecology and Evolution, 10, 280–290, 2019.

Malchow, A.-K., Bocedi, G., Palmer, S. C. F., Travis, J. M. J., and Zurell, D.: RangeShiftR: an R package for individual-based simulation of spatial eco-evolutionary dynamics and species' responses to environmental changes, Ecography, 44, 1443–1452, https://doi.org/10.1111/ecog.05689, 2021.

Malchow, A.-K., Fandos, G., Kormann, U. G., Grüebler, M. U., Kéry, M., Hartig, F., and Zurell, D.: Fitting individual-based models of spatial population dynamics to long-term monitoring data, Ecological Applications, 34, e2966, https://doi.org/10.1002/eap.2966, 2024.

Mendes, P., Velazco, S. J. E., Andrade, A. F. A. de, and De Marco, P.: Dealing with overprediction in species distribution models: How adding distance constraints can improve model accuracy, Ecological Modelling, 431, 109180, https://doi.org/10.1016/j.ecolmodel.2020.109180, 2020.

Meyer, A. V., Sakairi, Y., Kearney, M. R., and Buckley, L. B.: A guide and tools for selecting and accessing microclimate data for mechanistic niche modeling, Ecosphere, 14, e4506, https://doi.org/10.1002/ecs2.4506, 2023.

de Oliveira Junior, A. C. and Velazco, S. J. E.: adm: An R package for constructing abundance-based species distribution models, Methods in Ecology and Evolution, 16, 1404–1412, https://doi.org/10.1111/2041-210X.70074, 2025.

Petchey, O. L., Pontarp, M., Massie, T. M., Kéfi, S., Ozgul, A., Weilenmann, M., Palamara, G. M., Altermatt, F., Matthews, B., Levine, J. M., Childs, D. Z., McGill, B. J., Schaepman, M. E., Schmid, B., Spaak, P., Beckerman, A. P., Pennekamp, F., and Pearse, I. S.: The ecological forecast horizon, and examples of its uses and determinants, Ecology Letters, 18, 597–611, https://doi.org/10.1111/ele.12443, 2015.

694 Piironen, J. and Vehtari, A.: Sparsity information and regularization in the horseshoe and
695 other shrinkage priors, Electron. J. Statist., 11, https://doi.org/10.1214/17-EJS1337SI, 2017.

696 Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein,
697 S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F.,
698 and Dormann, C. F.: Cross-validation strategies for data with temporal, spatial, hierarchical,
699 or phylogenetic structure, Ecography, 40, 913–929, https://doi.org/10.1111/ecog.02881,
700 2017.

701 Rojas-Soto, O., Forero-Rodríguez, J. S., Galindo-Cruz, A., Mota-Vargas, C., Parra-Henao, K.
702 D., Peña-Peniche, A., Piña-Torres, J., Rojas-Herrera, K., Sánchez-Rodríguez, J. D., Toro-
703 Cardona, F. A., and Trinidad-Domínguez, C. D.: Calibration areas in ecological niche and
704 species distribution modelling: Unravelling approaches and concepts, Journal of
705 Biogeography, n/a, https://doi.org/10.1111/jbi.14834, 2024.

706 Rose, M. B., Velazco, S. J. E., Regan, H. M., and Franklin, J.: Rarity, geography, and plant
707 exposure to global change in the California Floristic Province, Global Ecology and
708 Biogeography, 32, 218–232, https://doi.org/10.1111/geb.13618, 2023.

709 Royle, J. A. and Kéry, M.: A Bayesian state-space formulation of dynamic occupancy
710 models, Ecology, 88, 1813–1823, https://doi.org/10.1890/06-0669.1, 2007.

711 Rushing, C. S., Royle, J. A., Ziolkowski, D. J., and Pardieck, K. L.: Modeling spatially and
712 temporally complex range dynamics when detection is imperfect, Sci Rep, 9, 12805,
713 https://doi.org/10.1038/s41598-019-48851-5, 2019.

714 Salguero-Gómez, R., Jones, O. R., Archer, C. R., Buckley, Y. M., Che-Castaldo, J., Caswell,
715 H., Hodgson, D., Scheuerlein, A., Conde, D. A., Brinks, E., de Buhr, H., Farack, C.,
716 Gottschalk, F., Hartmann, A., Henning, A., Hoppe, G., Römer, G., Runge, J., Ruoff, T.,
717 Wille, J., Zeh, S., Davison, R., Vieregg, D., Baudisch, A., Altwegg, R., Colchero, F., Dong,
718 M., de Kroon, H., Lebreton, J.-D., Metcalf, C. J. E., Neel, M. M., Parker, I. M., Takada, T.,
719 Valverde, T., Vélez-Espino, L. A., Wardle, G. M., Franco, M., and Vaupel, J. W.: The
720 compadre Plant Matrix Database: an open online repository for plant demography, Journal of
721 Ecology, 103, 202–218, https://doi.org/10.1111/1365-2745.12334, 2015.

Salguero-Gómez, R., Jones, O. R., Archer, C. R., Bein, C., de Buhr, H., Farack, C., Gottschalk, F., Hartmann, A., Henning, A., Hoppe, G., Römer, G., Ruoff, T., Sommer, V., Wille, J., Voigt, J., Zeh, S., Vieregg, D., Buckley, Y. M., Che-Castaldo, J., Hodgson, D., Scheuerlein, A., Caswell, H., and Vaupel, J. W.: COMADRE: a global data base of animal demography, Journal of Animal Ecology, 85, 371–384, https://doi.org/10.1111/1365-2656.12482, 2016.

Santini, L., Benítez-López, A., Maiorano, L., Čengić, M., and Huijbregts, M. A. J.: Assessing the reliability of species distribution projections in climate change research, Divers Distrib, ddi.13252, https://doi.org/10.1111/ddi.13252, 2021.

Serra-Diaz, J. M., Andrews, L. C., Schwarz Meyer, A., Carlson, B. S., Maitner, B., Pinilla-Buitrago, G. E., Pigot, A. L., Trisos, C. H., Wilson, A. M., Urban, M. C., and Merow, C.: A global early warning system for predicting exposure of biodiversity to extreme heat, Nat. Clim. Chang., 1–8, https://doi.org/10.1038/s41558-026-02642-9, 2026.

Thuiller, W., Guéguen, M., Renaud, J., Karger, D. N., and Zimmermann, N. E.: Uncertainty in ensembles of global biodiversity scenarios, Nat Commun, 10, 1446, https://doi.org/10.1038/s41467-019-09519-w, 2019.

Varela, S., Anderson, R. P., García-Valdés, R., and Fernández-González, F.: Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models, Ecography, 37, 1084–1091, https://doi.org/10.1111/j.1600-0587.2013.00441.x, 2014.

Velazco, S. J. E., Ribeiro, B. R., Laureto, L. M. O., and De Marco Júnior, P.: Overprediction of species distribution models in conservation planning: A still neglected issue with strong effects, Biological Conservation, 252, 108822, https://doi.org/https://doi.org/10.1016/j.biocon.2020.108822, 2020.

Waldock, C., Stuart-Smith, R. D., Albouy, C., Cheung, W. W. L., Edgar, G. J., Mouillot, D., Tjiputra, J., and Pellissier, L.: A quantitative review of abundance-based species distribution models, Ecography, 2022, https://doi.org/10.1111/ecog.05694, 2022.

Ye, T., Lachos, V. H., Wang, X., and Dey, D. K.: Comparisons of zero-augmented continuous regression models from a Bayesian perspective, Statistics in Medicine, 40, 1073–1100, https://doi.org/10.1002/sim.8795, 2021.

751 Zipkin, E. F., Grant, E. H. C., and Fagan, W. F.: Evaluating the predictive abilities of
752 community occupancy models using AUC while accounting for imperfect detection,
753 Ecological Applications, 22, 1962–1972, https://doi.org/10.1890/11-1936.1, 2012.