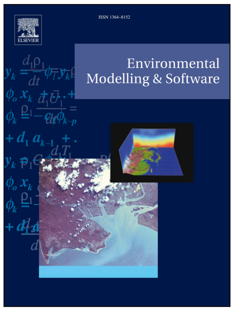


ENMTML: An R package for a straightforward construction of complex ecological niche models

André Felipe Alves de Andrade, Santiago José Elías Velazco, Paulo De Marco Júnior

PII: S1364-8152(19)31042-4

DOI: https://doi.org/10.1016/j.envsoft.2019.104615

Reference: ENSO 104615

To appear in: *Environmental Modelling and Software*

Received Date: 29 October 2019

Revised Date: 19 December 2019

Accepted Date: 26 December 2019

Title:

ENMTML: An R package for a straightforward construction of complex Ecological Niche Models

Authors:

André Felipe Alves de Andrade[1]

Santiago José Elías Velazco[2]

Paulo De Marco Júnior[1]

Address:

[1] Theory, Metacommunity and Landscape Ecology Lab, ICB V, Universidade Federal de Goiás, CP 131, 74.001-970, Goiânia, GO, Brazil.

[2] Instituto de Biología Subtropical, Universidad Nacional de Misiones-CONICET, Bertoni 85, CP 3370, Puerto Iguazú, Misiones, Argentina

ORCiD IDs:

André Felipe Alves de Andrade: 0000-0002-6134-3176

Santiago José Elías Velazco: 0000-0002-7527-0967

Paulo De Marco Júnior: 0000-0002-3628-6405

Corresponding author:

André Felipe Alves de Andrade

Theory, Metacommunity and Landscape Ecology Lab, ICB V, Universidade Federal de Goiás, CP 131, 74.001-970, Goiânia, GO, Brazil.

Email:andrefaandrade@gmail.com

Telephone: +553135211373

Abstract

Ecological niche models (ENMs) is a popular method in ecology, mostly due to its broad applicability and the fact that required data is simple and easily accessible from digital databases. Nevertheless, there is an underlying methodological complexity, often overlooked by many scientists that rely on ENMs to achieve other objectives. We present here the package ENMTML, an Open Source R package. The main purpose of this package is to assemble all this methodological complexity spread over several papers and bring it into the spotlight in a simple way for people not used to the details of ENMs. The package contains several alternatives to different methodological steps, e.g., pseudo-absence allocation and accessible area delimitation, formulated within a single function, to make it accessible for people not used to the programming environment.

Keywords: Species distribution model; open-source software; Niche modelling; Model evaluation;

Software and data availability

Availability of Software	License GPL
Name of software: ENMTML	Code repository
Type of software: Add-on package for R https://cran.r-project.org	https://github.com/andrefaa/ENMTML
	Installation in R *install_github*("andrefaa/ENMTML")
First available 2019	
Program language: R	
Requires R version 3.6.0 or later	

1. Introduction

Ecological Niche and Species Distribution Models (ENMs and SDMs, respectively), are widely applied in ecology, providing important basal information for the most diverse fields, such as conservation (e.g. Keppel *et al.*, 2012; Razgour *et al.*, 2018), biological invasions (Peterson, 2003; Campos *et al.*, 2014; Lins *et al.*, 2018), phylogenetic/evolutionary studies (e.g. Carstens & Richards, 2007; Chifflet *et al.*, 2016) and disease management (Peterson & Shaw, 2003). While there are theoretical differences among ENMs and SDMs (see Peterson & Soberón, 2012), we will adopt the nomenclature ENM from now on as most studies are closer to estimating species' niche. Such broad applicability is related to two significant properties of ENMs: (i) a simple underlying model that requires only occurrence data and environmental variables, and (ii) a huge effort employed by researchers to develop robust methods and software. There is a significant change in methods from first ENMs studies, that uses one to few algorithms and do not explore other steps that could influence the result (Peterson & Holt, 2003), to current studies, which use several algorithms and diverse steps to fit models, such as pseudo-absence allocation and accessible area definition (e.g., Velazco *et al.*, 2019).

One of the major assets of ENMs is its community, with several researchers dedicated to delving into specific methodological aspects of the modeling process. Some noteworthy aspects involve the control of collinearity among environmental variables (De Marco & Nóbrega, 2018), different strategies for the allocation of pseudo-absences (Engler *et al.*, 2004; Barbet-Massin *et al.*, 2012; Senay *et al.*, 2013); careful definition of the accessible area (Peterson *et al.*, 2001; Soberón, 2010; Barve *et al.*, 2011; Cooper & Soberón, 2018); ensemble of different algorithms (Marmion *et al.*, 2009; Thuiller *et al.*, 2009; Hao *et al.*, 2019); different evaluation metrics (Allouche *et al.*, 2006; Leroy *et al.*, 2018) and diverse methods to partition the occurrence data for fitting and evaluating the model (Muscarella *et al.*, 2014; Roberts *et al.*, 2017). Given the wide variety of methods for each one of the several steps of fitting ENMs and the possible interactions that may arise, the number of models produced for a single species may easily surpass a thousand.

The great diversity of choices creates a duality in ENMs: while models are simple to fit and the required data is easily available, several decisions should be made regarding methodological steps that must be done judiciously and are not as readily available as the data. As a result, studies that rely on ENMs usually do not have the same methodological rigor as studies that focus on developing ENMs, i.e., several studies still apply (Area Under the Curve) AUC as an evaluation metric, even though it has been demonstrated for over 10 years that the metric is deeply affected by prevalence (Lobo *et al.*, 2008) or the extent of the accessible area (Peterson *et al.*, 2008; Barve *et al.*, 2011). On the other side, there has been a great effort to develop alternatives for the AUC and several other methodological aspects, which have been implemented in several R packages and ENMs software (Thuiller *et al.*, 2009; Guo & Liu, 2010; Naimi & Araújo, 2016; Hijmans *et al.*, 2017; Golding *et al.*, 2018; Kass *et al.*, 2018; Sánchez-Tapia *et al.*, 2018; Cobos *et al.*, 2019).

Ideally, ENMs should be fit-for-purpose, which means that fitting ENMs is a process that must be thought carefully, as there is not a single correct way to fit models (Guillera-Arroita *et al.*, 2015; Qiao *et al.*, 2015). Due to the great variety of methodological choices and the velocity that new alternatives arise, it may be hard to keep up with novelties within the ENMs' field. As a result, people who are not involved in the methodological developments within the field or do not have connections to developers have small participation in all the published papers (Ahmed *et al.*, 2015). We introduce here *ENMTML,* a new R package to fit ENMs. The main objective of this package is to put together all this methodological diversity developed within the ENM field and present it to users simply and transparently. Despite being an R package, we also made it friendly for non-programmers and summarized the whole fitting process into a single function with several arguments that correspond to the methodological alternatives.

2. Methods description

2.1 Arguments and settings

The *ENMTML* package and its processes can be divided into three major stages: pre-processing, processing, and post-processing. This division in three stages is familiar to most ENMs routines. Identifying the stage in which each methodological step will be performed may help users to understand the connections among the different methodological steps and provides an overview that assists the decision-making process (Figure 1).

In the pre-processing stage, the data is input (species occurrences and predictors variables), and a series of steps can be performed before fitting the model. Occurrence data is input as a tab-separated text file (TXT). The program automatically uses unique occurrences per cell. In addition, the user can control two steps acting over the occurrence dataset: i) the minimum number of occurrences valid for model fitting and ii) perform a thinning process to reduce sampling bias. Regarding predictors, there are three methods to control for collinearity and the possibility to include predictors for other time or geographic windows. As for pre-processing steps, there are five different strategies for pseudo-absence allocation with the option to control for presence-absence ratio; four methods to partition the data into subsets, with the possibility to provide a specific dataset for independent evaluation (a useful asset when studying biotic invasions); two methods to create species-specific accessible areas; and it is also possible to identify extrapolation areas based on a Mobility-Oriented Parity analysis (Owens *et al.*, 2013).

The processing stage is when algorithms will fit models, and the suitability maps generated. For starters, the user can choose if both partial and final suitability maps will be generated or not. There are thirteen algorithms available for model fitting: Bioclim (Nix, 1986), Mahalanobis Distance (Farber & Kadmon, 2003), Domain (Carpenter *et al.*, 1993), Ecological Niche Factor Analysis (Hirzel *et al.*, 2002), Generalized Linear Models (McCullagh & Nelder, 1989), Generalized Additive Models (Hastie & Tibshirani, 1990), Boosted Regression Tree (Friedman, 2001), Random Forests (Prasad *et al.*, 2006), Support Vector Machine (Guo *et al.*, 2005), Maximum Entropy with quadratic and linear (Anderson & Gonzalez, 2011) and default features (Phillips *et al.*, 2006; Phillips, 2017), Maximum Likelihood (Royle *et al.*, 2012) and Gaussian Process (Golding & Purse, 2016).

Finally, in the post-processing stage, the suitability maps generated from the different algorithms are evaluated using seven different metrics (AUC, True Skill Statistics (TSS), Kappa, Jaccard, Sorensen, Boyce, and F_{pb}). When multiple models are fitted for the same species (i.e., several replicates or geographical partitions), the evaluation output result is the mean and standard deviation of the partial models. Other post-processing options include the creation of binary maps based on five different thresholds; six different ways to generate ensemble models; and the application of spatial restrictions to reduce model commission and bring the result closer to an estimation of the species realized distribution (MSDM).

All features are organized in a single R function with multiple arguments the user needs to fill according to the specific purpose. We chose not to establish default arguments, so users must think carefully about the choices. To provide support, we briefly explain the methodological steps and indicate relevant studies for each alternative.

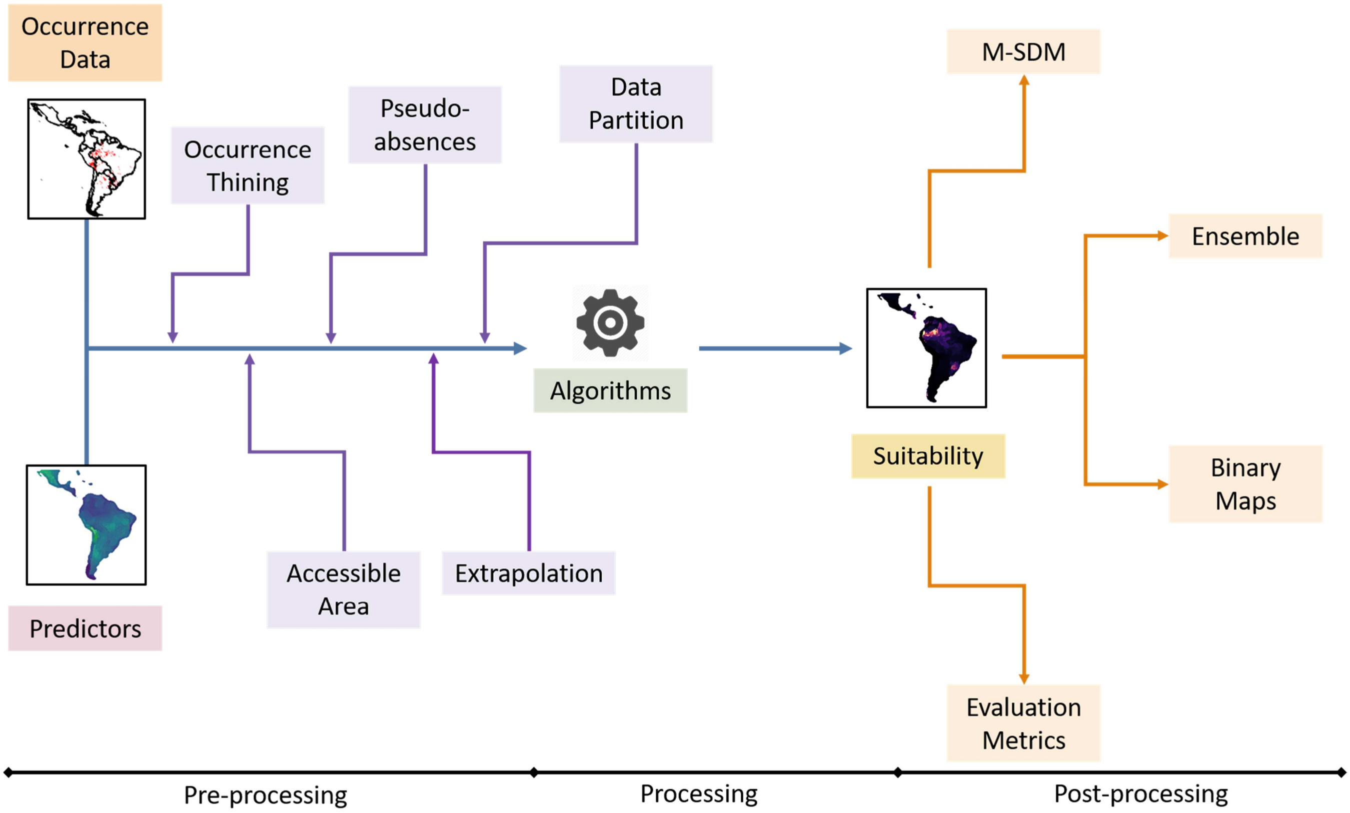


Figure 1: General workflow of the *ENMTML* package and all steps that can be taken at each stage of the modeling routine. In the pre-processing stage occurrence and predictors are imported and the user may take six different steps before fitting the models. The processing stage involves fitting the models and producing the suitability maps, which can be made using thirteen different algorithms. In the post-processing stage, the results are evaluated and the user may perform analysis upon the different suitability maps produced.

2.2. Occurrence data processing

`Arguments involved: (occ_file/ Sp / x / y / min_occ / thin_occ)`

Occurrence data is imported as a tab-separated TXT file that needs to be specified by the user as the file path of the file in the argument `occ_file`. This file must contain information about species name, longitude, and latitude (in decimal degrees), and the name of those columns must be provided in the arguments `Sp, x, y`.

The user must also provide the minimum number of unique occurrences valid for model fitting in the argument `min_occ`, species below this number will be excluded from the analysis. There is not a rule for the definition of a minimum number of occurrences, but there are several studies that indicate that model accuracy is directly related to sample size (Wisz *et al.*, 2008). There are several factors that affect the viable minimum number of occurrences(Mateo *et al.*, 2010), but a good framework for exploring this subject is the one developed by van Proosdij *et al.* (2015).

Finally, users might opt to reduce autocorrelation in occurrence data and possible sampling bias by a thinning technique (argument `thin_occ`), performed using the package spThin (Aiello-Lammens *et al.*, 2015). There are three alternatives for defining the thinning distance: i) based on the distance of a Moran's I Variogram that minimizes the spatial autocorrelation; ii) retaining unique cells that fall within a grid two times greater than the original cellsize; and iii) based on a minimum distance defined by the user (Table 1). For a better comprehension of the topic see (Aiello-Lammens *et al.*, 2015).

Tables 1: Thinning alternatives in the *ENMTML* package (references for each method indicate its original development or an example of its best practice use).

Occurrence thinning method	Acronym used in the `thin_occ` argument	Method description	Additional arguments	References
None	NULL	Use original unique occurrences	-	-
Moran Variogram	MORAN	Choose from pairs of occurrences which are within a distance defined by a Moran Variogram	-	Veloz (2009)
2x cell-size	CELLSIZE	Choose from pairs of occurrences which are within a distance defined by 2x cellsize	-	Velazco *et al.*(2019)
User-defined	USER-DEFINED	Choose from pairs of occurrences which are within a distance defined by the user (in km)	Distance	Aiello-Lammens *et al.*(2015)

2.3 Predictors input and collinearity reduction

`Arguments involved: (pred_dir / proj_dir / colin_var)`

Predictors are imported in the argument `pred_dir`, which specifies the folder path of the predictors, and should be in any of the given formats: BIL, TIF, ASC, TXT. Predictors for projection also accept the same formats and should be included in nested folders, with a major folder including all the projections datasets each with its respective sub-folder (Figure 2).

Collinearity in predictors can be controlled using three different strategies: i) Pearson correlation with a threshold defined by the user; ii) Variance Inflation Factor (VIF; Marquaridt, 1970) and; Principal Component Analysis (PCA), using the axis that account for 95% of the total variance in the predictors as the new predictors (Heikkinen *et al.*, 2006; De Marco & Nóbrega, 2018). Predictors eliminated by the Pearson and VIF will also be eliminated for projections datasets. When users choose to perform a PCA and have datasets for projection, the linear relationship between the predictors and the principal components is projected onto the new datasets to create the principal components for the projection datasets (see De Marco & Nóbrega, 2018).

Table 2: Methods to reduce predictor collinearity available in the *ENMTML* package.

Variable collinearity reduction method	Acronym used in the `colin_var` argument	Method description	Additional arguments	References
None	NULL	Use original variables provided by the user	-	-
Pearson Correlation	PEARSON	Eliminates correlated variables according to a chosen threshold	Threshold	Dormann *et al.* (2013)
Variance Inflation Factor	VIF	Eliminates correlated variables based on VIF	-	Marquaridt (1970)
Principal Components Analysis	PCA	Performs a PCA on variables and use the principal components as variables	-	De Marco & Nóbrega (2018)

2.4 Pseudo-absences and background points allocation

`Arguments involved: (pseudoabs_method / pres_abs_ratio)`

The program allocates pseudo-absences and background points within the area used to calibrate the models (Table 3). Such allocation will be particular for those geographical partitioning method (such us block- and band-cross validation) in which pseudo-absences and background points are created after performing such partition, in order to maintain a homogeneous distribution of background points between partitions, as well as a constant prevalence (conceived here as the relationship between presences and pseudo-absences). Since algorithm's performance may be sensible to the way pseudo-absences are distributed throughout the calibration area (Wisz & Guisan, 2009; Barbet-Massin *et al.*, 2012), the program offers five pseudo-absences allocation methods: i) 'single random' distribution (Zaniewski *et al.*, 2002); ii) 'geographically constrained method', i.e., pseudo-absences are allocated outside a buffer around presences (Barbet-Massin *et al.*, 2012); iii) 'environmental constrained methods' based on the lowest suitable region predicted by a Bioclim model (Engler *et al.*, 2004); iv) 'geographical and environmental constrained method'(Lobo *et al.*, 2010) and; v) a three-step method which combine environmental and geographical approach plus a k-mean non-agglomerative cluster process to distribute homogeneously on environmental space (Senay *et al.*, 2013).

The program also allows for the user to define the ratio between presences and absences (argument `pres_abs_ratio`), a methodological step that received considerable focus from researchers and affects algorithm performance (Barbet-Massin *et al.*, 2012).

Table 3: Pseudo-absence allocation methods available in the *ENMTML* package.

Pseudo-absence allocation method	**Acronym used in the `pseudoabs_method` argument**	**Description of restriction**	**Reference**
Random	RND	None	Zaniewski *et al.* (2002)
Geographical Constrain	GEO_CONST	Outside a distance buffer	Barbet-Massin *et al.* (2012)
Environmental Constrain	ENV_CONST	Within lowest suitability areas predicted by a Bioclim	Engler *et al.* (2004)
Environmental and Geographical Constrain	GEO_ENV_CONST	Combination of Geographical and Environmental	Lobo *et al.* (2010)
Three-Step Constrain	GEO_ENV_KM_CONST	Combination of Environmental, Geographical and k-mean cluster	Senay *et al.* (2013)

2.5 Methods to define the accessible area

`Arguments involved: (sp_accessible_area)`

A crucial decision at the moment to construct ENMs is the hypothesized accessible area, i.e., the geographical region used by a species throughout a relevant period of time (Barve *et al.*, 2011), also known as the movement component of the BAM diagram (Soberon & Peterson, 2005). Such an accessible area can be delimited based on the knowledge of species ecology, dispersal ability, geographical barriers, and ancient region were species inhabited (Soberón, 2010; Peterson *et al.*, 2011). Nonetheless, this information is often missing for most species; therefore, different techniques act as an approximation of the accessible area. *ENMTML* account with four option to define accessible areas: i) no restriction, i.e., the entire predictors extent will be used as accessible area; ii) define an accessible area based on a buffer around occurrence data; iii) define the accessible area based on a mask, e.g., using a shapefile for biogeographical ecoregions, or; iv) accessible are defined by the user (supported formats: SHP/TIF/BIL/ASC/TXT; Table 4).

Table 4: Methods to delimit species accessible area available at the *ENMTML* package.

Accessible area definition method	**Acronym used in the `sp_accessible_area` argument**	**Type of data required**	**Reference**
Whole predictors extent	NULL	no data	-
Buffer	BUFFER	Type 1 = buffer radius based on occurrence data Type 2 = buffer radius defined by the user	Barve *et al.* (2011)
Mask	MASK	Single shapefile or raster (BIL/ASC/TIF/SHP/TXT) mask from which boundaries will be extracted	Peterson *et al.* (2001)
User-delimited	USER-DELIMITED	Folder with multiple shapefile or raster (BIL/ASC/TIF/SHP/TXT) masks, one for each species	-

2.6 Data partition

```
Arguments involved: (eval_occ / part)
```

An ideal model evaluation requires a dataset in which occurrences are independent of the ones used to fit the model; this independent dataset can be supplied as the path to a TXT file in the argument `eval_occ`.

Nevertheless, the most common evaluation method is to partition occurrence data in two subsets, one to fit the model and another for evaluation. For this option (argument `part`), the package offers four methods for data partitioning, two based on random partitions and two on geographical partitions (Table 5). Among random partition methods the user can choose: i) bootstrap, in which users specify the number of replicates and proportion of the dataset used for fitting the model, e.g., 10 replicates each with 70% for training models, the remaining 30% is used for validation; and ii) k-fold, in which the dataset is split into a chosen number of folds, and on each run the model is fit using k-1 folds and evaluated on the folder left out. As alternatives for geographical partitions, the dataset can be split based on bands (latitudinal/longitudinal) or based on a checkerboard (blocks), with occurrence data being split into two subsets, alternatively used for fitting and evaluating the model. The optimal band or checkerboard is found based on the size which presents (i) the lower spatial autocorrelation, based on Moran's I, (ii) the maximum environmental similarity, based on Multivariate Environmental Similarity Surface metric (MESS) and (iii) the minimum difference in the number of records between subsets (Velazco *et al.*, 2019). The importance of carefully delimiting blocks for fitting and evaluating the models is discussed by Roberts *et al.* (2017).

Table 5: Data partition methods available in the *ENMTML* package.

Data partition method	Type of partition	Acronym used in the `part` argument	Method description	Additional arguments	References
Bootstrap	Random	BOOT	Random partition between training and test subsets	replicates and proportion	Fielding & Bell (1997)
K-Fold	Random	KFOLD	Random partition of occurrences in folds	folds	Fielding & Bell (1997)

Bands	Geographical	BAND	Geographical partition in one-dimension bands	type=1(Latitude) type=2(Longitude)	Bahn & McGill (2013)
Block	Geographical	BLOCK	Geographical partition in two-dimensions (checkerboard)	-	Roberts *et al.* (2017)

2.7 Measure of models' extrapolation

```
Arguments involved: (extrapolation)
```

ENMs are fitted based on conditions found in occurrences and absence/pseudo-absence/background data. When making predictions, it is not uncommon for models to predict onto new conditions (non-analog climates), especially when performing projections to other time periods or geographical regions. In those situations, models will perform extrapolations, which means that there is some uncertainty as models were not fitted on those environmental conditions (Fitzpatrick & Hargrove, 2009). To identify geographical locations in which models are performing extrapolations, we included a Mobility-Oriented Parity analysis (MOP; Owens *et al.*, 2013), which is based on the defined accessible area for each species. If there is no accessible area, the program calculates MOP based on all conditions within the geographical extent of predictors. Example of articles that discuss the main issues caused by model extrapolation are discussed by Elith *et al.* (2010) and Owens *at al.* (2013).

2.8 Modeling algorithms

```
Arguments involved: (algorithm)
```

As one of the primary sources of ENMs/SDMs uncertainty is the method used to construct them (Watling *et al.*, 2015; Thuiller *et al.*, 2019), and assuming that no single methods can lead with all modeling situation (Qiao *et al.*, 2015), our ***ENMTML*** package fit 13 algorithms that range different statistical techniques and type of data used to fit the models (Table 5).

Table 6: Algorithms used by the *ENMTML* package to construct ecological niche and species distribution models.

Algorithm	**Acronym used in the `algorithms` argument**	**Package**	**Data used to create models**	**Reference package**
Bioclim (Envelope Score)	BIO	*dismo*	Presences	Hijmans *et al.* (2017)
Mahalanobis	MAH	*dismo*	Presences	Hijmans *et al.* (2017)
Domain	DOM	*dismo*	Presences	Hijmans *et al.* (2017)
Generalized Linear Models	GLM	*stats*	Presences and pseudo-absences	R Core Team (2018)
Generalized Additive Models	GAM	*gam*	Presences and pseudo-absences	Hastie (2018)
Support Vector Machine	SVM	*kernlab*	Presences and pseudo-absences	Karatzoglou *et al.* (2004)
Boosted Regression Trees	BRT	*dismo*	Presences and pseudo-absences	Hijmans *et al.* (2017)
Random Forest	RDF	*randomForest*	Presences and pseudo-absences	Liaw & Wiener (2002)
Maximum Likelihood	MLK	*maxlike*	Presences and background points	Royle *et al.* (2012)
Bayesian Gaussian Process	GAU	GRaF	Presences and pseudo-absences	Golding (2014)
Maximum Entropy simple (only linear and quadratic features)	MXS	maxnet	Presences and background points	Phillips (2017)
Maximum Entropy default (all features)	MXD	maxnet	Presences and background points	Phillips (2017)
Ecological Niche Factor Analysis	ENF	adehabitatHS	Presences and background points	Calenge (2006)

2.9 Model evaluation

Model evaluation is performed using seven different metrics: Area Under the Curve (AUC, (Fielding & Bell, 1997), Kappa (Cohen, 1960), True Skill Statistic (Allouche *et al.*, 2006), Jaccard (Leroy *et al.*, 2018), Sorensen (Leroy *et al.*, 2018), F_{pb} (Li & Guo, 2013), Boyce (Boyce *et al.*, 2002), partial ROC and its respective p-value (Peterson *et al.*, 2008), omission rate (OR; Fielding & Bell, 1997) and proportion of the total area in which species is considered to be present (Peterson, 2001). The values at the table are an average of the several replicates (if the bootstrap partition was chosen), folds (if random k-folds were chosen), or geographical subsets (if bands or block partition was chosen), accompanied by the respective standard deviation. Metrics are given for each algorithm used to fit models for each species, and each threshold chosen to create binary maps. The type of partition used to create occurrence subsets is also indicated (Table 7).

Table 7: Example of an evaluation table output for models created using the algorithm Maxent for two different species and two different thresholds evaluated by a random Bootstrap partition.

Sp	**Alg**	**Part**	**Thr**	**AUC**	**Kappa**	**TSS**	**Jaccard**	**Sorensen**	**Fpb**	**pROC**	**OR**	**%Area**	**Boyce**
Sp_18	MXS	BOOT	MAX_TSS	0.995	0.950	0.950	0.954	0.976	1.909	1.754	0.240	65.765%	1.000
Sp_18	MXS	BOOT	LPT	0.990	0.929	0.928	0.938	0.966	1.875	1.675	0.000	78.345%	0.831
Sp_34	MXS	BOOT	MAX_TSS	0.998	0.966	0.966	0.969	0.984	1.938	1.876	0.120	72.972%	0.807
Sp_34	MXS	BOOT	LPT	0.990	0.929	0.928	0.938	0.966	1.875	1.290	0.000	87.029%	0.831

Sp	**AUC_SD**	**Thr**	**Kap_SD**	**TSS_SD**	**Jacc_SD**	**Sor_SD**	**Fpb_SD**	**pROC_SD**	**OR_SD**	**%Area_SD**	**Boyce_SD**
Sp_18	0.007	MAX_TSS	0.071	0.071	0.064	0.034	0.129	0.023	0.120	2.875%	0.002
Sp_18	0.014	LPT	0.101	0.101	0.088	0.047	0.177	0.042	0.014	5.897%	0.015
Sp_34	0.003	MAX_TSS	0.047	0.047	0.044	0.023	0.088	0.054	0.028	3.471%	0.023
Sp_34	0.003	LPT	0.047	0.047	0.044	0.023	0.088	0.076	0.270	9.743%	0.042

2.10 Threshold for binary maps

```
Arguments involved: (thr)
```

The different thresholds are used to create binary maps, being that more than one option can be chosen, which results in different sets of binary maps created within a single script run (Table 8). The thresholds are chosen based on the suitability value that maximizes a given metric. For instance, the MAX_TSS threshold uses the suitability value that gives the highest TSS value to create binary maps. This is the common threshold at which the sum of Specificity and Sensitivity is maximum. The same logic stands for all the other alternatives, except for Lowest Presence Threshold (LPT; Pearson, 2007) and Sensitivity. LPT threshold establishes a threshold value in which suitability is the lowest among all occurrence data. Sensitivity requires users to specify a desired sensitivity value for the resulting binary map (Table 8).

Table 8: Threshold for binary maps available in the *ENMTML* package.

Chosen	**Acronym**	**Method description**	**Additional arguments**	**References**

metric for threshold definition	used in the `thr` argument			
Least Presence Threshold	LPT	Lowest suitability value among occurrence data	-	Pearson (2007)
True Skill Statistics	MAX_TSS	Suitability value that maximizes the TSS	-	Allouche *et al.* (2006)
Kappa	MAX_KAPPA	Suitability value that maximizes the Kappa	-	Allouche *et al.* (2006)
Sensitivity	SENSITIVITY	Suitability value that results in the specified sensitivity value	sens	-
Jaccard	JACCARD	Suitability value that maximizes the Jaccard Index	-	Leroy *et al.* (2018)
Sorensen	SORENSEN	Suitability value that maximizes the Sorensen Index	-	Leroy *et al.* (2018)

2.11 Ensemble methods

`Arguments involved: (ensemble)`

The major source of model uncertainty is caused by the different algorithms used to fit ENMs (Diniz-Filho *et al.*, 2009; Thuiller *et al.*, 2019). A commonly used method to deal with this is to create an ensemble model of different algorithms (Araújo & New, 2007; Marmion *et al.*, 2009). *ENMTML* offers six ensemble methods, three based on different ways to calculate models` average and three based on PCA derived from the models. Average-based ensembles can be created using: i) a simple average of all models, ii) weighted average, in which models` suitability is weighted by how well that algorithm performed and iii) superior average, in which a simple average is calculated only for those algorithms that performed better than the average of all algorithms. PCA-based ensemble performs a principal components analysis on suitability maps and uses the first component as the final map, this can be performed: i) using all models, ii) using only the superior models, selected similarly to the superior average, and iii) principal components are calculated using only suitability values above the threshold for each algorithm, values below the threshold are set to zero (Table 9).

Table 9: Ensemble methods available in the *ENMTML* package.

Ensemble method	Acronym used in the `ensemble` argument	Method description	Reference
None	NULL	No ensemble is performed	-
Mean	MEAN	Simple average of suitability predicted by different algorithms	Thuiller *et al.* (2009)
Weighted mean	W_MEAN	Average of suitability values weighted by the performance of the algorithms (TSS)	Thuiller *et al.* (2009)
Mean of the best models	SUP	Average of the best algorithms, i.e., those with TSS over the average for a single species	Velazco *et al.*(2019)
Principal Component Analysis (PCA)	PCA	Performs a PCA with algorithms suitability and returns the eigenvalues of the first principal component	Thuiller (2004)
Principal Component Analysis with the best models	PCA_SUP	Performs a PCA with the suitability of the best algorithms, i.e., those with TSS over the average for a single species, and returns the eigenvalues of the first principal component	-
Principal Component Analysis with threshold	PCA_THR	Performs a PCA with suitability values above thresholds used to binarize each algorithm	-

2.12 Methods to constrain ENMs

`Arguments involved: (msdm)`

There is an underlying difference between ecological niche models (ENMs) and species distribution models (SDMs), being that both the niche and the distribution are more suitable to answer different questions (Peterson & Soberón, 2012). Usually, models' output represents the niche (ENMs), being that methods that bring ENMs closer to SDMs, called here MSDM, is a topic lightly treated on species distribution (Mendes et al., *in prep*). MSDM procedures are grouped in two approaches, *a priori* and *a posteriori* methods. The first set of techniques creates geographic variables that are incorporated as predictors for ENMs fitting (Allouche *et al.*, 2008). The second set of methods constrains generated species suitability patterns using estimates of site accessibility, not being included as predictors while fitting models (Mendes et al., *in prep*).

Table 10: Spatial restriction (MSDM) methods available in the *ENMTML* package.

Method type	Method names	Acronym used in the `msdm` argument	Characteristic	Reference
None	None	NULL	Does not constrain ENMs	-
A priori	Latlong	XY	Create two layers with latitude and longitude values	Allouche *et al.* (2008)
	Minimum distance	MIN	Create a layer with the distance of each cell to the closest occurrence	Allouche *et al.* (2008)
	Cumulative distance	CML	Create a layer with information of the summed distance from each cell to all occurrences	Allouche *et al.* (2008)
	Kernel	KER	Create a layer with a Gaussian-Kernel on the occurrence data	Allouche *et al.* (2008)
A posteriori	Occurrences Based Restriction	OBR	Uses the distance between points to exclude far suitable patches	Mendes *et al.* (*in prep*)
	Lower Quantile	LR	Select 25% of suitability patches without presences that are nearest suitability patches with presences	Mendes *et al.* (*in prep*)
	Presence	PRES	Select only the patches with confirmed occurrence data	Mendes *et al.* (*in prep*)
	Minimum Convex Polygon	MCP	Excludes suitable cells outside the minimum convex polygon of the occurrence data	Kremen *et al.* (2008)
	Buffered Minimum Convex Polygon	MCP-B	Creates a buffer around the MCP	Kremen *et al.* (2008)

2.13 Parallel processing

`Arguments involved: (cores)`

The *ENMTML* package has the option to fit models using parallel processing, which accelerates the process. However, as this is computation-intensive, we chose to leave it open for users to decide the number of computer cores allocated for fitting ENMs. If the users do not specify the number of cores, only a single core will be used.

2.14 Output & Folders

`Arguments involved: (save_part / save_final)`

There are several possible outputs for a single run of the *ENMTML* package. All the outputs produced by the fitting process are within a *Result* folder, which is created at the same level as the Predictors folders (Figure 2). Within the *Result* folder, there is a sub-folder named *Algorithm* that contains the suitability and binary maps produced for each algorithm for each species. If the user chose to create ensemble models, there is another subfolder named *Ensemble*, with the combined maps created for each ensemble type chosen by the user. If the user chose to perform projections to different geographical regions or time periods there will also be a sub-folder named *Projection*, within which are the sub-folders for each projection scenario, with contains suitability maps generated for all the algorithms and the ensemble of those algorithms, if the user-specified an ensemble method. Users can control if partial and final models will be saved, altering the arguments `save_part` and `save_final` (TRUE/FALSE).

Files generated at the pre-processing stage are also within the *Results* folder. Accessible area masks for each species are found within the *Extent_Masks* sub-folder. Masks used to constrain pseudo-absence allocation are also saved within *Results*, i.e., if the user chose to restrict pseudo-absences allocation using an environmental constraint, there will be a sub-folder named *Env_Constrain* which indicates valid areas for pseudo-absence allocation. Finally, if the user chose to perform a geographical partition of the occurrence dataset, there will be a corresponding sub-folder named *BLOCK* or *BANDS*, with the areas used to delimit each occurrence subset.

Other than the folders, there is also a series of TXT (tab-delimited) files within the *Results* folder. The main ones are the *Evaluation_Table*, which contains the results for model evaluation; *Thresholds* contains the suitability values used to create the binary maps, and *InfoModelling* provides a summary of the arguments used to fit the model. Other than those, other useful files are *Number_Unique_Occurrences*, which specifies the number of unique occurrences for each species; *Occurrences_Cleaned* and *Occurrences_Filtered* returns the datasets produced after occurrences went through the unique occurrences and thinning steps;

Occurrences_Fitting and *Occurrences_Evaluation* returns the dataset used for fitting and evaluating the models; *Moran_and_Mess* files have information about the Moran`s I and the environmental similarity (MESS) calculated between subsets, available both for random and geographical partition.

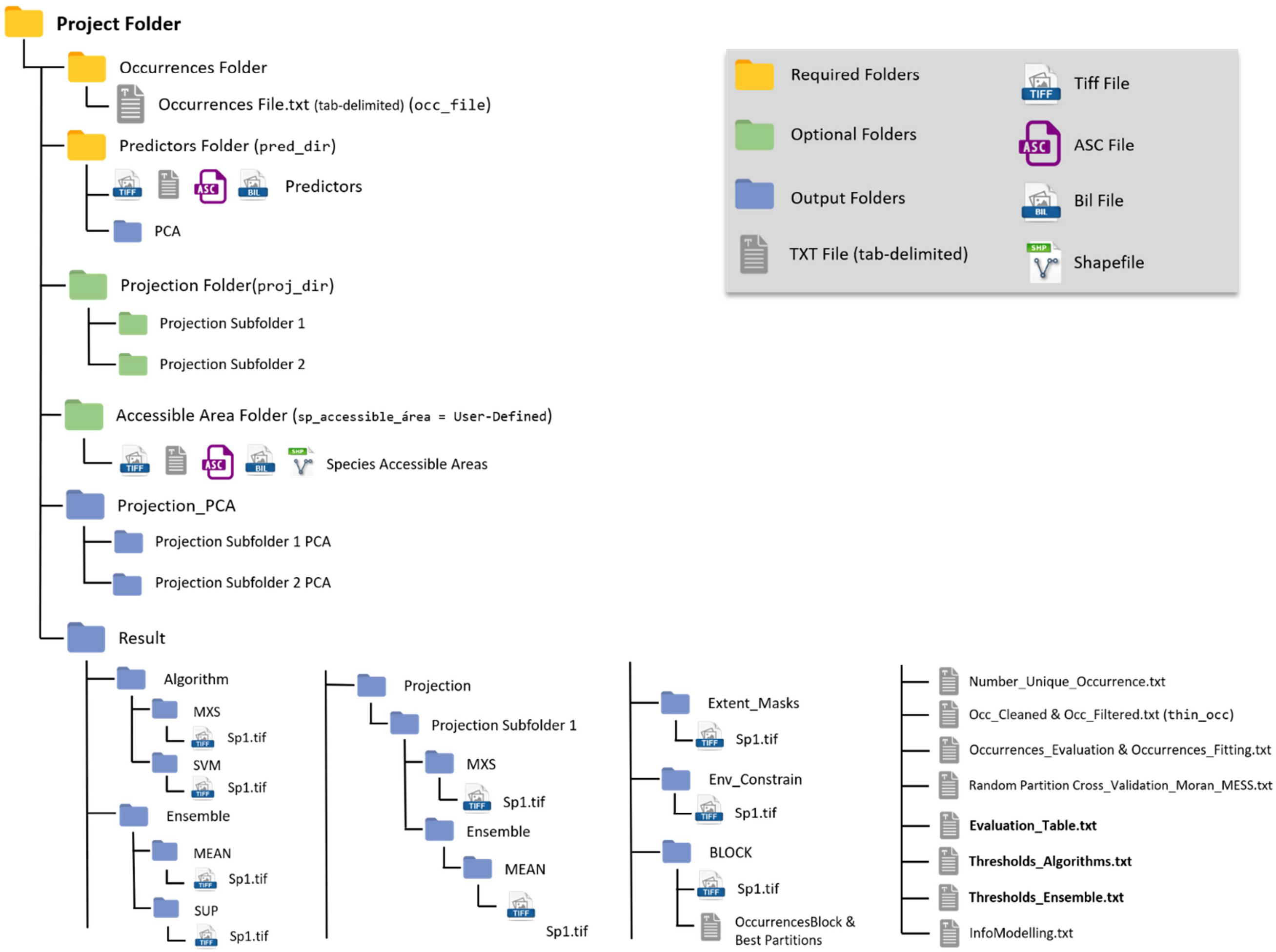


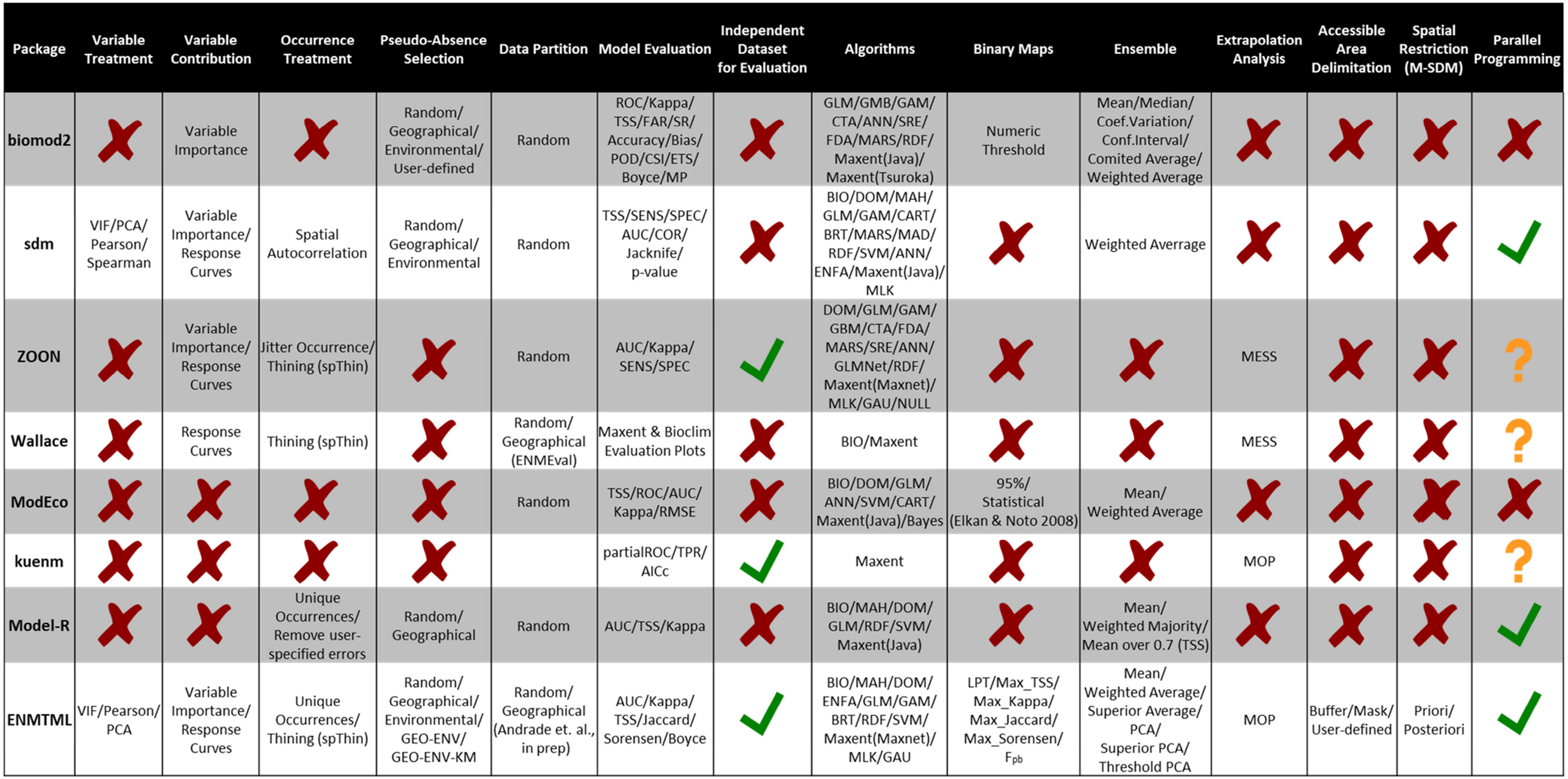
Figure 2: All folders and subfolders involved in a single run of the *ENMTML* package. Yellow folders (occurrence and predictors) are mandatory to run the main function. Green folders (projection and accessible area) are optional and will be required according to the modeling objective. Blue folders are produced by the script, is that most outputs are within the main Results folders, which contains a set of TXT files with model evaluation and information and sub-folders with the models produced by each algorithm and ensemble methods. Folders related to the accessible area, pseudo-absence allocation, and geographical partition are also created to avoid repeating those analyses in the future.

3. Comparison with other packages and innovations

There are several R packages to fit ENMs. We performed a literature search and found seven alternatives: *biomod* (Thuiller *et al.*, 2009), *ModEco* (Guo & Liu, 2010), *sdm* (Naimi & Araújo, 2016), *Model-R* (Sánchez-Tapia *et al.*, 2018), *Wallace* (Kass *et al.*, 2018), *ZOON* (Golding *et al.*, 2018), and *kuenm* (Cobos *et al.*, 2019). We summarize those packages in a table, highlighting each package features and contrasting them with the features available at *ENMTML* (Table 10). Most packages focus on the development of a specific aspect of the modeling process, e.g., the package *biomod* was proposed as a platform for creating ensemble models, while the package *kuenm* is heavily focused towards accurately developing Maxent models; therefore a crucial aspect of software/package selection lies on the study objective.

We introduce the package *ENMTML,* which proposes to integrate complex methodological developments in the ENMs' field, published from several different sources, in a single package and make them visible for users, which are not accustomed to the methodological details of ENMs. Our secondary objective was to make the package user-friendly, even for people not comfortable with the programming environment; therefore, we summarized the whole process into one single function with arguments that must be filled by the user according to the study objectives. We covered the majority of the ENMs process, from pre-processing occurrences and predictors to post-processing suitability models into ensembles or MSDM and provided several methodological alternatives to the different modeling steps (Table 10).

Table 10: Comparison of features included in seven R packages used for fitting ENMs and the *ENMTML* package.



Package	Variable Treatment	Variable Contribution	Occurrence Treatment	Pseudo-Absence Selection	Data Partition	Model Evaluation	Independent Dataset for Evaluation	Algorithms	Binary Maps	Ensemble	Extrapolation Analysis	Accessible Area Delimitation	Spatial Restriction (M-SDM)	Parallel Programming
biomod2	✗	Variable Importance	✗	Random/ Geographical/ Environmental/ User-defined	Random	ROC/Kappa/ TSS/FAR/SR/ Accuracy/Bias/ POD/CSI/ETS/ Boyce/MP	✗	GLM/GMB/GAM/ CTA/ANN/SRE/ FDA/MARS/RDF/ Maxent(Java)/ Maxent(Tsuroka)	Numeric Threshold	Mean/Median/ Coef.Variation/ Conf.Interval/ Comited Average/ Weighted Average	✗	✗	✗	✗
sdm	VIF/PCA/ Pearson/ Spearman	Variable Importance/ Response Curves	Spatial Autocorrelation	Random/ Geographical/ Environmental	Random	TSS/SENS/SPEC/ AUC/COR/ Jacknife/ p-value	✗	BIO/DOM/MAH/ GLM/GAM/CART/ BRT/MARS/MAD/ RDF/SVM/ANN/ ENFA/Maxent(Java)/ MLK	✗	Weighted Averrage	✗	✗	✗	✓
ZOON	✗	Variable Importance/ Response Curves	Jitter Occurrence/ Thining (spThin)	✗	Random	AUC/Kappa/ SENS/SPEC	✓	DOM/GLM/GAM/ GBM/CTA/FDA/ MARS/SRE/ANN/ GLMNet/RDF/ Maxent(Maxnet)/ MLK/GAU/NULL	✗	✗	MESS	✗	✗	?
Wallace	✗	Response Curves	Thining (spThin)	✗	Random/ Geographical (ENMEval)	Maxent & Bioclim Evaluation Plots	✗	BIO/Maxent	✗	✗	MESS	✗	✗	?
ModEco	✗	✗	✗	✗	Random	TSS/ROC/AUC/ Kappa/RMSE	✗	BIO/DOM/GLM/ ANN/SVM/CART/ Maxent(Java)/Bayes	95%/ Statistical (Elkan & Noto 2008)	Mean/ Weighted Average	✗	✗	✗	✗
kuenm	✗	✗	✗	✗		partialROC/TPR/ AICc	✓	Maxent	✗	✗	MOP	✗	✗	?
Model-R	✗	✗	Unique Occurrences/ Remove user-specified errors	Random/ Geographical	Random	AUC/TSS/Kappa	✗	BIO/MAH/DOM/ GLM/RDF/SVM/ Maxent(Java)	✗	Mean/ Weighted Majority/ Mean over 0.7 (TSS)	✗	✗	✗	✓
ENMTML	VIF/Pearson/ PCA	Variable Importance/ Response Curves	Unique Occurrences/ Thining (spThin)	Random/ Geographical/ Environmental/ GEO-ENV/ GEO-ENV-KM	Random/ Geographical (Andrade et. al., in prep)	AUC/Kappa/ TSS/Jaccard/ Sorensen/Boyce	✓	BIO/MAH/DOM/ ENFA/GLM/GAM/ BRT/RDF/SVM/ Maxent(Maxnet)/ MLK/GAU	LPT/Max_TSS/ Max_Kappa/ Max_Jaccard/ Max_Sorensen/ F_{pb}	Mean/ Weighted Average/ Superior Average/ PCA/ Superior PCA/ Threshold PCA	MOP	Buffer/Mask/ User-defined	Priori/ Posteriori	✓

4. Example

We used the *ENMTML* package to fit current and future distribution for five virtual species. We only present here the results produced for a single species (full models' outputs can be found in Appendix A). For this example, we used five bioclimatic variables (bio1, bio3, bio4, bio12, and bio15) from the WorldClim database v2.0 (https://www.worldclim.org). We projected the models to 2080 climatic conditions with a Representative Concentration Pathway (RCP) of 8.5. We used the MOHC HadGEM2-ES model and the same bioclimatic variables used in current conditions sourced by GCM Downscaled Data Portal (http://ccafs-climate.org). Current and future variables had ten arcmins of resolution. We performed a Principal Component Analysis (PCA) in the environmental data in order to reduce predictors collinearity (see the details of this procedure in the Methods sub-section "*Predictors input and collinearity reduction*"). We employed Support Vector Machine (SVM), Random Forests (RDF), and Maximum Entropy with default tuning (MXD) as algorithms. We used an equal number of absences and presences (i.e., presences/absences ratio equal to 1), which were randomly allocated within a calibration area (i.e., species accessible area) delimited by a buffer of 500 km around the presences. Models were validated by spatial block cross-validation. For the current condition we constrained the models using the method MCP-B (see Methods sub-section *"Methods to constrain ENMs"*) with a buffer of 200 km around the MCP. Final models were constructed by ensembling all the algorithms with a PCA (see details in Methods sub-section "*Ensemble methods*"). We calculated models' extrapolation for current and future conditions based on Mobility-Oriented Parity (MOP) metric. The total time used for fitting and processing the models of five species employing four cores was 2.545 minutes.

All these procedures are expressed in R command line below:

```
ENMTML(pred_dir = d_env, proj_dir = d_fut, occ_file = d_occ,
       sp = 'species', x = 'x', y = 'y', min_occ = 10, thin_occ = NULL,
       eval_occ = NULL, colin_var = c(method = 'PCA'), imp_var = FALSE,
       sp_accessible_area = c(method='BUFFER', type= '2' , width = '500'),
       pseudoabs_method = c(method = 'RND'), pres_abs_ratio = 1,
       part = c(method= 'BLOCK'), save_part = FALSE,
       save_final = TRUE, algorithm = c('SVM', 'RDF', 'MXD'),
       thr = c(type = 'MAX_TSS'), msdm = NULL,
ensemble = c(method = 'PCA'), extrapolation = FALSE, cores = 1)
```

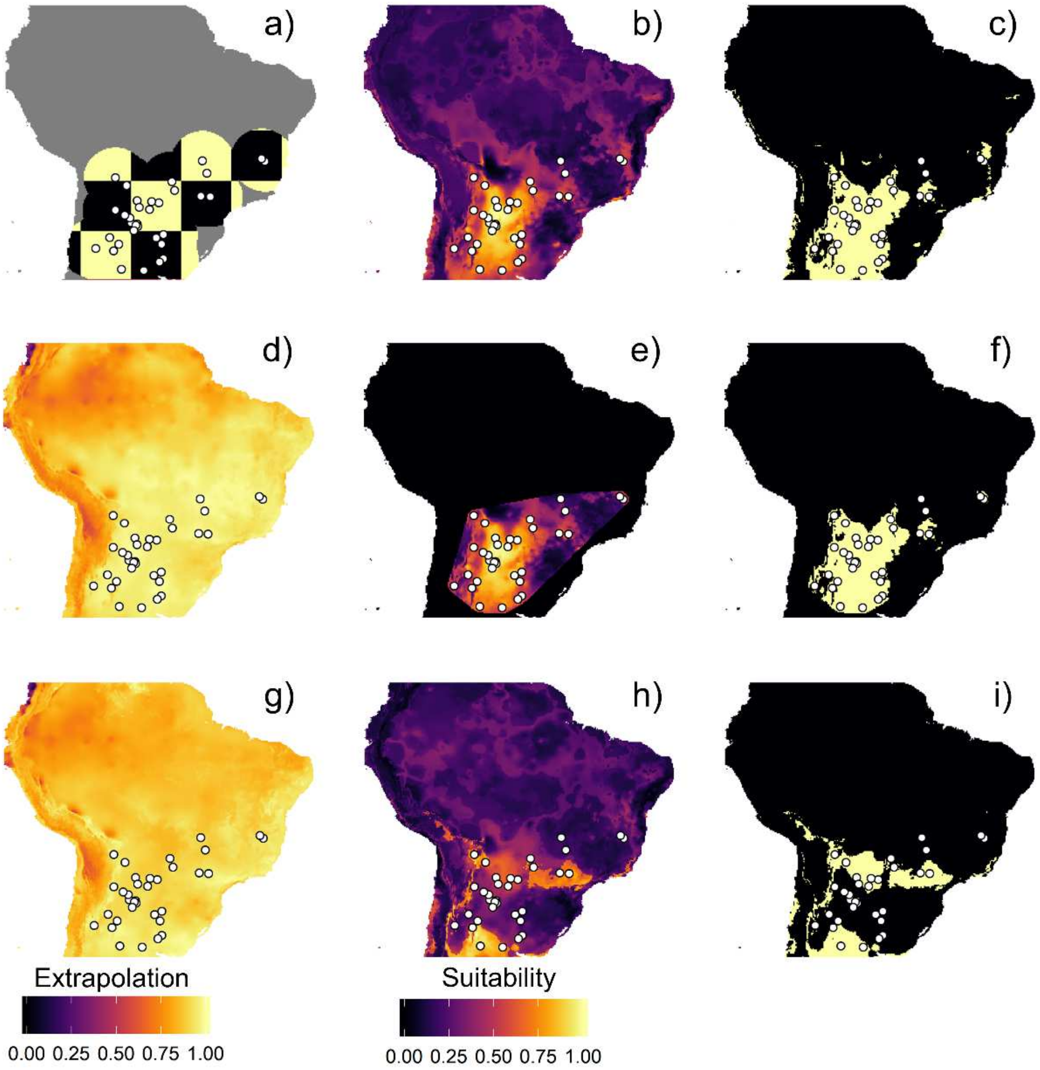


Fig. 3: Some output layers generated by *ENMTML* package. a) The calibration area used to construct the models based on a 200 km buffer around presences (white dots). Black and yellow checkerboard shows the best geographic block partition found for this species occurrences. b) and c) depict continuous and binary suitability patterns without restriction, respectively. d) and g) represent models' extrapolation for current and 2080 (RCP 8.5) environmental conditions, respectively. Extrapolation is based on the Mobility-Oriented Parity metric. The closer to zero, the higher the extrapolation. e) and f) depict continuous and binary suitability patterns constrained by a Minimum Convex Polygon plus a buffer of 200km. h) and i) represent a continuous and binary suitability pattern for 2080 environmental conditions (RCP 8.5). Current and 2080 suitability patterns are ensembled models perfomed by Principal Component Analysis.

5. Future Prospects

We present the release of the *ENMTML* package, but we already have in mind ideas for future implementations. As the main objective of the package is to approach complex methodological developments to people that rely on ENMs but do not focus the development of new methods and are not comfortable using R, in the next update we expect to launch a web platform using Shiny. On the other hand, we also believe that *ENMTML* package might be of great use for the whole ENMs' community, as it centers on methodological developments scattered around the literature, and not always implemented in R, in one single location. With that in mind, we also look forward to providing further options for people who are interested in the fine-tuning of models. One of the first additions already planned is the possibility for users to change algorithms parameters. In addition, we also plan to explore in-depth the ensemble field and include more ensemble alternatives and uncertainty maps. Finally, we believe an important aspect of ENMs is to be clear about model uncertainty; therefore, in the upcoming update, we will implement metrics to calculate source of uncertainty for each species in a way similar to Watling *et al.* (2015). Other than the already planned improvements, users can expect novel methodological approaches published in the literature to be implemented in the future versions of the package and are welcome to contribute with the development of the package and suggest new features.

6.Acknowledgments

We like to use this chance to thank the many people who helped us by testing the package throughout its development. There is not enough space to name them all, but we are greatly indebted to a large number of people. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001. A.F.A.A thanks CAPES for his doctoral research grant. S.J.E.V. thanks to Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) for the postdoctoral fellowship. P.D.M is continuously supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) productivity grants (308694/2015-5). A.F.A.A and S.J.E.V were responsible for the code; A.F.A.A, S.J.E.V, and P.D.M discussed the methodological steps and the respective implementation; S.J.E.V created the package structure and the help file; A.F.A.A led the writing of the manuscript. All authors contributed to the draft, improvement of the ideas and gave final approval for publication.

7.References

Ahmed, S.E., Mcinerny, G., O’Hara, K., Harper, R., Salido, L., Emmott, S. & Joppa, L.N. (2015) Scientists and software - surveying the species distribution modelling community. *Diversity and Distributions*, **21**, 258–267.

Aiello-Lammens, M.E., Boria, R.A., Radosavljevic, A., Vilela, B. & Anderson, R.P. (2015) spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, **38**, 541–545.

Allouche, O., Steinitz, O., Rotem, D., Rosenfeld, A. & Kadmon, R. (2008) Incorporating distance constraints into species distribution models. *Journal of Applied Ecology*, **45**, 599–609.

Allouche, O., Tsoar, A. & Kadmon, R. (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, **43**, 1223–1232.

Anderson, R.P. & Gonzalez, I. (2011) Species-specific tuning increases robustness to sampling bias in models of species distributions: An implementation with Maxent. *Ecological Modelling*, **222**, 2796–2811.

Araújo, M.B. & New, M. (2007) Ensemble forecasting of species distributions. *Trends in ecology & evolution*, **22**, 42–7.

Bahn, V. & McGill, B.J. (2013) Testing the predictive performance of distribution models. *Oikos*, **122**, 321–331.

Barbet-Massin, M., Jiguet, F., Albert, C.H. & Thuiller, W. (2012) Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, **3**, 327–338.

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S.P., Peterson, A.T., Soberón, J. & Villalobos, F. (2011) The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, **222**, 1810–1819.

Boyce, M.S., Vernier, P.R., Nielsen, S.E. & Schmiegelow, F.K.. (2002) Evaluating resource selection functions. *Ecological Modelling*, **157**, 281–300.

Calenge, C. (2006) The package adehabitat for the R software: a tool for the analysis of space and habitat use by animals. *Ecological Modelling*, **197**, 516–519.

Campos, M., de Andrade, A.F.A. & Kunzmann, B. (2014) Modelling of the potential distribution of Limnoperna fortunei (Dunker, 1857) on a global scale. *Aquatic Invasions*, **9**, 253–265.

Carpenter, G., Gillison, A.N. & Winter, J. (1993) DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation*, **2**, 667–680.

Carstens, B.C. & Richards, C.L. (2007) Iintegrating coalescent and ecological niche modeling in comparative phylogeography. *Evolution*, **61**, 1439–1454.

Chifflet, L., Rodriguero, M.S., Calcaterra, L.A., Rey, O., Dinghi, P.A., Baccaro, F.B., Souza, J.L.P., Follett, P. & Confalonieri, V.A. (2016) Evolutionary history of the little fire ant Wasmannia auropunctata before global invasion: Inferring dispersal patterns, niche requirements and past and present distribution within its native range. *Journal of Evolutionary Biology*, **29**, 790–809.

Cobos, M.E., Peterson, A.T., Barve, N. & Osorio-Olvera, L. (2019) kuenm: an R package for detailed development of ecological niche models using Maxent. *PeerJ*, **7**, e6281.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **XX**, 37–46.

Cooper, J.C. & Soberón, J. (2018) Creating individual accessible area hypotheses improves stacked species distribution model performance. *Global Ecology and Biogeography*, **27**, 156–165.

Diniz-Filho, J.A.F., Mauricio Bini, L., Fernando Rangel, T., Loyola, R.D., Hof, C., Nogués-Bravo, D. & Araújo, M.B. (2009) Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. *Ecography*, **32**, 897–906.

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. & Lautenbach, S. (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, **36**, 27–46.

Elith, J., Kearney, M. & Phillips, S. (2010) The art of modelling range-shifting species. *Methods in Ecology and Evolution*, **1**, 330–342.

Engler, R., Guisan, A. & Rechsteiner, L. (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, **41**, 263–274.

Farber, O. & Kadmon, R. (2003) Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. *Ecological Modelling*, **160**, 115–130.

Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence / absence models. *Environmental Conservation*, **24**, 38–49.

Fitzpatrick, M.C. & Hargrove, W.W. (2009) The projection of species distribution models and the problem of non-analog climate. *Biodiversity and Conservation*, **18**, 2255–2261.

Friedman, J. (2001) Greedy Function Approximation : A Gradient Boosting Machine. *The Annals of Statistics*, **29**, 1189–1232.

Golding, N. (2014) GRaF: Species distribution modelling using latent Gaussian random fields.

Golding, N., August, T.A., Lucas, T.C.D., Gavaghan, D.J., van Loon, E.E. & McInerny, G. (2018) The ZOON R package for reproducible and shareable species distribution modelling. *Methods in Ecology and Evolution*, **9**, 260–268.

Golding, N. & Purse, B. V. (2016) Fast and flexible Bayesian species distribution modelling using Gaussian processes. *Methods in Ecology and Evolution*, **7**, 598–608.

Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., Gordon, A., Kujala, H., Lentini, P.E., Mccarthy, M.A., Tingley, R. & Wintle, B.A. (2015) Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography*, **24**, 276–292.

Guo, Q., Kelly, M. & Graham, C.H. (2005) Support vector machines for predicting distribution of Sudden Oak Death in California. *Ecological Modelling*, **182**, 75–90.

Guo, Q. & Liu, Y. (2010) ModEco: an integrated software package for ecological niche modeling. *Ecography*, **33**, 637–642.

Hao, T., Elith, J., Guillera□Arroita, G. & Lahoz□Monfort, J.J. (2019) A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD. *Diversity and Distributions*, 1–14.

Hastie, T. (2018) gam: Generalized Additive Models.

Hastie, T. & Tibshirani, R. (1990) *Generalised additive models*, Chapman and Hall.

Heikkinen, R.K., Luoto, M., Araújo, M.B., Virkkala, R., Thuiller, W. & Sykes, M.T. (2006) Methods and uncertainties in bioclimatic envelope modelling under climate change. *Progress in Physical Geography*, **30**, 751–777.

Hijmans, R.J., Phillips, S., Leathwick, J. & Elith., J. (2017) dismo: Species Distribution Modeling. *Http://Cran.R-Project.Org/Web/Packages/Dismo/*.

Hirzel, a. H., Hausser, J., Chessel, D. & Perrin, N. (2002) Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? *Ecology*, **83**, 2027–2036.

Karatzoglou, A., Smola, A., Hornik, K. & Zeileis, A. (2004) kernlab-an S4 package for kernel methods in R. *Journal of Statistical Software*, **11**, 1–20.

Kass, J.M., Vilela, B., Aiello-Lammens, M.E., Muscarella, R., Merow, C. & Anderson, R.P. (2018) Wallace: A flexible platform for reproducible modeling of species niches and distributions built for community expansion. *Methods in Ecology and Evolution*, **9**, 1151–1156.

Keppel, G., Van Niel, K.P., Wardell-Johnson, G.W., Yates, C.J., Byrne, M., Mucina, L., Schut, A.G.T., Hopper, S.D. & Franklin, S.E. (2012) Refugia: identifying and understanding safe havens for biodiversity under climate change. *Global Ecology and Biogeography*, **21**, 393–404.

Kremen, C., Cameron, A., Moilanen, A., Phillips, S.J., Thomas, C.D., Beentje, H., Dransfield, J., Fisher, B.L., Glaw, F., Good, T.C., Harper, G.J., Hijmans, R.J., Lees, D.C., Louis, E., Nussbaum, R.A., Raxworthy, C.J., Razafimpahanana, A., Schatz, G.E., Vences, M., Vieites, D.R., Wright, P.C. & Zjhra, M.L. (2008) Aligning Conservation Priorities Across Taxa in Madagascar with High-Resolution Planning Tools. *Science*, **320**, 222–226.

Leroy, B., Delsol, R., Hugueny, B., Meynard, C.N., Barhoumi, C., Barbet□Massin, M. & Bellard, C. (2018) Without quality presence – absence data , discrimination metrics such as TSS can be misleading measures of model performance. *Journal of Biogeography*, **45**, 1994–2002.

Li, W. & Guo, Q. (2013) How to assess the prediction accuracy of species presence-absence models without absence data? *Ecography*, **36**, 788–799.

Liaw, A. & Wiener, M. (2002) Classification and Regression by randomForest. *R News*, **2**, 18–22.

Lins, D.M., de Marco, P., Andrade, A.F.A. & Rocha, R.M. (2018) Predicting global ascidian invasions. *Diversity and Distributions*, **24**, 692–704.

Lobo, J.M., Jiménez-Valverde, A. & Hortal, J. (2010) The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, **33**, 103–114.

Lobo, J.M., Jiménez-Valverde, A. & Real, R. (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.

De Marco, P. & Nóbrega, C.C. (2018) Evaluating collinearity effects on species distribution models: An approach based on virtual species simulation. *PLOS ONE*, **13**, e0202403.

Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R.K. & Thuiller, W. (2009) Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions*, **15**, 59–69.

Marquaridt, D.W. (1970) Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, **12**, 591–612.

Mateo, R.G., Felicísimo, Á.M. & Muñoz, J. (2010) Effects of the number of presences on reliability and stability of MARS species distribution models: The importance of regional niche variation and ecological heterogeneity. *Journal of Vegetation Science*, **21**, 908–922.

McCullagh, P. & Nelder, J.A. (1989) *Generalized Linear Models*, Chapman and Hall.

Muscarella, R., Galante, P.J., Soley-Guardia, M., Boria, R.A., Kass, J.M., Uriarte, M. & Anderson, R.P. (2014) ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for MAXENT ecological niche models. *Methods in Ecology and Evolution*, **5**, 1198–1205.

Naimi, B. & Araújo, M.B. (2016) sdm: a reproducible and extensible R platform for species distribution modelling. *Ecography*, **39**, 368–375.

Nix, H.A. (1986) *A biogeographic analysis of Australian elapid snakes*. *Australian Flora and Fauna Series No. 7*, pp. 4–15. Australian Government Publishing Service.

Owens, H.L., Campbell, L.P., Dornak, L.L., Saupe, E.E., Barve, N., Soberón, J., Ingenloff, K., Lira-Noriega, A., Hensz, C.M., Myers, C.E. & Peterson, A.T. (2013) Constraints on interpretation of ecological niche models by limited environmental ranges on calibration areas. *Ecological Modelling*, **263**, 10–18.

Pearson, R.G. (2007) Species' Distribution Modeling for Conservation Educators and Practitioners. *Lessons in Conservation*, **3**, 54–89.

Peterson, A.T. (2003) Predicting the Geography of Species' Invasions Via Ecological Niche Modeling. *The Quarterly Review of Biology*, **78**, 419–433.

Peterson, A.T. & Holt, R.D. (2003) Niche differentiation in Mexican birds: Using point occurrences to detect ecological innovation. *Ecology Letters*, **6**, 774–782.

Peterson, A.T., Papeş, M. & Soberón, J. (2008) Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling*, **213**, 63–72.

Peterson, A.T., Sánchez-Cordero, V., Soberón, J., Bartley, J., Buddemeier, R.W. & Navarro-Sigüenza, A.G. (2001) Effects of global climate change on geographic distributions of Mexican Cracidae. *Ecological Modelling*, **144**, 21–30.

Peterson, A.T. & Shaw, J. (2003) Lutzomyia vectors for cutaneous leishmaniasis in Southern Brazil: Ecological niche models, predicted geographic distributions, and climate change effects. *International Journal for Parasitology*, **33**, 919–931.

Peterson, A.T. & Soberón, J. (2012) Species distribution modeling and ecological niche modeling: Getting the Concepts Right. *Natureza e Conservacao*, **10**, 102–107.

Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R.P., Martínez-Meyer, E., Nakamura, M. & Bastos Araujo, M. (2011) *Ecological niches and geographic distributions*, Princeton University Press.

Peterson, T.A. (2001) Predicting Species' Geographic Distributions Based on Ecological Niche Modeling. *The Condor*, **103**, 599–605.

Phillips, S. (2017) maxnet: Fitting "Maxent" Species Distribution Models with "glmnet."

Phillips, S., Anderson, R. & Schapire, R. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Prasad, A.M., Iverson, L.R. & Liaw, A. (2006) Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems*, **9**, 181–199.

van Proosdij, A.S.J., Sosef, M.S.M., Wieringa, J.J. & Raes, N. (2015) Minimum required number of specimen records to develop accurate species distribution models. *Ecography*, n/a-n/a.

Qiao, H., Soberón, J. & Peterson, A.T. (2015) No silver bullets in correlative ecological niche modelling: insights from testing among many potential algorithms for niche estimation. *Methods in Ecology and Evolution*, **6**, 1126–1136.

R Core Team (2018) R: A language and environment for statistical computing. *R Foundation for Statistical Computing*.

Razgour, O., Taggart, J.B., Manel, S., Juste, J., Ibáñez, C., Rebelo, H., Alberdi, A., Jones, G. & Park, K. (2018) An integrated framework to identify wildlife populations under threat from climate change. *Molecular Ecology Resources*, **18**, 18–31.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F. & Dormann, C.F. (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 1–17.

Royle, J.A., Chandler, R.B., Yackulic, C. & Nichols, J.D. (2012) Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution*, **3**, 545–554.

Sánchez-Tapia, A., de Siqueira, M.F., Lima, R.O., Barros, F.S.M., Gall, G.M., Gadelha, L.M.R., da Silva, L.A.E. & Osthoff, C. (2018) Model-R: A framework for scalable and reproducible ecological niche modeling. *Communications in Computer and Information Science*, **796**, 218–232.

Senay, S.D., Worner, S.P. & Ikeda, T. (2013) Novel Three-Step Pseudo-Absence Selection Technique for Improved Species Distribution Modelling. *PLoS ONE*, **8**, e71218.

Soberon, J. & Peterson, A.T. (2005) Interpretation of Models of Fundamental Ecological Niches and Species' Distributional Areas. *Biodiversity Informatics*, **2**, 1–10.

Soberón, J.M. (2010) Niche and area of distribution modeling: a population ecology perspective. *Ecography*, **33**, 159–167.

Thuiller, W. (2004) Patterns and uncertainties of species' range shifts under climate change. *Global Change Biology*, **10**, 2020–2027.

Thuiller, W., Guéguen, M., Renaud, J., Karger, D.N. & Zimmermann, N.E. (2019) Uncertainty in ensembles of global biodiversity scenarios. *Nature Communications*, **10**, 1446.

Thuiller, W., Lafourcade, B., Engler, R. & Araújo, M.B. (2009) BIOMOD - A platform for ensemble forecasting of species distributions. *Ecography*, **32**, 369–373.

Velazco, S.J.E., Villalobos, F., Galvão, F. & De Marco Júnior, P. (2019) A dark scenario for Cerrado plant species: Effects of future climate, land use and protected areas ineffectiveness. *Diversity and Distributions*, **25**, 660–673.

Veloz, S.D. (2009) Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *Journal of Biogeography*, **36**, 2290–2299.

Watling, J.I., Brandt, L.A., Bucklin, D.N., Fujisaki, I., Mazzotti, F.J., Romañach, S.S. & Speroterra, C. (2015) Performance metrics and variance partitioning reveal sources of uncertainty in species distribution models. *Ecological Modelling*, **309–310**, 48–59.

Wisz, M.S. & Guisan, A. (2009) Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. *BMC Ecology*, **9**, 8.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., Elith, J., Dudík, M., Ferrier, S., Huettmann, F., Leathwick, J.R., Lehmann, A., Lohmann, L., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M.C., Phillips, S.J., Richardson, K.S., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S.E. & Zimmermann, N.E. (2008) Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, **14**, 763–773.

Zaniewski, A.E., Lehmann, A. & Overton, J.M. (2002) Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157**, 261–280.

Highlights

- We present *ENMTML*, an open source R package to fit ecological niche models (ENMs)
- The package covers a wide variety of methodological aspects gathered from several studies
- Complex methodological features, which were not readily available in R, are now easily accessible to users
- We condense all this complexity in a single function to make it easier for users to follow a workflow
- We demonstrate an example of fitting models for four species with complex methodological choices and its interactions

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: