

EcoLab: A Tool for Modeling and Analyzing Potential Species Distribution

Ane C. Simões¹, Aline C. Martins², Leyza B. Dorini¹, Luiz Gomes-Jr¹

¹PPGCA - UTFPR – Curitiba – Brazil

²Coordenação de Biodiversidade - INPA – Manaus – Brazil

Abstract. *Species distribution modeling (SDM) is widely used in ecological research; however, its traditional implementation in R still presents limitations, such as the difficulty of selecting the appropriate packages and the absence of biotic variables (data on species that interact with the target species). Given this context, this work presents the development of an application that automates and integrates all stages of building an SDM, encompassing steps from obtaining occurrence data to extracting environmental variables, modeling, and evaluating results.*

1. Introduction

The current global decline in biodiversity has raised significant concern among researchers, especially in the biological sciences [Sousa-Silva et al. 2014]. Human activity is considered the main driver of this reduction, promoting climate change, habitat alteration, and increasing the risk of extinction for numerous animal and plant species [Lima and Marchioro 2021].

To mitigate these impacts, technology has played a fundamental role in the development of tools that improve the understanding of species distribution, supporting conservation planning and the assessment of environmental changes [Dogbo et al. 2025]. In this context, *Species Distribution Modeling (SDM)* combines occurrence data, biotic and abiotic variables, and machine learning algorithms to generate potential distribution maps. These maps help researchers identify priority areas and define conservation strategies.

Although *SDM* is widely used in biological research, its implementation is often complex and relies heavily on R¹. Researchers usually need to integrate multiple packages, such as *rgbif* for occurrence data, *bioclim* for climate data, and *Biomod2* for model execution and evaluation. This fragmentation requires technical expertise, may limit study replicability, and restrict access for researchers unfamiliar with programming.

According to [Kass et al. 2025], the large number of libraries available in R makes it difficult to select the most appropriate tools, often resulting in redundant packages with similar functionalities. The authors also highlight the importance of incorporating biotic variables to improve model accuracy, a feature that is still absent from most *SDM* packages. These limitations reveal a gap in the lack of an integrated, programming-free tool

*Video demo: https://youtu.be/hPBY4_qipXE

GitHub Repo: <https://github.com/EcoToolBox/EcoLab>

¹R is a free and open-source software environment for statistical computing, data analysis, and graphical visualization. Additional information is available at <https://www.r-project.org>.

that combines occurrence data, environmental variables, and biotic interactions in a single workflow.

Given this scenario, the present study proposes the development of an application capable of integrating and automating the steps required to build a *SDM*. The tool includes the acquisition of occurrence data from: (i) *Global Biodiversity Information Facility (GBIF)*; (ii) *iNaturalist*; (iii) *SpeciesLink*. The system also extracts environmental variables, including climate data from *WorldClim* and ecological data derived from *Sentinel-2* imagery, and implements machine learning models to generate potential distribution maps.

Although several tools can apply *SDM* techniques, the proposed system distinguishes itself by integrating all stages of the process while eliminating the need for R programming knowledge, which can be a barrier for many researchers. Furthermore, the system allows the integration of occurrence data from interacting species to improve model performance. These data are obtained from the *Global Biotic Interactions* platform, which provides records of more than 40 ecological interaction types, including predation, pollination, and mutualism. Based on this information, it is possible to identify species that interact with the target organism and use their occurrences as additional variables. This integration addresses the gap highlighted by [Kass et al. 2025] by increasing model accuracy and making predictions more consistent with ecological conditions.

For model evaluation, the system provides the metrics *Area Under the Curve (AUC)*, *True Skill Statistic (TSS)*, and the *Boyce Index*, which are widely used in the literature. The project also aims to contribute to reducing biodiversity knowledge gaps regarding the analyzed species. Through the generated models, it becomes possible to examine potential distribution maps, expanding the understanding of species geographic occurrence and the factors influencing their distribution.

2. Methodology

2.1. System Architecture

EcoLab is implemented as a desktop application, its backend is developed using *FastAPI* and *Uvicorn*, responsible for orchestrating data acquisition, processing, and modeling tasks. The frontend is built with *React*, providing an accessible graphical interface that allows researchers to interact with the system without programming knowledge. The application is distributed as a standalone executable using *PyInstaller* and is available for Windows, Mac, and Linux.

Figure 1 illustrates the system architecture and the data sources used throughout the modeling process. From the user's perspective, the system requires the selection of the target species, environmental variables, biotic interaction types, and modeling algorithms. Users may also provide their own occurrence and environmental datasets in a structured format. The output consists of potential distribution maps and evaluation metrics generated through the integrated workflow.

2.2. Occurrence Data Acquisition

The first step of an *SDM* is obtaining occurrence records for the target species. To retrieve and unify data from different public biodiversity repositories, we developed *EcoObs*², a

²*EcoObs*: <https://pypi.org/project/ecobs/>

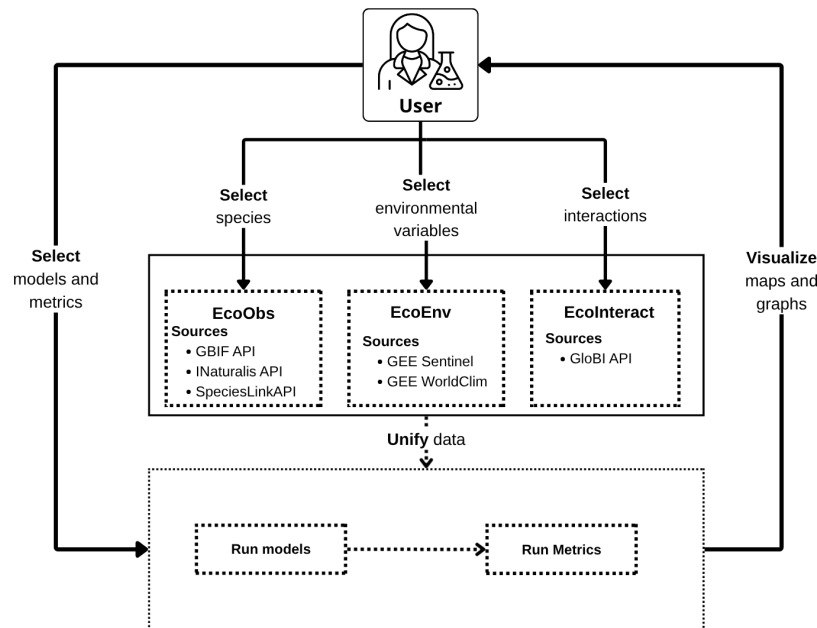


Figure 1. Architecture diagram of the proposed application

Python library that connects to multiple *APIs* and collects occurrence data in a standardized format.

After collection, the records undergo a curation process that includes duplicate removal, exclusion of entries without valid geographic coordinates, and verification of possible taxonomic inconsistencies.

This work adopts presence and pseudo-absence data. Pseudo-absences are generated using the *background* approach, where points not classified as presences are considered absences.

2.3. Environmental and Biotic Variables

The environmental variables used in this work include both environmental and ecological factors associated with species distribution.

To automate the acquisition of abiotic variables, we developed *EcoEnv*³, a library responsible for retrieving environmental data from *Google Earth Engine*. The library uses *Sentinel-2* imagery to extract *NDVI* and *NDWI*, while *WorldClim* data are used to obtain average precipitation and temperature values.

To incorporate ecological interactions into the modeling process, we developed *EcoInteract*⁴, which uses the *Global Biotic Interactions* (GloBI) platform as a source of interaction records. After identifying interacting species, the system uses *EcoObs* to retrieve occurrence records for these organisms, allowing their occurrences to be included as additional variables in the models.

³*EcoEnv*: <https://pypi.org/project/ecoenvironment/>

⁴*EcoInteract*: <https://pypi.org/project/ecointeract/>

2.4. Modeling

After data consolidation, the dataset is balanced between presences and pseudo-absences to reduce training bias and avoid *overfitting*. The data are then divided into training and testing subsets for model calibration and evaluation.

The modeling process employs five machine learning algorithms commonly used in *SDM*: *Maximum Entropy* (MaxEnt), *Random Forest* (RF), *Generalized Additive Model* (GAM), *Boosted Regression Trees* (BRT), and *Support Vector Machine* (SVM). These algorithms generate potential distribution maps indicating areas of higher habitat suitability. The generated models are evaluated using *AUC*, *TSS*, and the *Boyce Index*, enabling comparison among different approaches.

3. Use Case

For the initial system evaluation, experiments were conducted using the species *Tetragonisca angustula*, popularly known as the Jataí bee. The decision to work with native bee species was motivated by the expertise of the biology researcher associated with the project. Figure 2 presents the parameters selected for the occurrence search. All available observation sources were selected, together with the largest available time range, while the search was restricted to the Brazilian territory.

The screenshot displays the EcoLab web application interface. On the left, a sidebar shows 'Selecionadas' with 'Tetragonisca angustula' selected. The main area, titled 'Filtros de Busca', contains a section 'Selecione as fontes de dados' where GBIF, iNaturalist, and SpeciesLink are all checked. Below this, the 'Intervalo de anos' is configured with a start date of 1800 and an end date of 2026, accompanied by a horizontal timeline slider. At the bottom, the 'PAIS' dropdown menu is set to 'Brasil'.

Figure 2. EcoLab occurrence search parameters configured for *Tetragonisca Angustula*: all available observation sources (GBIF, iNaturalist and SpeciesLink) selected across the full time range, restricted to Brazilian territory.

Next, all available environmental variables were selected. For the biotic interactions, the relations *eats*, *preysOn*, and *kills* were included. For the purposes of this experiment, all available modeling algorithms and evaluation metrics were selected.

After the data cleaning and curation process, approximately 6.5 thousand occurrence records were obtained by combining the selected data sources (Figure 3).

After this stage, the environmental and biotic variables were retrieved, and pseudo-absence points were generated across the Brazilian territory. The selected machine learning models were then trained to generate potential species distribution maps.

Ocorrências Encontradas

6580 registros encontrados para 1 espécie(s).

scientificName	country	latitude	longitude	day	month	year	source	eventDate
Tetragonisca angus...	BR	-12.9711	-38.5108	1	1	1811	gbif	1811-01-01
Tetragonisca angus...	BR	-13.1822	-39.425	1	1	1811	gbif	1811-01-01
Tetragonisca angus...	BR	-12.67	-39.1019	1	1	1811	gbif	1811-01-01
Tetragonisca angus...	BR	-14.0694	-42.475	1	1	1811	gbif	1811-01-01
Tetragonisca angus...	BR	-14.9261	-42.8192	1	1	1811	gbif	1811-01-01
Tetragonisca angus...	BR	-14.8497	-42.4333	1	1	1811	gbif	1811-01-01
Tetragonisca angus...	BR	-12.5631	-41.39	1	1	1811	gbif	1811-01-01
Tetragonisca angus...	BR	-13.1144	-40.7009	1	1	1811	gbif	1811-01-01

Rows per page: 100 1-100 of 6580

Figure 3. Occurrence records retrieved for *Tetragonisca Angustula* using EcoObs. Approximately 6,500 records are displayed across Brazilian territory after data cleaning and curation.

Figures 4 present the habitat suitability maps generated by EcoLab. According to the evaluation indices (AUC, TSS, and Boyce Index), the Random Forest model showed better overall performance than MaxEnt. In the maps, MaxEnt predicted suitable habitats over a larger area of Brazil, resulting in a more generalized distribution. In contrast, Random Forest produced a more restricted suitability map, with predictions concentrated closer to the known occurrence records.

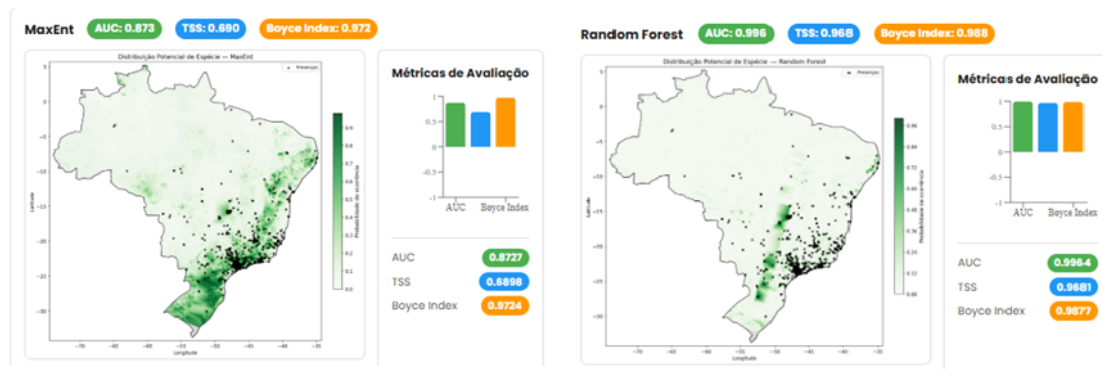


Figure 4. Habitat suitability maps for *Tetragonisca Angustula* generated by EcoLab using MaxEnt and Random Forest. Darker colors indicate higher predicted suitability across Brazilian territory.

4. Conclusion

The system was developed to support ecology researchers in performing *Species Distribution Modeling (SDM)* analyses by integrating species occurrence records, environmental variables, and model generation into a single platform, reducing the need for manual database extraction and programming knowledge. By simplifying these processes, the proposed tool aims to make ecological modeling more accessible to researchers with different levels of technical expertise.

The main expectation for the system is to provide a practical and extensible environment capable of assisting researchers in biodiversity analysis, species conservation

studies, and ecological decision-making. In addition to automating data acquisition and model generation, the platform is expected to facilitate the exploration of environmental variables, species interactions, and distribution patterns in a more intuitive and reproducible way.

The tool was also evaluated by researchers from the biology field, who recognized its applicability and potential contribution to ecological studies. The feedback obtained during this evaluation revealed important opportunities for improvement, including additional export functionalities, support for user-provided datasets, refinements in occurrence-data filtering, and the incorporation of new interaction databases. These suggestions provide valuable directions for future development and reinforce the potential of the system as a collaborative platform for ecological and biodiversity research.

This work also highlights the relevance of database integration, data quality, and user-oriented analytical systems applied to biodiversity research. The proposed system integrates data sources and combines multiple modeling approaches, including structured environmental data, geospatial information, and graph-based interaction data, into a unified platform for ecological analysis. By automating data processing workflows and providing an accessible interface for complex analytical tasks, the system demonstrates how database technologies, data integration, and hybrid modeling approaches can support real-world scientific challenges in biodiversity and environmental research.

References

- Dogbo, S. F., Salako, K. V., Agoundé, G., Dimobe, K., Adiko, A. E. G., Gebauer, J., Yao, C. Y. A., and Glèlè Kakaï, R. (2025). Potential impacts of future climate on twelve key multipurpose tree species in benin: Insights from species distribution modeling for biodiversity conservation. *Trees, Forests and People*, 19:100744.
- Kass, J. M., Smith, A. B., Warren, D. L., Vignali, S., Schmitt, S., Aiello-Lammens, M. E., Arlé, E., Márcia Barbosa, A., Broennimann, O., Cobos, M. E., Guéguen, M., Guisan, A., Merow, C., Naimi, B., Nobis, M. P., Ondo, I., Osorio-Olvera, L., Owens, H. L., Pinilla-Buitrago, G. E., Sánchez-Tapia, A., Thuiller, W., Valavi, R., Velazco, S. J. E., Zizka, A., and Zurell, D. (2025). Achieving higher standards in species distribution modeling by leveraging the diversity of available software. *Ecography*, 2025(2):e07346.
- Lima, V. P. and Marchioro, C. A. (2021). Brazilian stingless bees are threatened by habitat conversion and climate change. *Regional Environmental Change*, 21(1):14.
- Sousa-Silva, R., Alves, P., Honrado, J., and Lomba, A. (2014). Improving the assessment and reporting on rare and endangered species through species distribution models. *Global Ecology and Conservation*, 2:226–237.

AI Usage

For the sake of transparency, *Generative Artificial Intelligence (AI)* tools were used during both the manuscript preparation and application development stages, assisting with English translation, grammatical and typographical revision, resolution of specific implementation issues, and repetitive programming tasks.