

Oct 03, 2025

Table protocol about species model distribution (SDM) and following the Zurell, et al 2020.

DOI

<https://dx.doi.org/10.17504/protocols.io.81wgbwqrogpk/v1>

Nathália ernandes Canassa¹, Carlos A. Peres², Célia ristina Clemente Machado³, Helder Farias P. de Araujo¹

¹Universidade Federal da Paraíba; ²University of East Anglia; ³Universidades Estadual da Paraíba



nfcanassabio

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.81wgbwqrogpk/v1>

Protocol Citation: Nathália ernandes Canassa, Carlos A. Peres, Célia ristina Clemente Machado, Helder Farias P. de Araujo 2025. Table protocol about species model distribution (SDM) and following the Zurell, et al 2020.. **protocols.io**

<https://dx.doi.org/10.17504/protocols.io.81wgbwqrogpk/v1>

License: This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited



Protocol status: Working

We use this protocol and it's working

Created: October 02, 2025

Last Modified: October 03, 2025

Protocol Integer ID: 228881

Keywords: table protocol about species model distribution, protocol about species model distribution, species model distribution, table protocol, sdm, species, protocol, table, model

Abstract

Protocol about species model distribution (SDM) and following the Zurell, et al 2020.

Model workflow

- 1 Model fitting was performed using multiple bootstrap replicates (10 repetitions), which enhances the model's reliability and reduces the impact of random variability in the data. Model performance was evaluated using the area under the curve (AUC) of the receiver operating characteristic (ROC), which ensures that the model is making accurate and reliable predictions. The correlation between environmental variables was tested, and highly correlated variables were removed to reduce collinearity, which could compromise model accuracy. To delimit the species' occurrence areas, a MaxEnt threshold was applied, which defines a cutoff value on the habitat suitability index. This threshold is used to convert the continuous predictions into a binary representation, where areas with values above the threshold are considered suitable for the species, while those below the cutoff are considered unsuitable. The threshold selection was based on model performance analysis (e.g., using ROC curves or maximizing accuracy), ensuring that the chosen cutoff realistically reflected the species' potential occurrence areas. This helps to identify regions where the species is most likely to occur, based on the predicted environmental conditions.

Software, codes and data

- 2 Software: Analyses were conducted in R version 3.5.3. The MaxEnt version 3.4.3 using packages "dismo" version 1.3-5
Data availability: Data are available in an open, online, digital repository (DOI: 10.17632/bdmyzxtkdm.1)

Biodiversity data

- 3 Taxon names: All species are listed in the Supplementary Information (Table S1)
Ecological level: Species level
Biodiversity data source: Data were derived from literature sources, but all data are openly available and platforms to supplement spatially explicit data on species occurrences such as Gbif and Species Link
Sampling design: Web of Science, and ScienceDirect databases, employing word combinations (in English and Portuguese) such as "medium- to large-bodied mammals", "checklist", "Caatinga", "hunting", and "ethnofauna" to access any available information on MLBM assemblages across the Caatinga. We also used the Global Biodiversity Information Facility (GBIF – <https://www.gbif.org>) and the Specieslink (<https://specieslink.net>) platform to supplement spatially explicit data on species occurrences.
Sample size per taxon: Total 51 species and 8169 geographic coordinates
Country/region: We compiled all the data for South America.

Details on scaling: All occurrence data were selected from across South America and later cropped to match the regional boundaries of the Caatinga domain, using de Software QGIS, Madeira version 4.4.14.

Details on data cleaning/filtering steps: We searched for the names of all 51 species known to occur in the Caatinga region of South America. We compiled all the data into a single table and verified the geographic coordinates for each species using QGIS software (Madeira version 4.4.14). We only used geographic coordinates within the South American limits, excluding any localities outside this boundary, duplicate localities, and locations within 2 km of each other for the same species.

Data partitioning

- 4 Selection of training data (for model fitting): We conducted 10 replicate analysis for each species based on a 25% bootstrap of available occurrence data.

Predictor variables

- 5 Diurnal amplitude of mean temperature (BIO2), temperature seasonality (BIO4), mean temperature of the wettest (BIO8) and driest (BIO09) quarter, annual precipitation (BIO12), precipitation seasonality (BIO15), precipitation of the driest (BIO17), warmest (BIO18) and coldest (BIO19) quarter, and terrain elevation
Details on data sources: Current WorldClim version 2.1 (www.worldclim.org)
Spatial resolution and spatial extent of raw data: Spatial resolution of 2.5 arcminutes ($\approx$ 4,630 m)
Map Spatial Reference: XY Coordinate System - GCS WGS 1984. Datum - D WGS 1984

Variable pre-selection

- 6 Details on pre-selection of variables: we used 19 climate variables along with elevation data extracted from WorldClim version 2.1 (www.worldclim.org), at a spatial resolution of 2.5 arcminutes ($\approx$ 4,630 m). WorldClim provides high-resolution interpolated climate data, derived from 9,000 to 60,000 weather stations worldwide, aggregated across a target temporal range of 1970–2000.

Multicollinearity

- 7 To avoid redundancy, we included in the models all variables that were correlated with each other by less than 0.8, using a Pearson correlation matrix calculated using the R package "vegan" (version 3.5.3).

Model settings



- 8 We used Receiver Operator Characteristic (ROC) statistics to assess model accuracy, with 10 replicates of 10,000 maximum iterations, 10% of the average replicates were randomized as test data, while the remainder were randomized to train the model during each replicate.

Model estimates

- 9 Assessment of variable importance: we used the jackknife option to identify variables not contributing importantly to model robustness.

Threshold selection

- 10 Details on threshold selection: We selected the thresholds for each species that defined the smallest potential habitat following a conservative approach to avoid overestimating species geographic distributions (Table S2).

Assessment

- 11 Performance statistics estimated on training data: performance statistics were estimated using the training data during each bootstrap replicate. For each of the 10 repetitions, MaxEnt evaluated model accuracy by comparing predicted and observed occurrences. Standard metrics such as AUC (Area Under the Curve) were calculated to assess model performance, ensuring robust and reliable predictions. The use of multiple replicates enhances the stability of the performance estimates and minimizes the risk of overfitting. Performance statistics estimated on validation data (from data partitioning): although no explicit data partitioning was applied, we used a random test point during each bootstrap replicate. These random test points served as a form of validation by evaluating the model's performance on data that was not used during model fitting. Performance statistics, such as AUC, were calculated based on these test points, providing an estimate of model accuracy and its ability to generalize to unseen data. Response plots: We used partial dependence plots to check the ecological plausibility of fitted relationships in MaxEnt models.

Prediction

- 12 Prediction unit: Predictions of relative probability of presence
Post-processing: after thresholds selection, clipping was performed to generate binary maps.

Uncertainty quantification

- 13 Algorithmic uncertainty, if applicable: None



Uncertainty in input data, if applicable: None

Effect of parameter uncertainty, error propagation, if applicable: None

Acknowledgements

Variable pre-selection

- **Details on pre-selection of variables:** we used 19 climate variables along with elevation data extracted from WorldClim version 2.1 (www.worldclim.org), at a spatial resolution of 2.5 arcminutes ($\approx 4,630$ m). WorldClim provides high-resolution interpolated climate data, derived from 9,000 to 60,000 weather stations worldwide, aggregated across a target temporal range of 1970–2000

Multicollinearity

- To avoid redundancy, we included in the models all variables that were correlated with each other by less than 0.8, using a Pearson correlation matrix calculated using the R package “vegan” (version 3.5.3)

Model settings

- We used Receiver Operator Characteristic (ROC) statistics to assess model accuracy, with 10 replicates of 10,000 maximum iterations, 10% of the average replicates were randomized as test data, while the remainder were randomized to train the model during each replicate

Model estimates

- **Assessment of variable importance:** we used the jackknife option to identify variables not contributing importantly to model robustness

Threshold selection

- **Details on threshold selection:** We selected the thresholds for each species that defined the smallest potential habitat following a conservative approach to avoid overestimating species geographic distributions (Table S2)

ASSESSMENT

Performance statistics

- **Performance statistics estimated on training data:** performance statistics were estimated using the training data during each bootstrap replicate. For each of the 10 repetitions, MaxEnt evaluated model accuracy by comparing predicted and observed occurrences. Standard metrics such as AUC (Area Under the Curve) were calculated to assess model performance, ensuring robust and reliable predictions. The use of multiple replicates enhances the stability of the performance estimates and minimizes the risk of overfitting.

- **Performance statistics estimated on validation data (from data partitioning):** although no explicit data partitioning was applied, we used a random test point during each bootstrap replicate. These random test points served as a form of validation by evaluating the model's performance on data that was not used during model fitting. Performance statistics, such as AUC, were calculated based on these test points, providing an estimate of model accuracy and its ability to generalize to unseen data.

Plausibility check

- **Response plots:** We used partial dependence plots to check the ecological plausibility of fitted relationships in MaxEnt models.



PREDICTION

Prediction output

- **Prediction unit:** Predictions of relative probability of presence
- **Post-processing:** after thresholds selection, clipping was performed to generate binary maps.

Uncertainty quantification

- **Algorithmic uncertainty, if applicable:** None
- **Uncertainty in input data, if applicable:** None
- **Effect of parameter uncertainty, error propagation, if applicable:** None