From hot to cold spots: climate change will modify diversity patterns of small mammals in a biodiversity hotspot

– ODMAP Protocol –

Gabriela Alves-Ferreira, Mauricio Vancine, Flavio Mota, Carolina Bello, Thadeu Sobral-Souza, Alexandre R. Percequillo, Thomas E. Lacher Jr, Mauro Galetti, Ricardo S. Bovendorp

2024-09-13

## Overview

#### Authorship

Contact : [gabriela-alves77@hotmail.com](mailto:gabriela-alves77@hotmail.com) [mauricio.vancine@gmail.com](mailto:mauricio.vancine@gmail.com) [flaviomoc@gmail.com](mailto:flaviomoc@gmail.com)

Study link: In review

#### Model objective

Model objective: Forecast and transfer

Target output: Suitable vs. unsuitable habitat

#### Focal Taxon

Focal Taxon: In AF there are around 30 species of marsupials (Didelphidae) and 97 species of small rodents (Caviidae, Cricetidae, Ctenomyidae, Echimyidae) (Bovendorp et al., 2017b; Dalapicolla et al., 2021, Abreu et al., 2022). We focus on 101 species of terrestrial small mammals.

#### Location

Location: Our main study area is the AF in South America. However, we conduct the calibration and projection of the SDMs considering the entire extent of South America (Longitude -109.446 to -26.241, Latitude -58.498 to 12.590), since many species we are evaluating occur not only in the AF, but also in other biomes in South America.

#### Scale of Analysis

Spatial extent: -109.446, -26.241, -58.498, 12.590 (xmin, xmax, ymin, ymax)

Spatial resolution: Community boundaries were defined as grid cell of 5km x 5km.

Temporal extent: We used occurrence records collected between 1970 and 2024.

Temporal resolution: We used the present (1970-2000) and two future climate scenarios: 2050 and 2070. For each future scenario, we considered two Shared Socioeconomic Pathways (SSPs) representing optimistic (SSP370) and pessimistic (SSP585) greenhouse gas scenarios.

Boundary: natural

#### Biodiversity data

Observation type: We compiled occurrence records from the literature (Dalapicolla et al., 2021) and from the online platforms

Response data type: presence/absence, presence/background

#### Predictors

Predictor types: climatic

#### Hypotheses

Hypotheses: We used Species distribution models (SDMs) through an ensemble approach to assess the potential distribution of species of small mammals using baseline (1970-2000) and future (2050 and 2070) climate scenarios for two greenhouse gas concentration scenarios (SSP370 and SSP585). We evaluate if protected areas and forest remnants in the Atlantic Forest of South America (AF) will ensure the taxonomic diversity (TD) and phylogenetic diversity (PD) of non-volant small mammals under scenarios of future climate change.

#### Assumptions

Model assumptions: • Occurrence records are accurate and representative. • Species are at equilibrium with their environment. • The niche is preserved over time. • Models do not consider potential for adaptation to climate change.

#### Algorithms

Modelling techniques: glm, gam, gau, randomForest, suport vector machines, maxent, Neural Network Models, Generalized Boosted Regression Models

Model complexity: The models were fitted and evaluated using the R package *flexsdm* (Velazco et al., 2022) with the following algorithms, fitted without tuning (as tunning is not available in the *flexsdm* package for these algorithms): Generalized Linear Models (GLM, poly = 2), Generalized Additive Models (GAM, binomial distribution family and thin plate regression spline as a smoothing basis), and Gaussian Process Models (GPM),; and with hyperparameter tuning: Random Forest (RF; using *mtry* = (1, 2, 3, 4, 5, 6, 7), i.e., the number of variables to randomly sample as candidates at each split), Support Vector Machines (SVM; using C = (2, 4, 8, 16, 20), and sigma = (0.01, 0.1, 0.2, 0.3, 0.4)), MaxEnt (using regularization multipliers = (0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4), and feature classes = (l, lq, h, lqh, lqhp, lqhpt)), Neural Networks Models (NNM; using size = (2, 4, 6, 8, 10), and decay distance = (0.001, 0.05, 0.1, 1, 3, 4, 5, and 10)), and Generalized Boosted Regression Models (GBM; using trees = (20, 50, 100), shrinkage = (0.1, 0.5, 1), and n.minobsinnode = (1, 3, 5, 7, 9), i.e., the minimum number of observations in the terminal nodes of the trees). We created SDMs with hyperparameters tuning because default hyperparameter values ​​often do not return the best models (Fourcade, 2021; Morales et al., 2017; Vignali et al., 2020).

Model averaging: We addressed algorithmic uncertainty by using an ensemble method that averaged the results from distinct algorithms. The ensemble models were computed as the weighted averages of the climatic suitability across all the algorithms, using the TSS values for “Maximum Sensitive plus Specificity” threshold as weight.

#### Workflow

Model workflow: We used SDMs for species with more than 20 occurrences (82 species, 81%) and Ensemble of Small Models (ESM) for species with less than 20 and more than 10 occurrences (19 species, 18%) (Breiner et al., 2015). For calibration area of each species we randomly sampled pseudo-absence (same number of presences) and background points (number of presences multiplied by 10; Whitford et al., 2024) throughout the area used for model fitting. The models were fitted and evaluated using the R package flexsdm (Velazco et al., 2022) with the following algorithms, fitted without tuning (as tunning is not available in the flexsdm package for these algorithms): Generalized Linear Models (GLM), Generalized Additive Models (GAM), and Gaussian Process Models (GPM),; and with hyperparameter tuning: Random Forest (RF; using mtry = (1, 2, 3, 4, 5, 6, 7), i.e., the number of variables to randomly sample as candidates at each split), Support Vector Machines (SVM; using C = (2, 4, 8, 16, 20), and sigma = (0.01, 0.1, 0.2, 0.3, 0.4)), MaxEnt (using regularization multipliers = (0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4), and feature classes = (l, lq, h, lqh, lqhp, lqhpt)), Neural Networks Models (NNM; using size = (2, 4, 6, 8, 10), and decay distance = (0.001, 0.05, 0.1, 1, 3, 4, 5, and 10)), and Generalized Boosted Regression Models (GBM; using trees = (20, 50, 100), shrinkage = (0.1, 0.5, 1), and n.minobsinnode = (1, 3, 5, 7, 9), i.e., the minimum number of observations in the terminal nodes of the trees). We created SDMs with hyperparameters tuning because default hyperparameter values ​​often do not return the best models (Fourcade, 2021; Morales et al., 2017; Vignali et al., 2020). We evaluated and calibrated the models using spatial block cross validation with four partitions (Santini et al., 2021). The spatial block validation is considered a more robust approach for evaluating model transferability (Roberts et al., 2017). We partitioned the data using presence and pseudo-absence data using Spatial block cross-validation, which were used to partition the background data. First, we tried to partition to k = 4 whenever possible; if the function returned an error, we decreased it and tried with k = 3 and if the error persisted, we did it with k = 2. For ESM, we used a random partition, with k = 3 and 5 replicates due to few points of occurrences. We accessed the model performance by calculating the True Still Statistic (TSS) and selected models fitted with TSS > 0, since TSS values vary from -1 to 1, where negative and close-to-zero values indicate models that are not different from randomly generated models, while values close to 1 indicate good models (Allouche et al., 2006). We also reported other evaluation metrics to facilitate comparisons in the literature such as Boyce Index, Area Under Curve (AUC), Omission Rate (OR), FPB, and Sorensen Index. We carefully inspected the response curves for bimodal responses or responses that were considered implausible by small mammal experts. We found no curves that fit these terms, and therefore no response curves were removed. We addressed algorithmic uncertainty by using an ensemble method that averaged the results from distinct algorithms. The ensemble models were computed as the weighted averages of the climatic suitability across all the algorithms, using the TSS values for “Maximum Sensitive plus Specificity” threshold as weight. We projected the models for the present, 2050 and 2070 in an optimistic and pessimistic scenario. We converted the final continuous models (present and future) into presence-absence maps (presence = 1 and absence = 0) using the threshold that maximizes the sum of sensitivity and specificity (Max SSS) (Liu et al., 2013). We chose this approach because our interest in this article was to analyze the composition of potential ecological communities (Stacked Species Distribution Modeling - SSDM; Calabrese et al., 2014), to calculate TD and PD for each community.

#### Software

Software: R 4.3.2 and R Packages flexsdm (Velazco et al., 2020), R package CoordinateCleaner (Zizka et al., 2019), R package spThin (Aiello-Lammens et al., 2015).

Code availability: <https://github.com/gabferreira/analysis_small_mammals/tree/main/00_scripts>

Data availability: <https://github.com/gabferreira/analysis_small_mammals/tree/main/01_data>

## Data

#### Biodiversity data

Taxon names: We focused on 101 species of terrestrial small mammals. In the AF are recognized around 30 species of marsupials (Didelphidae) and 97 species of small rodents (Caviidae, Cricetidae, Ctenomyidae, Echimyidae) (Bovendorp et al., 2017b; Dallapicola et al., (2021); Abreu et al., 2022). The taxon names can be accessed in the supplementary table S1.

Taxonomic reference system: Current names in use are based on the Dallapicola et al., (2021). The taxonomic issues in species names were corrected by specialists in small mammals.

Ecological level: species, communities

Data sources: We compiled occurrence records from the literature (Dalapicolla et al., 2021) and from the online platforms: GBIF (www.gbif.org, 2024), speciesLink (www.specieslink.net, 2024), VertNet (<https://www.vertnet.org>), iDigBio (<https://www.idigbio.org>), Sistema da Informação sobre a Biodiversidade Brasileira (www.sibbr.gov.br, 2024), iDigBio, (www.idigbio.org/, 2024), and ICMBio (<https://portaldabiodiversidade.icmbio.gov.br/portal/>, 2024).

Sample size: The sample size per taxon ranges from 10 to 1,410 occurrences per species. Most species (76) have between 10 and 150 occurrence records. The number of occurrences per species can be seen in Table S1.

Clipping: Our main study area is the AF in South America. However, we conduct the calibration and projection of the SDMs considering the entire extent of South America (Longitude -109.446 to -26.241, Latitude -58.498 to 12.590), since many species we are evaluating occur not only in the AF, but also in other biomes in South America.

Scaling: We thinned occurrences to reduce spatial bias using 5 kilometers filtering through R package spThin (Aiello-Lammens et al., 2015). We also applied an environmental filter with the bioclimatic variables to eliminate occurrence records with the same value for the same environmental condition (using 12 classes) with the R package flexsdm (Velazco et al., 2022).

Cleaning: We performed a quality control on the occurrences to eliminate duplicates, unlikely or impossible records, georeferenced in centroids of countries, states, and cities, or georeferenced in institutions (i.e., universities, zoos), filter only occurrence in South America spatial limit, and located over the ocean through R package CoordinateCleaner (Zizka et al., 2019). We used occurrence records collected between 1970 and 2024. The taxonomic issues in species names were corrected by specialists in small mammals. We selected species with more than 10 occurrence records to conduct further analysis. Our final database comprises 12,166 occurrence records for 102 species of small mammals in the Atlantic Forest (Supplementary Table S1).

Absence data: We randomly sampled pseudo-absence (same number of presences) throughout the area used for model fitting.

Background data: We randomly sampled background points (number of presences multiplied by 10) throughout the area used for model fitting.

Errors and biases: We performed a quality control on the occurrences to eliminate duplicates, unlikely or impossible records, georeferenced in centroids of countries, states, and cities, or georeferenced in institutions (i.e., universities, zoos), filter only occurrence in South America spatial limit, and located over the ocean through R package CoordinateCleaner (Zizka et al., 2019). We used occurrence records collected between 1970 and 2024. The taxonomic issues in species names were corrected by specialists in small mammals.

#### Data partitioning

Validation data: We evaluated and calibrated the models using spatial block cross validation with four partitions (Santini et al., 2021). The spatial block validation is considered a more robust approach for evaluating model transferability (Roberts et al., 2017). We partitioned the data using presence and pseudo-absence data using Spatial block cross-validation, which were used to partition the background data. First, we tried to partition to k = 4 whenever possible; if the function returned an error, we decreased it and tried with k = 3 and if the error persisted, we did it with k = 2. For ESM, we used a random partition, with k = 3 and 5 replicates due to few points of occurrences.

Test data: There was no truly independent dataset available.

#### Predictor variables

Predictor variables: We adjusted the bioclimatic variables for calibration area for each species (i.e. buffer with ~300 km radius), and selected a set of variables with the VIF lower than 2.0 (Dormann et al., 2013). The same variables were used to project the estimated niche for future scenarios of climate change.

Data sources: URL: <https://chelsa-climate.org/bioclim/> Version 2.1 Accession date: 08/23/2024

Spatial extent: -109.446, -26.241, -58.498, 12.590 (xmin, xmax, ymin, ymax)

Spatial resolution: 5 km²

Coordinate reference system: The coordinate reference system is WGS84 (EPSG:4326).

Temporal extent: We obtained bioclimatic variables from the CHELSA v.2.1 for the present (1970-2000).

Dimension reduction: We followed a more restrictive recommendation for selecting variables (VIF < 2) for the environment where each species occurs (300 km buffer). It would be very costly to consider the individual biological response of our more than 100 species of small mammals. However, the Atlantic Forest environment is highly diverse in terms of climatic conditions, as we have a biome with different ecoregions and types of vegetation (Joly et al., 2014), largely due to its large latitudinal extension (which we mentioned in the methods), but also altitudinal extension, with strong effects of maritime activity due to its proximity to the Atlantic Ocean. Thus, it is expected that species respond to the bioclimatic patterns captured by the bioclimatic variables.

#### Transfer data

Data sources: URL: <https://chelsa-climate.org/bioclim/> Version 2.1 Accession date: 08/23/2024

Spatial resolution: 5 km² and same spatial extent of the raw data

Temporal extent: Two future climate scenarios: 2050 and 2070.

Models and scenarios: We considered two Shared Socioeconomic Pathways (SSPs) representing optimistic (SSP370) and pessimistic (SSP585) greenhouse gas scenarios. We used all Global Circulation Models (GCM) available in CHELSA: GFDL-ESM4, MPI-ESM1-2-HR, MRI-ESM2-0, IPSL-CM6A-LR, and UKESM1-0-LL (Collins et al., 2011; Navarro-Racines et al., 2020; Canon et al., 2020), to minimize the uncertainty about the choice of just one GCM (Thuiller et al., 2019).

## Model

#### Variable pre-selection

Variable pre-selection: We selected a set of bioclimatic variables for each species with the Variance Inflation Factor (VIF) lower than 2.0 (Dormann et al., 2013).

#### Multicollinearity

Multicollinearity: We used the Variance Inflation Factor (VIF) less than 2.0 (Dormann et al., 2013) to choose the least correlated variables.

#### Model settings

GLM: poly = 2

GAM: binomial distribution

randomForest: mtry (1, 2, 3, 4, 5, 6, 7)

suport vector machines: C (2, 4, 8, 16, 20), sigma (0.01, 0.1, 0.2, 0.3, 0.4)

maxent: featureRule (“l”, “lq”, “h”, “lqh”, “lqhp”, “lqhpt”), regularizationMultiplierSet (0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5)

Neural Network Models: size (2, 4, 6, 8, 10), decay (0.001, 0.05, 0.1, 1, 3, 4, 5, 10)

Generalized Boosted Regression Models: n.trees (20, 50, 100), shrinkage (0.1, 0.5, 1), n.minobsinnode (1, 3, 5, 7, 9)

Model settings (extrapolation): The extrapolation was limited, as we applied a posteriori method on final outputs which only retains those pixels in suitability patches intercepting known occurrences records (Mendes et al., 2020).

#### Model estimates

<Coefficients>

Variable importance:

#### Model selection - model averaging - ensembles

Model selection: We accessed the model performance by calculating the True Still Statistic (TSS) and selected models fitted with TSS > 0, since TSS values vary from -1 to 1, where negative and close-to-zero values indicate models that are not different from randomly generated models, while values close to 1 indicate good models (Allouche et al., 2006). We also reported other evaluation metrics to facilitate comparisons in the literature such as Boyce Index, Area Under Curve (AUC), Omission Rate (OR), FPB, and Sorensen Index.

Model averaging: We addressed algorithmic uncertainty by using an ensemble method that averaged the results from eight distinct algorithms. The ensemble models were computed as the weighted averages of the climatic suitability across all the algorithms, using the TSS values as weight.

Model ensembles: We addressed algorithmic uncertainty by using an ensemble method that averaged the results from eight distinct algorithms. The ensemble models were computed as the weighted averages of the climatic suitability across all the algorithms, using the TSS values as weight.

#### Analysis and Correction of non-independence

Spatial autocorrelation: We used spatial block cross validation to address spatial autocorrelation between calibration and validation occurrence records.

#### Threshold selection

Threshold selection: We transformed the final continuous models (present and future) into presence-absence maps (presence = 1 and absence = 0) using the threshold that maximizes the sum of sensitivity and specificity (Max SSS) (Liu et al., 2013).

## Assessment

#### Performance statistics

Performance on training data: TSS, AUC, Sorensen Index, Boyce Index, FPB

Performance on validation data: TSS, AUC, Boyce Index, FPB, Sorensen index

Performance on test data: There was no truly independent dataset available.

#### Plausibility check

Response shapes: We carefully inspected the response curves for bimodal responses or responses that were considered implausible by small mammal experts. We found no curves that fit these terms, and therefore no response curves were removed.

Expert judgement: The presence-absence and the continous maps were carefully checked by specialists in terrestrial small mammals.

## Prediction

#### Prediction output

Prediction unit: The models produced continuous suitability maps and presence-absence maps.

Post-processing: We clipped the models by the extent of the Atlantic Forest.

#### Uncertainty quantification

Algorithmic uncertainty: We addressed algorithmic uncertainty by using an ensemble method that averaged the results from eight distinct algorithms. We computed the ensemble models as the weighted averages of the climatic suitability across all the algorithms, using the TSS values as weight (Leroy et al., 2018).

Scenario uncertainty: We used all Global Circulation Models (GCM) available in CHELSA: GFDL-ESM4, MPI-ESM1-2-HR, MRI-ESM2-0, IPSL-CM6A-LR, and UKESM1-0-LL (Collins et al., 2011; Navarro-Racines et al., 2020; Canon et al., 2020), to minimize the uncertainty about the choice of just one GCM (Thuiller et al., 2019). We calculated the mean of all GCMs to use in the SDMs.

Novel environments: We applied spatial constraints based on a posteriori methods to reduce overprediction in presence-absence maps. We used the method “pres” (only occurrences based restriction), which only retains those pixels in suitability patches intercepting known occurrences records to remove the overprediction (Mendes et al., 2020).