

## APPLICATION

# GLOSSA: A user-friendly R Shiny application for Bayesian machine learning analysis of marine species distribution

Jorge Mestre-Tomás<sup>1,2</sup>  | Alba Fuster-Alonso<sup>1,3</sup>  | José M. Bellido<sup>4</sup>  | Marta Coll<sup>1,5</sup> 

<sup>1</sup>Renewable Marine Resources Department, Institut de Ciències del Mar (ICM), CSIC, Barcelona, Spain

<sup>2</sup>Department of Applied Statistics and Operational Research, and Quality, Universitat Politècnica de València (UPV), Valencia, Spain

<sup>3</sup>Department of Statistics and Operations Research (VaBar), Universitat de València (UV), Valencia, Spain

<sup>4</sup>Centro Oceanográfico de Murcia, Instituto Español de Oceanografía (COMU-IEO), CSIC, San Pedro del Pinatar, Spain

<sup>5</sup>Ecopath International Initiative (EII), Barcelona, Spain

## Correspondence

Jorge Mestre-Tomás

Email: [jorgemestre@icm.csic.es](mailto:jorgemestre@icm.csic.es)

## Funding information

Ministerio de Ciencia e Innovación, Grant/Award Number: PID2020-118097RB-I00 and PID2021-124831OA-I00; Horizon Europe, Grant/Award Number: 101059877; Agència de Gestió d'Ajuts Universitaris i de Recerca, Grant/Award Number: 2021 SGR 00435

Handling Editor: Aldo Compagnoni

## Abstract

1. Species distribution models (SDMs) are one of the most common statistical methods to assess species occupancy and geographic distribution patterns. With the increasing complexity and availability of ecological data in the marine context, many methodological approaches have been developed to support SDM analysis. However, their application often requires expertise in data analysis, statistical modelling and programming, which limits their accessibility for broader use.
2. Here we introduce GLOSSA, an open-source R package and Shiny application designed to make marine species distribution modelling more accessible. GLOSSA provides a user-friendly interface for fitting Bayesian Additive Regression Trees (BART) SDMs using species occurrence and environmental data.
3. GLOSSA guides users through key SDM steps, including data uploading, filtering occurrence data, harmonizing environmental layers, generating pseudo-absences, tuning BART complexity, performing spatial and temporal block cross-validation, visualizing predictions and uncertainty and exporting configuration files to ensure reproducibility.
4. We demonstrate the functionality of GLOSSA through three marine case studies, addressing a range of ecological scenarios at regional and global scales. Along with detailed documentation, examples and tutorials, GLOSSA provides an example of how an intuitive graphical interface can make species distribution modelling accessible to a broad audience.

## KEYWORDS

Bayesian Additive Regression Trees, biogeography, habitat suitability model, probability of occurrence, R Shiny, software, species distribution model

## 1 | INTRODUCTION

Understanding the geographic distribution of species over time and the effects of environmental conditions on them is crucial for ecological research, biodiversity conservation and resource

management (Franklin, 2013; Guisan et al., 2013; Pulliam, 2000; Srivastava et al., 2019). Species distribution models (SDMs), ecological niche models (ENMs) and habitat suitability models (HSMs) are widely used tools to estimate the relationship between species occurrence and environmental conditions and predict their

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society.

distributions and potential suitable habitats. These models have a wide range of applications, including evaluating climate change impact on species distributions (Booth, 2018; Brown et al., 2016; Melo-Merino et al., 2020), assessing invasive species risk (Fournier et al., 2019; Lyons et al., 2020), understanding ecological interactions (Ge et al., 2024; Hill et al., 2024; Liu et al., 2023) and planning marine protected areas (Ferrari et al., 2018; Hunt et al., 2020).

The increasing availability of georeferenced species and environmental data, together with advances in computational power, has led to an increase in the use of SDMs (GBIF.org, 2020; Tyberghein et al., 2012). However, these models often require expertise in data analysis and statistical modelling to address challenges such as data quality, sampling bias, autocorrelation and proper model validation (Kass et al., 2025), especially in marine contexts where datasets are often heterogeneous, data coverage is sparse, and spatio-temporal biases are common (Klaassen et al., 2025; Melo-Merino et al., 2020; Schickele et al., 2025). While custom scripts for SDM fitting offer flexibility, they are often poorly documented, designed for specific tasks and require programming skills. These challenges highlight the need for user-friendly tools to lower technical barriers (Wallace et al., 2012) in marine ecological research.

To meet the growing demand for easy-to-use methods, several tools now provide graphical user interfaces (GUIs) that simplify model use—see, for example, Osorio-Olvera et al. (2020), Kass et al. (2023), Coro et al. (2025) or Figueira et al. (2024). Despite these advances, ongoing work aims to improve documentation, reproducibility and transparency of workflows (Mislán et al., 2016), while also achieving higher standards in SDM practices, such as better sampling bias correction, quantification of model uncertainty and expanding the range of supported methods (Araújo et al., 2019; Kass et al., 2025).

Here, we present GLOSSA (Global Ocean Species Spatio-temporal Analysis), an open-source R Shiny application designed to make species distribution modelling using occurrence data and environmental predictors more accessible, particularly in marine environments. GLOSSA implements an intuitive graphical interface that guides users through a streamlined SDM workflow from data upload and preprocessing to model fitting, evaluation, prediction, visualization and export. The current modelling approach is based on Bayesian Additive Regression Trees (BART; Chipman et al., 2010), a flexible, non-parametric method useful for capturing complex ecological relationships (Fuster-Alonso et al., 2024; Thompson et al., 2023). GLOSSA supports multiple pseudo-absence generation strategies, spatial and temporal cross-validation, functional response curves and variable importance plots, and uncertainty quantification through summaries of posterior distributions (point estimates and credible intervals). Interactive visualizations help explore predictions and diagnostics, and all results can be exported with a configuration file to ensure reproducibility. The entire workflow is accessible through a user-friendly interface that requires no coding expertise.

The following sections describe GLOSSA's workflow and implementation, illustrate its application through case studies, and discuss its current capabilities, limitations and future development.

## 2 | OVERVIEW OF GLOSSA

GLOSSA is a GUI-based application for species distribution modelling that currently uses correlative models based on BART to estimate species response to the environment by using georeferenced species occurrences and environmental data. GLOSSA's analysis workflow is organized into four main sequential steps (Figure 1), supported by an intuitive user interface with a simple learning curve: *Data upload*, *Data processing*, *Model fitting, evaluation and prediction*, and *Visualization and export*. Users interact directly with the shiny app during the upload and visualization steps, while data processing and modelling are handled automatically. Although some preprocessing steps may still need to be performed externally, GLOSSA integrates the core elements of SDM fitting and validation into an accessible interface designed for usability and reproducibility. Below, we describe each step in greater detail.

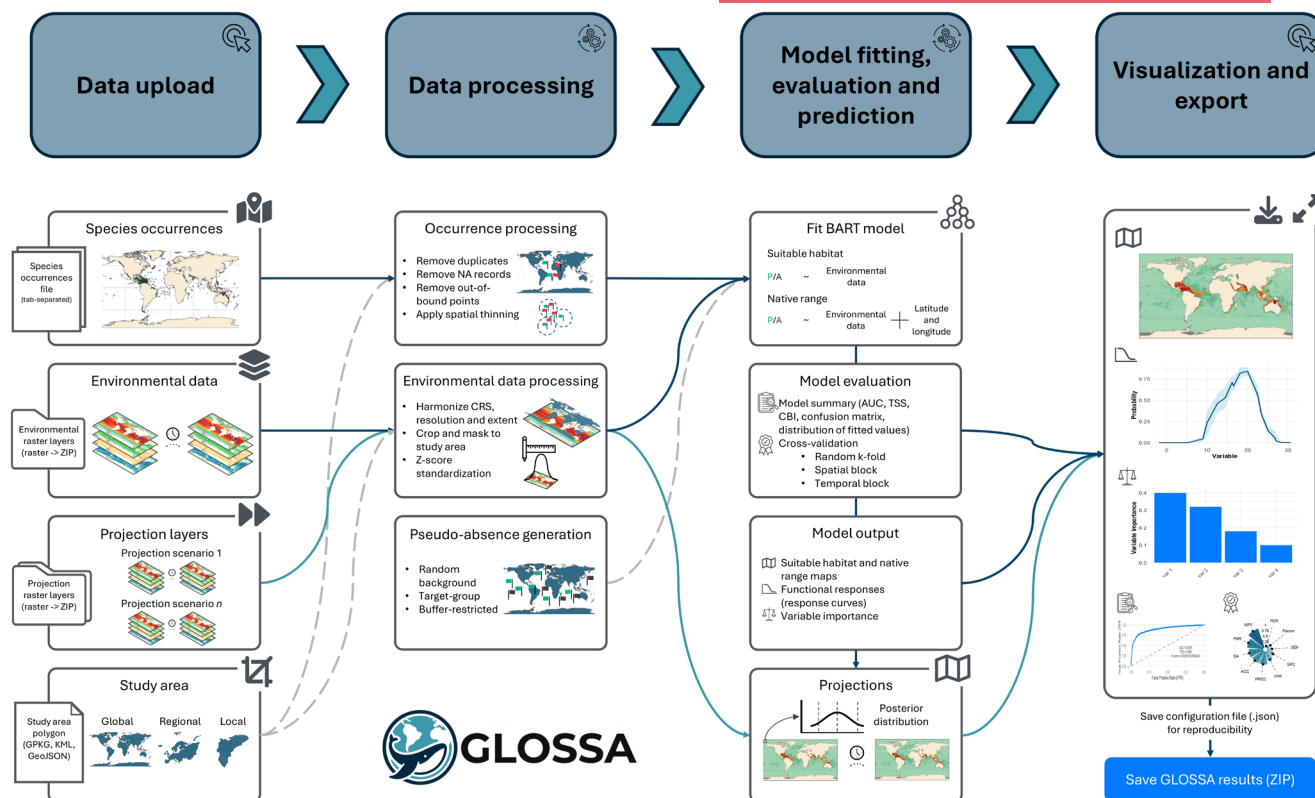
### 2.1 | Data upload and processing

#### 2.1.1 | Species occurrences

GLOSSA supports both presence/(pseudo-)absence data and presence-only data, for which pseudo-absences will be generated. Input files must contain longitude and latitude in the WGS84 reference system, and a column indicating presence or absence. A timestamp is optional but required when using time-varying predictors. Occurrence records may come from different sources such as survey programs, research studies, fisheries logbooks, satellite tracking or public data repositories such as the Global Biodiversity Information Facility (GBIF, <https://www.gbif.org/>) and the Ocean Biogeographic Information System (OBIS, <https://obis.org/>) (Melo-Merino et al., 2020). GLOSSA automatically removes duplicates, filters points outside the study area and can apply thinning, but any additional preprocessing steps not implemented in GLOSSA should be performed before using the application. Users can analyse multiple species independently within the same session.

#### 2.1.2 | Environmental data

Environmental predictor variables should be provided as raster layers. Predictor values are extracted at each occurrence point matching the timestamps if multiple temporal layers are provided. Only areas with complete environmental data are used, removing occurrences with missing predictor values. Input data should be in



**FIGURE 1** GLOSSA workflow overview. The main steps are indicated at the top and box labels match section headings in the text for clarity.

the WGS84 coordinate reference system, and if raster layers differ in resolution or extent, GLOSSA harmonizes them by aggregating to the coarsest resolution and adjusting spatial extents. Predictors may come from global databases such as the National Oceanic and Atmospheric Administration (NOAA, <https://www.noaa.gov/>), the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP, <https://www.isimip.org/>) or Bio-ORACLE (<https://www.bio-oracle.org/>) or represent custom variables such as fishing effort (e.g. Global Fishing Watch, <https://globalfishingwatch.org/>) or covariates representing sampling bias.

### 2.1.3 | Study area

Users can optionally provide a polygon to define the extent of the study area. This polygon will be used to filter species occurrences and to crop and mask environmental layers. If the polygon is not provided, the analysis is performed where environmental data are available.

### 2.1.4 | Projection layers

To support forecasting and scenario analysis, users can project the fitted model to other areas or time periods. Users can provide

environmental datasets for different time periods or climate scenarios, and GLOSSA projects the model to each and computes summary metrics across scenarios to facilitate comparisons of predicted distributions under varying environmental conditions and future climate scenarios.

### 2.1.5 | Occurrence processing

GLOSSA automatically removes occurrence records with missing or duplicated coordinates and those outside the study area or with missing predictor values. Occurrence datasets with heterogeneous sampling efforts or compiled from different sources often show sampling bias. To help mitigate this, GLOSSA includes spatial thinning (filtering) using a distance threshold or a regular grid (Mestre-Tomás, 2025) and also adjusting pseudo-absences, such as target-group background (Phillips et al., 2009). However, as marine occurrence data is prone to sampling bias and errors, additional preprocessing may be required outside of GLOSSA, using tools such as the *CoordinateCleaner* package (Zizka et al., 2019) or GUI-based platforms like OCCUR (Ronquillo et al., 2024) for more detailed cleaning of occurrence records or to apply alternative bias correction methods like environmental thinning (Moudry et al., 2024).

## 2.1.6 | Environmental data processing

Although users are encouraged to harmonize raster layers before upload, GLOSSA automatically aligns projections, resolutions and extents internally. In addition, optional Z-score standardization can be applied to continuous predictors to have a zero mean and unit variance, which may help to bring predictor variables to a similar scale.

## 2.1.7 | Pseudo-absence generation

Pseudo-absences are not true absences, generally representing the available environment in the study area, and their sampling strategy, number and distribution can affect model performance (Barbet-Massin et al., 2012; Phillips et al., 2009; Whitford et al., 2024). When true absences are not provided, GLOSSA supports four pseudo-absence generation methods: (i) random background sampling, (ii) target-group background, (iii) buffer-restricted sampling, which samples from a convex hull of the presences, excluding areas within a user-defined distance from presences and (iv) sampling based on the environmental space constrained to lower suitability regions using the *flexsdm* package (Velazco et al., 2022). Users can define the ratio between presences and pseudo-absences. By default, GLOSSA generates balanced pseudo-absences (1:1 ratio), which has been shown to perform well in many classification machine learning algorithms (Barbet-Massin et al., 2012). However, pseudo-absences must be created with caution as they can influence model outcomes. Some considerations include the generation method (Broussin et al., 2024; Iturbide et al., 2015), location of points (VanDerWal et al., 2009), the number of pseudo-absences (Barbet-Massin et al., 2012), the modelling technique used (Barbet-Massin et al., 2012) and the sources of biases in the data (Hertzog et al., 2014; Phillips et al., 2009), among others. For more customization, users may also upload externally generated pseudo-absences created using alternative strategies (Barbet-Massin et al., 2012), although the use of real absences when available is preferred (Sillero et al., 2021).

## 2.2 | Modelling approach

### 2.2.1 | Fit BART model

GLOSSA implements species distribution modelling using BART (Chipman et al., 2010), which leverage the Bayesian framework for model fitting regularization and uncertainty estimation. Although relatively new in SDMs, BART has shown to produce promising results (Baquero et al., 2021; Carlson, 2020; Konowalik & Nosol, 2021), including when using pseudo-absences Konowalik and Nosol (2021); Fuster-Alonso et al. (2024), and also in the marine context at regional

and global scales (Fuster-Alonso et al., 2024; Rubbens et al., 2023; Thompson et al., 2023).

Chipman et al. (2010) introduced BART as a non-parametric approach to regression modelling based on a sum-of-trees model where priors are used to regularize inference and avoid overfitting. By using regularization priors, BART ensures each tree acts as a 'weak learner', contributing only a small part of the overall model (Chipman et al., 2010). This makes BART particularly suitable for species distribution modelling, which often involves complex, high-dimensional ecological data.

Since GLOSSA is based on presence and absence data, we use a logistic regression with a Bernoulli distribution to model the probability of presence. Therefore, the model applied is as follows:

$$Y_i \sim \text{Ber}(\pi_i), \quad i = 1, \dots, n,$$

$$\phi^{-1}(\pi_i) = \sum_j^m g_j(\mathbf{X}; T_j, M_j), \quad (1)$$

where  $Y_i$  represents the presence/(pseudo-)absence of species for observation  $i$ , and  $\pi_i$  is the parameter of interest representing the probability of presence linked to the predictor by a link function  $\phi^{-1}$  that denotes the standard normal cdf (probit link function). Each  $g_j$  corresponds to the  $j$ -th ( $j = 1, \dots, m$ ) tree of the form  $g_j(\mathbf{X}; T_j, M_j)$ , where  $m$  is the total number of trees,  $\mathbf{X}$  is a vector of multiple covariates,  $T_j$  is a binary tree structure consisting of a set of interior node decision rules and a set of terminal nodes, and  $M_j = \{\mu_j, \dots, \mu_{jb}\}$  represents a set of parameter values associated with each of the  $b_j$  terminal nodes of  $T_j$ . GLOSSA currently uses the default priors recommended by (Chipman et al., 2010) to regularize the fit. Finally, an MCMC algorithm is used to sample from the posterior distribution.

GLOSSA allows fitting two types of models based on the predictor structure. The 'suitable habitat' model is based only on environmental predictors and key drivers (e.g. temperature, salinity) to estimate where conditions are appropriate for the target species, even if the species may not currently be present there for other reasons. The 'native range' model also includes spatial coordinates (latitude and longitude), capturing the observed distribution with spatial smoothing.

Although BART effectively reduces overfitting through an ensemble of decision trees with regularization prior distributions (Carlson, 2020; Chipman et al., 2010), sparse occurrence data or sampling bias can still lead to overfitting. To mitigate this, GLOSSA allows users to choose model parameters like the number of trees ( $m$ ) which controls representational flexibility and the amount of shrinkage in the node parameters by tuning the regularization parameter  $k$ . The prior for node values follows a  $\mathcal{N}(0, \sigma_\mu^2)$ , such that higher values of  $k$  produce more conservative fits (Chipman et al., 2010; Dorie, 2024; see File S1). These values may be chosen by cross-validation (Chipman et al., 2010; Dorie, 2024).

## 2.3 | Model evaluation

### 2.3.1 | Model evaluation

After model fitting, GLOSSA determines the presence/absence classification threshold using the True Skill Statistic (TSS or Youden's Index, Allouche et al., 2006; Youden, 1950). The model summary includes a ROC curve with area under the curve (AUC), TSS, F1-score and diagnostic plots with the confusion matrix and the distribution of fitted values. Because these metrics can be misleading when using pseudo-absences instead of true absences (Konowalik & Nosol, 2021), GLOSSA also computes the Continuous Boyce Index (CBI), an evaluation metric for models fitted using presence and background data (Boyce et al., 2002; Hirzel et al., 2006).

Model predictive performance can be assessed using cross-validation, with a flexible number of folds and data partitioning. Random k-fold, spatial block and temporal split cross-validation are supported, where the spatial block size can be defined manually or estimated from the predictor or residual range of spatial autocorrelation (Valavi et al., 2018). These options allow users to account for spatial or temporal autocorrelation in the data and reduce overoptimistic performance estimates due to autocorrelation.

## 2.4 | Model output and projections

### 2.4.1 | Prediction and uncertainty

GLOSSA uses the fitted BART model to generate predictions from the posterior predictive distribution, allowing users to summarize predictions and quantify uncertainty using point estimates (e.g. mean, median, standard deviation) and credible intervals. By default, when fitting a model, GLOSSA projects the prediction of the suitable habitat and/or native range models on an averaged environmental scenario calculated as the mean of the fitting period. This approach provides a general prediction to explore the model output.

### 2.4.2 | Projections

When projection layers are provided (e.g. future scenarios), GLOSSA applies the fitted model to each scenario independently and computes summary metrics from the predictive posterior distribution across projections.

### 2.4.3 | Other model outputs

Additionally, GLOSSA estimates the functional responses (response curves, the relationship between the occurrence of a species and each environmental variable) as partial dependence plots (Friedman, 2001). These plots show the marginal effect that each variable has on the

predicted outcome of the BART model. The response curves are calculated for the suitable habitat model without variable standardization so they can be interpreted on the original scale. In addition, variable importance is computed using a permutation-based approach (Wei et al., 2015), measuring the change in the prediction error of the model using the F-score after we permuted the variable's values.

### 2.4.4 | Visualization and export

After completing the analysis, users can explore the results through interactive visualizations, including predicted maps, model summary and diagnostic plots, response curves and variable importance. All outputs can be exported for further analysis, including a configuration JSON file with all analysis settings and input files for reproducibility.

## 2.5 | User interface and implementation

GLOSSA is implemented as an R package named *glossa*, accessible on the Comprehensive R Archive Network (CRAN, <https://cran.r-project.org/package=glossa>) and maintained on GitHub (<https://github.com/iMARES-group/glossa>). Its intuitive user interface is built using *shiny* (Chang et al., 2024), *bs4Dash* (Granjon, 2024) and *shinyWidgets* (Perrier et al., 2024), while the BART modelling approach relies on the *dbarts* package (Dorie, 2024).

The app is structured into four main tabs: *Home*, *New Analysis*, *Reports* and *Export*. The *Home* tab offers access to detailed documentation, quick-start guides for first-time users, and tutorials and examples. The *New Analysis* tab (Figure 2a) allows users to upload input data, select the analysis options and validate and pre-visualize the loaded data, including species occurrences, environmental data and the study area boundaries. Additionally, users can define different predictor variables for each species within the same session, providing flexibility in modelling more than one species dataset.

Once the analysis is complete, the *Reports* tab (Figure 2b) provides interactive visualizations of model predictions, occurrences, environmental layers, functional responses, variable importance and validation metrics. Users can explore summary and uncertainty metrics from the posterior predictive distribution (e.g. mean, median, credible intervals) and analyse spatial and temporal trends in habitat suitability. Cross-validation results show the performance metrics in radar charts and occurrence maps with fold partitioning. Functional responses are represented using the mean response curve, with a shaded area representing the 95% credible interval, and variable importance is illustrated in boxplots, summarizing the results of the iterations of permutation variable importance, sorted according to the mean importance. All plot options can be accessed via the three-dot button in the top-right corner of each plot. Also, users are able to maximize plots for better visualization and export them in various formats.

The *Export* tab enables users to download all results, including model outputs, and a JSON configuration file containing input meta-data and analysis settings. This file ensures full reproducibility and

(a)

**Data upload**  
Upload your species occurrence data, environmental rasters, projection layers, and an optional study area polygon.

**Select analysis**  
Choose the model (native range/suitable habitat) and enable projections or additional outputs.

**Predictor variables**  
Assign environmental predictors for each species.

**Advanced options**  
Access advanced settings for further customization.

**Preview input data**  
Interactively preview your occurrence and environmental data. Open additional options using the button

**Run your analysis**  
Click to start the analysis when everything is ready.

**Validate input data**  
Ensure your data is correctly formatted.

| File name               | File size | Format Validation |
|-------------------------|-----------|-------------------|
| 1. sp1.txt              | 11823     | ✓                 |
| 2. sp2.csv              | 13622     | ✓                 |
| 3. env_layers.zip       | 47740     | ✓                 |
| 4. project_layers_L.zip | 48465     | ✓                 |

(b)

**Select species**  
If multiple occurrence files are uploaded, select which species results to display.

**Potential suitability**  
Trend of projected mean suitable probability and suitable area over time.

**P/A records**  
Number of presence and (pseudo-)absence records used to fit the model.

**GLOSSA predictions**  
Projection maps showing model predictions. You can choose the model (SH or NR), projection scenario, timestamp, and posterior distribution metric to plot.

**Environmental variables**  
Visualization of the environmental predictor variables used in the model.

**Presence validation**  
Visualization of filtered occurrence records.

**Plot options**  
Use the three dots icon to customize visualizations and explore the results in more detail.

**Export plots**  
Use the download icon to export any plot.

**Additional analysis**  
Explore functional responses, variable importance, and cross-validation results.

**Model summary**  
Validate your model by checking the ROC curve, AUC, TSS, CBI, confusion matrix, and distribution of fitted values.

**FIGURE 2** Screenshots of GLOSSA's user interface. (a) Data input and settings in the *New Analysis* tab. (b) Interactive visualizations in the *Reports* tab.

allows users to replicate the analysis under the same conditions. Outputs are provided in standard and interoperable formats to ensure compatibility with other downstream applications such as GIS platforms and Marine Ecosystem Models (MEMs).

### 3 | WORKED EXAMPLES

We demonstrate the use of GLOSSA through three case studies involving marine species at global, regional and local scales. Complete details for these examples, including data and code, are available in

the vignettes provided in Files [S1–S3](#) and on the package website (<https://imares-group.github.io/glossa/>). The associated datasets can also be accessed via Zenodo (<https://doi.org/10.5281/zenodo.13833556>). Below, we provide a brief overview of each case study.

#### 3.1 | Case study 1. Climate change impacts on yellowfin tuna global distribution

In our first example, we modelled the global habitat suitability of yellowfin tuna (*Thunnus albacares*) under two different climate

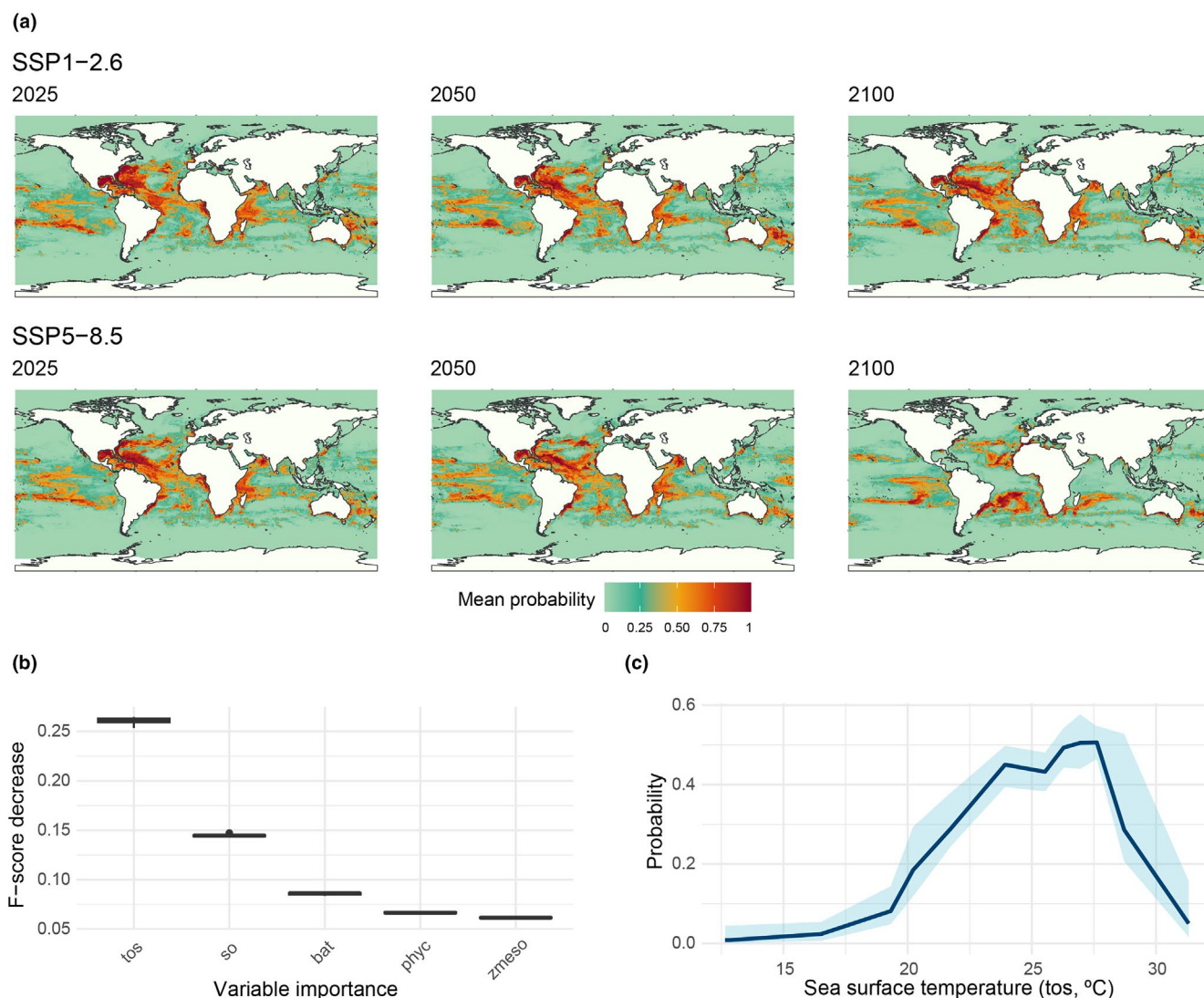
scenarios: SSP1-2.6, representing a sustainable future with reduced CO<sub>2</sub> emissions, and SSP5-8.5, a high-emission scenario characterized by increased global warming. Yellowfin tuna is an important migratory predator in the world's oceans and accounts for a significant portion of the global catch (FAO, 2024), making it essential to understand how climate change could impact its large-scale distribution (Erauskin-Extramiana et al., 2019).

Occurrence data were obtained from the OBIS (OBIS, 2025), while environmental variables, including sea surface temperature (°C), sea surface salinity (psu), primary productivity (mol/m<sup>2</sup>) and mesozooplankton (mol/m<sup>2</sup>) were obtained from the ISIMIP3b climate dataset using the GFDL-ESM4 Earth System Model (Warszawski et al., 2014). Bathymetry data were downloaded from the ETOPO 2022 Global Relief Model developed by the National Oceanic and Atmospheric Administration (NOAA,

<https://www.ncei.noaa.gov/products/etopo-global-relief-model>). Before fitting the model, a z-score standardization was applied to the variables.

The results show a decline in suitable habitat for *Thunnus albacares* from 2025 to 2100 under both scenarios, with a 7.9% reduction under SSP1-2.6 and a more intense 37.4% reduction under SSP5-8.5 (Figure 3). Under the high-emission scenario, suitable habitats shift from equatorial and tropical regions toward the subtropics, reflecting the impact of increased warming on the distributional shift of the species.

Variable importance analysis, measured through permutation accuracy using the F-score, shows that sea surface temperature is the most influential variable in predicting species occurrences, followed by salinity and bathymetry. The response curve for temperature shows a peak occurrence probability around 26–27°C.



**FIGURE 3** (a) Mean posterior probability of projected suitable habitats for *Thunnus albacares* under the SSP1-2.6 and SSP5-8.5 climate scenarios on a global scale. (b) Variable importance obtained from a permutation-based approach using the *F*-score (tos: Sea surface temperature, so: Salinity, bat: Bathymetry, phyc: Primary productivity, zmeso: Mesozooplankton). (c) Response curve (partial dependence plot) for sea surface temperature with 95% credible intervals.

These results highlight the role of temperature in shaping species distribution and the potential shifts and losses of suitable habitats under future climate scenarios of a widely distributed species. Additional results are provided in File S1.

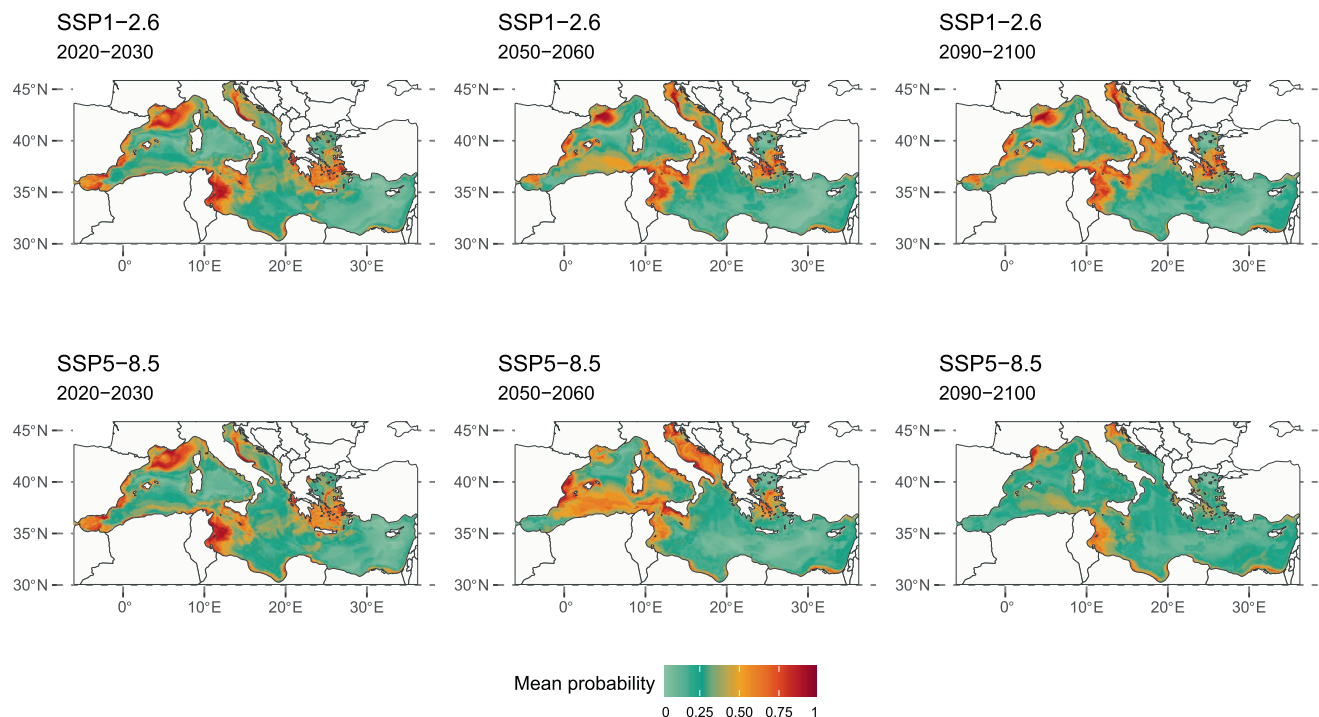
### 3.2 | Case study 2. Suitable habitat of *Caretta caretta* in the Mediterranean Sea

This case study illustrates the application of GLOSSA within a user-defined region, focusing on the loggerhead turtle (*Caretta caretta*) in the Mediterranean Sea. While globally listed as vulnerable by the IUCN Red List, the species is classified as least concern in the Mediterranean. However, climate change and global warming represent a risk on their life cycle and ecological niche (Mancino et al., 2022). Here, we analyse how different future climate scenarios will reshape the habitat suitability of the Mediterranean Sea for *Caretta caretta*. We used occurrence data from GBIF (GBIF.org, 2024), and we filtered records from 2000 to 2020. These data were combined with environmental variables from Bio-ORACLE v3.0 (Assis et al., 2024, <https://www.bio-oracle.org/>) from the 2000–2010 and 2010–2020 decades. The environmental variables used in the analysis included sea surface temperature (°C), primary productivity (mol/m<sup>2</sup>), salinity (psu) and bathymetry (m) from the ETOPO 2022 Global Relief Model by NOAA. After fitting the SDM with the standardized variables, we projected the suitable habitat under two different climate scenarios: SSP1-2.6 (sustainable development) and SSP5-8.5 (high emissions).

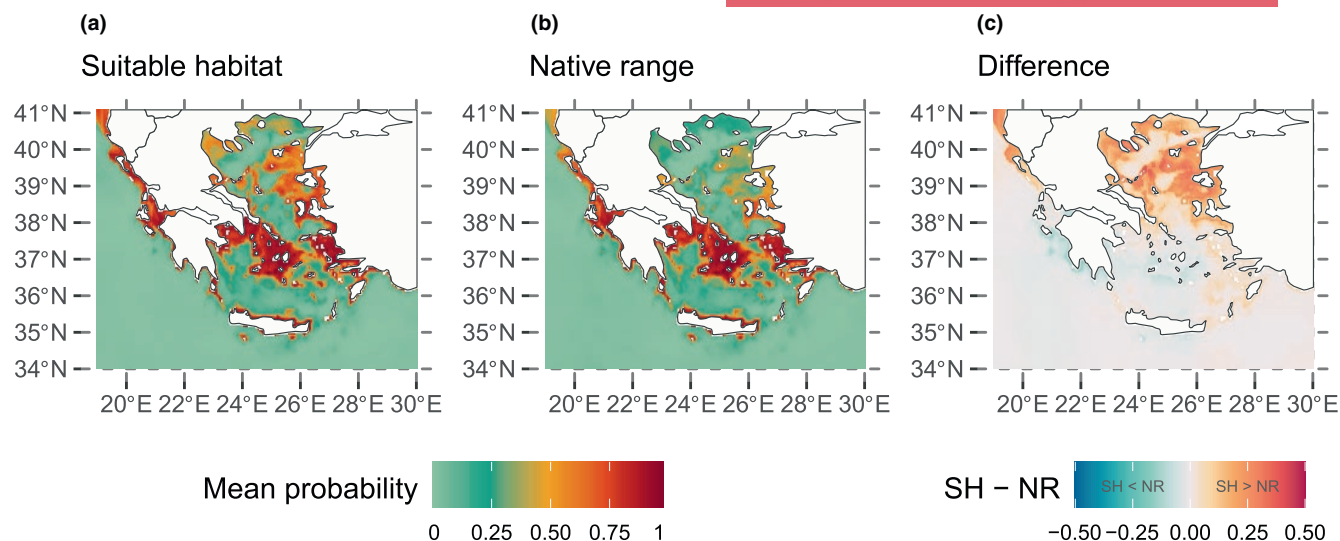
The results show a loss of habitat suitability in both scenarios from the 2020s to the 2090s (Figure 4). However, the decline is much more pronounced under SSP5-8.5 and during the last period, where up to 79% of suitable habitat area is lost. Additionally, habitat suitability in both scenarios tends to be higher in the western Mediterranean and the Adriatic Sea than in the eastern basin, where the loss of habitat is clear. Further details on the analysis and results can be found in File S2.

### 3.3 | Case study 3. Potential risk areas of *Siganus luridus* in the Greek Seas

Our third case study demonstrates GLOSSA's application at a local scale, assessing potential invasion risks of the rabbit fish *Siganus luridus* in the Greek Seas. This species is one of many Lessepsian migrants entering the Mediterranean through the Suez Canal, posing challenges to native species and making it crucial to predict their potential spread to inform management strategies (Solanou et al., 2023). Occurrence data from 2000 to 2020 was obtained from an individual dataset from OBIS (Sini et al., 2024). The environmental variables used as predictors were sea surface temperature (°C), sea water salinity (psu), net primary production of biomass (mg/m<sup>3</sup>/day) and dissolved molecular oxygen (mol/m<sup>2</sup>), obtained from the Copernicus Marine Environment Monitoring Service (<https://marine.copernicus.eu/>). We also included bathymetry (m) from the ETOPO 2022 Global Relief Model by NOAA. These variables were standardized to have a zero mean and unit variance. We used the GLOSSA application to predict both the native range and the suitable habitat of



**FIGURE 4** Mean posterior probability of projected suitable habitats for *Caretta caretta* under the SSP1-2.6 and SSP5-8.5 climate scenarios in the Mediterranean Sea.



**FIGURE 5** Mean posterior probability for the (a) suitable habitat and (b) native range of *Siganus luridus* in the Greek seas in 2020 and (c) the difference between both predictions (SH: Suitable habitat, NR: Native range).

*Siganus luridus* in the year 2020. While the suitable habitat model is based only on environmental variables, the native range model also includes latitude and longitude as predictors.

Figure 5 shows the mean probability of the posterior predictive distribution for the native range and the suitable habitat, revealing differences between both models. The spatially smoothed native range predicts where the species is currently present, while the suitable habitat identifies areas with environmental conditions favourable for the species, even where it has not yet been recorded. For example, the Thracian Sea and northeast Aegean islands were identified as high-risk areas for potential invasion, even though the species is not currently established there according to the predicted native range in 2020. However, there is high uncertainty in the Thracian Sea predictions as the environmental conditions are outside the range of the ones matching the occurrence records (Solanou et al., 2023). More information on this case study can be found in File S3.

### 3.4 | Memory usage and computation time

The runtime and memory usage of GLOSSA are influenced by several factors that increase the complexity of the analysis, including the number of species analysed in a single run, the number of species occurrences, the number of predictor variables, the size and resolution of the study area and the number of requested projections. High-resolution rasters and large spatial extents can increase memory requirements and slow performance. We therefore recommend first testing with cropped or aggregated rasters. To provide some benchmarks, we report the computation time from the analysis of the worked examples. The analyses were conducted on a single Windows 11 machine equipped with 64GB of RAM and an Intel(R) Core(TM) i7-1165G7 processor. This processor features 4 cores and 8 threads, with a base clock speed of 2.80GHz. The global analysis

of *Thunnus albacares* took 2h and 13 min, while the regional models for *Caretta caretta* and *Siganus luridus* took 20 and 4 minutes, respectively. Detailed computation times for each step of the analysis are provided in the Files S1–S3.

## 4 | CONCLUSIONS AND FUTURE DIRECTIONS

We introduced GLOSSA, an open-source R Shiny application that lowers technical barriers for species distribution modelling through an intuitive GUI and comprehensive documentation, making it accessible to a broader audience. It provides tools for occurrence and environmental data processing, pseudo-absence generation, model fitting using BART, model evaluation and interactive visualizations within a reproducible framework.

Currently, GLOSSA relies on BART, a promising method for SDMs based on an ensemble of trees, similar to boosted regression trees and Bayesian inference. This allows modelling complex non-linear relationships while generating posterior distributions for predictions, improving the estimation of uncertainty. However, the use of BART in SDMs is relatively new, and questions remain regarding its behaviour with different pseudo-absence strategies, sample sizes, multicollinearity and its transferability across space and time. Recent developments, such as BART extensions for supporting continuous spatial effects (Jiang & Wakefield, 2023) and improving extrapolation (Maia et al., 2024), may be particularly relevant in the SDM context.

We demonstrated the use of GLOSSA through three marine case studies addressing different ecological scenarios and spatial scales, such as evaluating the effects of climate change on *Thunnus albacares* and *Caretta caretta* range shifts and identifying potential invasion risk areas for *Siganus luridus*. Although GLOSSA is designed for SDM workflows, its outputs, such as suitability maps and functional

responses, can be integrated into broader analysis frameworks. For example, Marine Ecosystem Models (MEMs), which often rely on extensive parametrization and require a large amount of data to operate effectively (Robson et al., 2018; Steenbeek et al., 2021), could benefit from SDM outputs (Coll et al., 2019, 2020). Future work should explore this link.

Users should remain aware of common challenges associated with SDMs, including sampling bias, autocorrelation, scale mismatches and spatial and temporal misalignment between occurrences and predictors (Araújo et al., 2019; Kass et al., 2025; Melo-Merino et al., 2020). While GLOSSA provides some tools for spatial thinning, pseudo-absences generation and model validation, it has limited functionalities, so preprocessing and thoughtful modelling decisions remain essential and may need to be performed outside the GLOSSA app, like detailed occurrence cleaning or environmental bias correction.

Future development will focus on expanding bias correction options, including additional pseudo-absence generation methods and thinning methods that account for the environmental space. Although users can adjust BART configuration parameters, adding simpler modelling alternatives and ensemble modelling may help reduce overfitting under biased or limited data. Additionally, extrapolation diagnostics such as Multivariate Environmental Similarity Surface (MESS) maps could complement current model uncertainty estimates of BART.

Finally, documentation, as well as updates, examples and tutorials, is available at the GLOSSA project site: <https://imares-group.github.io/glossa/>.

## AUTHOR CONTRIBUTIONS

Jorge Mestre-Tomás, Alba Fuster-Alonso, José M. Bellido and Marta Coll conceptualized and designed the GLOSSA application and the manuscript. Jorge Mestre-Tomás developed and implemented GLOSSA, performed the analyses and generated the visualizations. Alba Fuster-Alonso contributed to the GLOSSA workflow, particularly in the modelling approach using BARTs. José M. Bellido and Marta Coll obtained the funding and provided supervision, offering insights and guidance throughout the project. Jorge Mestre-Tomás led the writing of the manuscript. All authors contributed critically to the drafts and approved the final version.

## ACKNOWLEDGEMENTS

We would like to thank the iMARES group for their helpful feedback on the app design. This study is part of the ProOceans project (PID2020-118097RB-I00) funded by the Spanish Ministry of Science and Innovation. We also acknowledge support from the SOSPen (PID2021-124831OA-I00) project funded by the Spanish Ministry of Science and Innovation and the GES4SEAS project (grant agreement no. 101059877) funded by the European Union under the Horizon Europe program. The authors also acknowledge the institutional support of the 'Severo Ochoa Center of Excellence' accreditation (CEX2024-001494-S funded by AEI [10.13039/501100011033](https://doi.org/10.13039/501100011033)) to the Institute of Marine Science (ICM-CSIC). Additionally, this research is part of the Integrated Marine Ecosystem Assessments (iMARES) research group, funded by the Agència de Gestió d'Ajuts

Universitaris i de Recerca of the Generalitat de Catalunya (2021 SGR 00435), and the Global Ecology and Fisheries (GLOFISH) research group of the IEO-CSIC. The authors acknowledge partial funding from Ministerio para la Transición Ecológica y el Reto Demográfico (MITECO).

## CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

## PEER REVIEW

The peer review history for this article is available at <https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210x.70167>.

## DATA AVAILABILITY STATEMENT

GLOSSA is developed openly on GitHub (<https://github.com/iMARES-group/glossa>), with releases archived on Zenodo (<https://doi.org/10.5281/zenodo.17161715>; Mestre-Tomás et al., 2025a). The R package `glossa` is also available via the Comprehensive R Archive Network (CRAN; <https://cran.r-project.org/package=glossa>). Documentation is available at <https://imares-group.github.io/glossa/>. Data and code used in the examples are available on Zenodo (<https://doi.org/10.5281/zenodo.17177436>; Mestre-Tomás et al., 2025b). Occurrence data for *Thunnus albacares* are available from OBIS (<https://obis.org/taxon/127027>; OBIS, 2025), for *Caretta caretta* from GBIF (<https://doi.org/10.15468/dl.es7562>; GBIF, 2024) and for *Siganus luridus* from the GreekMarineICAS geo-dataset (<https://doi.org/10.25607/t2smha>; Sini et al., 2024).

## ORCID

Jorge Mestre-Tomás  <https://orcid.org/0000-0002-8983-3417>  
 Alba Fuster-Alonso  <https://orcid.org/0000-0002-7283-291X>  
 José M. Bellido  <https://orcid.org/0000-0002-6887-4391>  
 Marta Coll  <https://orcid.org/0000-0001-6235-5868>

## REFERENCES

- 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(6), 1223–1232. <https://doi.org/10.1111/j.1365-2664.2006.01214.x>
- Araújo, M. B., Anderson, R. P., Márcia Barbosa, A., Beale, C. M., Dormann, C. F., Early, R., García, R. A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R. B., Zimmermann, N. E., & Rahbek, C. (2019). Standards for distribution models in biodiversity assessments. *Science Advances*, 5(1), eaat4858. <https://doi.org/10.1126/sciadv.aat4858>
- Assis, J., Fernández Bejarano, S. J., Salazar, V. W., Schepers, L., Gouvêa, L., Fragkopoulou, E., Leclercq, F., Vanhoorne, B., Tyberghein, L., Serrão, E. A., Verbruggen, H., & De Clerck, O. (2024). Bio-ORACLE v3.0. Pushing marine data layers to the CMIP6 Earth System Models of climate change research. *Global Ecology and Biogeography*, 33(4), e13813. <https://doi.org/10.1111/geb.13813>
- Baquero, R. A., Barbosa, A. M., Ayllón, D., Guerra, C., Sánchez, E., Araújo, M. B., & Nicola, G. G. (2021). Potential distributions of invasive vertebrates in the Iberian Peninsula under projected changes in climate extreme events. *Diversity and Distributions*, 27(11), 2262–2276. <https://doi.org/10.1111/ddi.13401>

- 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(2), 327–338. <https://doi.org/10.1111/j.2041-210X.2011.00172.x>
- Booth, T. H. (2018). Species distribution modelling tools and databases to assist managing forests under climate change. *Forest Ecology and Management*, 430, 196–203. <https://doi.org/10.1016/j.foreco.2018.08.019>
- Boyce, M. S., Vernier, P. R., Nielsen, S. E., & Schmiegelow, F. K. (2002). Evaluating resource selection functions. *Ecological Modelling*, 157(2–3), 281–300. [https://doi.org/10.1016/S0304-3800\(02\)00200-4](https://doi.org/10.1016/S0304-3800(02)00200-4)
- Broussin, J., Mouchet, M., & Goberville, E. (2024). Generating pseudo-absences in the ecological space improves the biological relevance of response curves in species distribution models. *Ecological Modelling*, 498, 110865. <https://doi.org/10.1016/j.ecolmodel.2024.110865>
- Brown, C. J., O'Connor, M. I., Poloczanska, E. S., Schoeman, D. S., Buckley, L. B., Burrows, M. T., Duarte, C. M., Halpern, B. S., Pandolfi, J. M., Parmesan, C., & Richardson, A. J. (2016). Ecological and methodological drivers of species' distribution and phenology responses to climate change. *Global Change Biology*, 22(4), 1548–1560. <https://doi.org/10.1111/gcb.13184>
- Carlson, C. J. (2020). embarcadero: Species distribution modelling with Bayesian additive regression trees in R. *Methods in Ecology and Evolution*, 11(7), 850–858. <https://doi.org/10.1111/2041-210X.13389>
- Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2024). shiny: Web application framework for R. R package version 1.8.1.9001. <https://github.com/rstudio/shiny>, <https://shiny.posit.co/>
- Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), 266–298. <https://doi.org/10.1214/09-AOAS285>
- Coll, M., Pennino, M. G., Steenbeek, J., Solé, J., & Bellido, J. M. (2019). Predicting marine species distributions: Complementarity of food-web and Bayesian hierarchical modelling approaches. *Ecological Modelling*, 405, 86–101. <https://doi.org/10.1016/j.ecolmodel.2019.05.005>
- Coll, M., Steenbeek, J., Pennino, M. G., Buszowski, J., Kaschner, K., Lotze, H. K., Rousseau, Y., Tittensor, D. P., Walters, C., Watson, R. A., & Christensen, V. (2020). Advancing global ecological modeling capabilities to simulate future trajectories of change in marine ecosystems. *Frontiers in Marine Science*, 7, 567877. <https://doi.org/10.3389/fmars.2020.567877>
- Coro, G., Sana, L., & Bove, P. (2025). An open science automatic workflow for multi-model species distribution estimation. *International Journal of Data Science and Analytics*, 20, 1131–1150. <https://doi.org/10.1007/s41060-024-00517-w>
- Dorie, V. (2024). dbarts: Discrete Bayesian additive regression trees sampler. R package version 0.9-28 <https://CRAN.R-project.org/package=dbarts>
- Erauskin-Extramiana, M., Arrizabalaga, H., Hobday, A. J., Cabré, A., Ibaibarriaga, L., Arregui, I., Murua, H., & Chust, G. (2019). Large-scale distribution of tuna species in a warming ocean. *Global Change Biology*, 25(6), 2043–2060. <https://doi.org/10.1111/gcb.14630>
- FAO. (2024). *The State of World Fisheries and Aquaculture 2024 – Blue Transformation in action*. Rome.
- Ferrari, R., Malcolm, H., Neilson, J., Lucieer, V., Jordan, A., Ingleton, T., Figueira, W., Johnstone, N., & Hill, N. (2018). Integrating distribution models and habitat classification maps into marine protected area planning. *Estuarine, Coastal and Shelf Science*, 212, 40–50. <https://doi.org/10.1016/j.ecss.2018.06.015>
- Figueira, M., Conesa, D., & López-Quílez, A. (2024). A shiny R app for spatial analysis of species distribution models. *Ecological Informatics*, 80, 102542. <https://doi.org/10.1016/j.ecoinf.2024.102542>
- Fournier, A., Penone, C., Pennino, M. G., & Courchamp, F. (2019). Predicting future invaders and future invasions. *Proceedings of the National Academy of Sciences of the United States of America*, 116(16), 7905–7910. <https://doi.org/10.1073/pnas.1803456116>
- Franklin, J. (2013). Species distribution models in conservation biogeography: Developments and challenges. *Diversity and Distributions*, 19(10), 1217–1223. <https://doi.org/10.1111/ddi.12125>
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29, 1189–1232. <https://www.jstor.org/stable/2699986>
- Fuster-Alonso, A., Mestre-Tomás, J., Baez, J. C., Pennino, M. G., Barber, X., Bellido, J. M., Conesa, D., López-Quílez, A., Steenbeek, J., Christensen, V., & Coll, M. (2024). Machine learning applied to global scale species distribution models (SDMs). *PREPRINT (Version 1) available at Research Square*. <https://doi.org/10.21203/rs.3.rs-4411399/v1>
- GBIF.org. (2020). GBIF home page. <https://www.gbif.org/>
- GBIF.org. (2024). GBIF occurrence download. <https://doi.org/10.15468/dl.es7562>
- Ge, X., Griswold, C. K., & Newman, J. A. (2024). Robust species distribution predictions of predator and prey responses to climate change. *Journal of Biogeography*, 51, 2047–2061. <https://doi.org/10.1111/jbi.14969>
- Granjon, D. (2024). bs4Dash: A 'Bootstrap 4' version of 'shinydashboard'. R package version 2.3.3. <https://github.com/Rinterface/bs4Dash>, <https://rinterface.github.io/bs4Dash/index.html>
- Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I., Regan, T. J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T. G., Rhodes, J. R., Maggini, R., Setterfield, S. A., Elith, J., Schwartz, M. W., Wintle, B. A., Broennimann, O., Austin, M., ... Buckley, Y. M. (2013). Predicting species distributions for conservation decisions. *Ecology Letters*, 16(12), 1424–1435. <https://doi.org/10.1111/ele.12189>
- Hertzog, L. R., Besnard, A., & Jay-Robert, P. (2014). Field validation shows bias-corrected pseudo-absence selection is the best method for predictive species-distribution modelling. *Diversity and Distributions*, 20(12), 1403–1413. <https://doi.org/10.1111/ddi.12249>
- Hill, J. L., Grismik, M., Hanscom, R. J., Sukumaran, J., Higham, T. E., & Clark, R. W. (2024). The past, present, and future of predator-prey interactions in a warming world: Using species distribution modeling to forecast ectotherm-endotherm niche overlap. *Ecology and Evolution*, 14(3), e11067. <https://doi.org/10.1002/ece3.11067>
- Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C., & Guisan, A. (2006). Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, 199(2), 142–152. <https://doi.org/10.1016/j.ecolmodel.2006.05.017>
- Hunt, T. N., Allen, S. J., Bejder, L., & Parra, G. J. (2020). Identifying priority habitat for conservation and management of Australian humpback dolphins within a marine protected area. *Scientific Reports*, 10(1), 14366. <https://doi.org/10.1038/s41598-020-69863-6>
- Iturbide, M., Bedia, J., Herrera, S., del Hierro, O., Pinto, M., & Gutiérrez, J. M. (2015). A framework for species distribution modelling with improved pseudo-absence generation. *Ecological Modelling*, 312, 166–174. <https://doi.org/10.1016/j.ecolmodel.2015.05.018>
- Jiang, A. Z., & Wakefield, J. (2023). BART-SIMP: A novel framework for flexible spatial covariate modeling and prediction using Bayesian additive regression trees. *arXiv preprint arXiv:2309.13270*. <https://doi.org/10.48550/arXiv.2309.13270>
- Kass, J. M., Pinilla-Buitrago, G. E., Paz, A., Johnson, B. A., Grisales-Betancur, V., Meenan, S. I., Attali, D., Broennimann, O., Galante, P. J., Maitner, B. S., Owens, H. L., Varela, S., Aiello-Lammens, M. E., Merow, C., Blair, M. E., & Anderson, R. P. (2023). wallace 2: A shiny app for modeling species niches and distributions redesigned to facilitate expansion via module contributions. *Ecography*, 2023(3), e06547. <https://doi.org/10.1111/ecog.06547>

- 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., ... Zurell, D. (2025). Achieving higher standards in species distribution modeling by leveraging the diversity of available software. *Ecography*, 2025(2), e07346. <https://doi.org/10.1111/ecog.07346>
- Klaassen, M., Marques, T. A., Alves, F., & Fernandez, M. (2025). Trends in marine species distribution models: A review of methodological advances and future challenges. *Ecography*, e07702. <https://doi.org/10.1002/ecog.07702>
- Konowalik, K., & Nosol, A. (2021). Evaluation metrics and validation of presence-only species distribution models based on distributional maps with varying coverage. *Scientific Reports*, 11(1), 1482. <https://doi.org/10.1038/s41598-020-80062-1>
- Liu, S., Tian, Y., Liu, Y., Alabía, I. D., Cheng, J., & Ito, S.-i. (2023). Development of a prey-predator species distribution model for a large piscivorous fish: A case study for Japanese Spanish mackerel *Scomberomorus niphonius* and Japanese anchovy *Engraulis japonicus*. *Deep Sea Research Part II: Topical Studies in Oceanography*, 207, 105227. <https://doi.org/10.1016/j.dsr2.2022.105227>
- Lyons, D. A., Lowen, J. B., Theriault, T. W., Brickman, D., Guo, L., Moore, A. M., Peña, M. A., Wang, Z., & DiBacco, C. (2020). Identifying marine invasion hotspots using stacked species distribution models. *Biological Invasions*, 22, 3403–3423. <https://doi.org/10.1007/s10530-020-02332-3>
- Maia, M., Murphy, K., & Parnell, A. C. (2024). GP-BART: A novel Bayesian additive regression trees approach using Gaussian processes. *Computational Statistics & Data Analysis*, 190, 107858. <https://doi.org/10.1016/j.csda.2023.107858>
- Mancino, C., Canestrelli, D., & Maiorano, L. (2022). Going west: Range expansion for loggerhead sea turtles in the Mediterranean Sea under climate change. *Global Ecology and Conservation*, 38, e02264. <https://doi.org/10.1016/j.gecco.2022.e02264>
- Melo-Merino, S. M., Reyes-Bonilla, H., & Lira-Noriega, A. (2020). Ecological niche models and species distribution models in marine environments: A literature review and spatial analysis of evidence. *Ecological Modelling*, 415, 108837. <https://doi.org/10.1016/j.ecolmodel.2019.108837>
- Mestre-Tomás, J. (2025). GeoThinnerR: An R package for efficient spatial thinning of species occurrences and point data. *arXiv preprint arXiv:2505.07867*. <https://doi.org/10.48550/arXiv.2505.07867>
- Mestre-Tomás, J., Fuster-Alonso, A., Bellido, J. M., & Coll, M. (2025a). GLOSSA (version all versions). *Zenodo*. <https://doi.org/10.5281/zenodo.17161714>
- Mestre-Tomás, J., Fuster-Alonso, A., Bellido, J. M., & Coll, M. (2025b). Supplementary materials for "GLOSSA: A user-friendly R Shiny application for Bayesian machine learning analysis of marine species distribution". *Zenodo*. <https://doi.org/10.5281/zenodo.17177436>
- Mislán, K., Heer, J. M., & White, E. P. (2016). Elevating the status of code in ecology. *Trends in Ecology & Evolution*, 31(1), 4–7. <https://doi.org/10.1016/j.tree.2015.11.006>
- Moudry, V., Bazzichetto, M., Remelgado, R., Devillers, R., Lenoir, J., Mateo, R. G., Lembrechts, J. J., Sillero, N., Lecours, V., Cord, A. F., Barták, V., Balej, P., Rocchini, D., Torresani, M., Arenas-Castro, S., Man, M., Pražlerová, D., Gdulová, K., Prošek, J., ... Šimová, P. (2024). Optimising occurrence data in species distribution models: Sample size, positional uncertainty, and sampling bias matter. *Ecography*, 2024(12), e07294. <https://doi.org/10.1111/ecog.07294>
- OBIS. (2025). Distribution records of *Thunnus albacares* (Bonnaterre, 1788). Ocean Biodiversity Information System. Intergovernmental Oceanographic Commission of UNESCO. <https://obis.org/taxon/127027>
- Osorio-Olvera, L., Lira-Noriega, A., Soberón, J., Peterson, A. T., Falconi, M., Contreras-Díaz, R. G., Martínez-Meyer, E., Barve, V., & Barve, N. (2020). ntbox: An R package with graphical user interface for modelling and evaluating multidimensional ecological niches. *Methods in Ecology and Evolution*, 11(10), 1199–1206. <https://doi.org/10.1111/2041-210X.13452>
- Perrier, V., Meyer, F., & Granjon, D. (2024). shinyWidgets: Custom Inputs Widgets for Shiny. R package version 0.8.6. <https://dreamrs.github.io/shinyWidgets/>, <https://github.com/dreamRs/shinyWidgets>
- Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological Applications*, 19(1), 181–197. <https://doi.org/10.1890/07-2153.1>
- Pulliam, H. R. (2000). On the relationship between niche and distribution. *Ecology Letters*, 3(4), 349–361. <https://doi.org/10.1046/j.1461-0248.2000.00143.x>
- Robson, B. J., Arhonditsis, G. B., Baird, M. E., Brebion, J., Edwards, K. F., Geoffroy, L., Hébert, M.-P., van Dongen-Vogels, V., Jones, E. M., Kruk, C., Mongin, M., Shimoda, Y., Skerratt, J. H., Trevathan-Tackett, S. M., Wild-Allen, K., Kong, X., & Steven, A. (2018). Towards evidence-based parameter values and priors for aquatic ecosystem modelling. *Environmental Modelling & Software*, 100, 74–81. <https://doi.org/10.1016/j.envsoft.2017.11.018>
- Ronquillo, C., Stropp, J., & Hortal, J. (2024). OCCUR Shiny application: A user-friendly guide for curating species occurrence records. *Methods in Ecology and Evolution*, 15(5), 816–823. <https://doi.org/10.1111/2041-210X.14271>
- Rubbens, P., Brodie, S., Cordier, T., Destro Barcellos, D., Devos, P., Fernandes-Salvador, J. A., Fincham, J. I., Gomes, A., Handegard, N. O., Howell, K., Jamet, C., Kartveit, K. H., Moustahfid, H., Parcerisas, C., Politikos, D., Sauzède, R., Sokolova, M., Uusitalo, L., van den Bulcke, L., ... Irsson, J. O. (2023). Machine learning in marine ecology: An overview of techniques and applications. *ICES Journal of Marine Science*, 80(7), 1829–1853. <https://doi.org/10.1093/icesjms/fsad100>
- Schickele, A., Clerc, C., Benedetti, F., De Angelis, D., Hofmann Elizondo, U., Münnich, M., Irsson, J.-O., & Vogt, M. (2025). CEPHALOPOD, a package to standardize marine habitat-modelling practices and enhance inter-comparability across biological observations. *Methods in Ecology and Evolution*, 16, 1126–1135. <https://doi.org/10.1111/2041-210X.70040>
- Sillero, N., Arenas-Castro, S., Enriquez-Urzelai, U., Vale, C. G., Sousa-Guedes, D., Martínez-Freiria, F., Real, R., & Barbosa, A. M. (2021). Want to model a species niche? A step-by-step guideline on correlative ecological niche modelling. *Ecological Modelling*, 456, 109671. <https://doi.org/10.1016/j.ecolmodel.2021.109671>
- Sini, M., Ragkousis, M., Katsanevakis, S., Zenetos, A., & Koukourouli, N. (2024). Marine impactful cryptogenic and alien species in the Greek Seas: A georeferenced dataset (1893–2020). Version 2.0. Hellenic Center for Marine Research. Occurrence dataset. <https://doi.org/10.25607/t2smha>
- Solanou, M., Valavanis, V. D., Karachle, P. K., & Giannoulaki, M. (2023). Looking at the expansion of three demersal leptocephalus fish immigrants in the Greek Seas: What can we get from spatial distribution modeling? *Diversity*, 15(6), 776. <https://doi.org/10.3390/d15060776>
- Srivastava, V., Lafond, V., & Griess, V. C. (2019). Species distribution models (SDM): Applications, benefits and challenges in invasive species management. *CABI Reviews*, 2019, 1–13. <https://doi.org/10.1079/PAVSNNR201914020>
- Steenbeek, J., Buszowski, J., Chagaris, D., Christensen, V., Coll, M., Fulton, E. A., Katsanevakis, S., Lewis, K. A., Mazaris, A. D., Macias, D., de Mutsert, K., Oldford, G., Pennino, M. G., Piroddi, C., Romagnoni, G., Serpetti, N., Shin, Y. J., Spence, M. A., & Stelzenmüller, V. (2021). Making spatial-temporal marine ecosystem modelling better—A perspective. *Environmental Modelling & Software*, 145, 105209. <https://doi.org/10.1016/j.envsoft.2021.105209>
- Thompson, M. S., Couce, E., Schratzberger, M., & Lynam, C. P. (2023). Climate change affects the distribution of diversity across marine

- food webs. *Global Change Biology*, 29(23), 6606–6619. <https://doi.org/10.1111/gcb.16881>
- Tyberghein, L., Verbruggen, H., Pauly, K., Troupin, C., Mineur, F., & De Clerck, O. (2012). Bio-ORACLE: A global environmental dataset for marine species distribution modelling. *Global Ecology and Biogeography*, 21(2), 272–281. <https://doi.org/10.1111/j.1466-8238.2011.00656.x>
- Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2018). block CV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, 10(2), 225–232. <https://doi.org/10.1111/2041-210X.13107>
- VanDerWal, J., Shoo, L. P., Graham, C., & Williams, S. E. (2009). Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? *Ecological Modelling*, 220(4), 589–594. <https://doi.org/10.1016/j.ecolmodel.2008.11.010>
- Velazco, S. J. E., Rose, M. B., de Andrade, A. F. A., Minoli, I., & Franklin, J. (2022). flexsdm: An R package for supporting a comprehensive and flexible species distribution modelling workflow. *Methods in Ecology and Evolution*, 13(8), 1661–1669. <https://doi.org/10.1111/2041-210X.13874>
- Wallace, B. C., Dahabreh, I. J., Trikalinos, T. A., Lau, J., Trow, P., & Schmid, C. H. (2012). Closing the gap between methodologists and end-users: R as a computational back-end. *Journal of Statistical Software*, 49, 1–15. <https://doi.org/10.18637/jss.v049.i05>
- Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., & Schewe, J. (2014). The inter-sectoral impact model intercomparison project (ISI-MIP): Project framework. *Proceedings of the National Academy of Sciences of the United States of America*, 111(9), 3228–3232. <https://doi.org/10.1073/pnas.1312330110>
- Wei, P., Lu, Z., & Song, J. (2015). Variable importance analysis: A comprehensive review. *Reliability Engineering & System Safety*, 142, 399–432. <https://doi.org/10.1016/j.ress.2015.05.018>
- Whitford, A. M., Shipley, B. R., & McGuire, J. L. (2024). The influence of the number and distribution of background points in presence-background species distribution models. *Ecological Modelling*, 488, 110604. <https://doi.org/10.1016/j.ecolmodel.2023.110604>
- Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32–35. [https://doi.org/10.1002/1097-0142\(1950\)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3](https://doi.org/10.1002/1097-0142(1950)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3)
- Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., Farooq, H., Herdean, A., Ariza, M., Scharn, R., Svantesson, S., Wengström, N., Zizka, V., & Antonelli, A. (2019). CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*, 10(5), 744–751. <https://doi.org/10.1111/2041-210X.13152>

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**File S1.** Worldwide suitable habitat of *Thunnus albacares*.

**File S2.** Distribution of the loggerhead sea turtle in the Mediterranean Sea.

**File S3.** Potential risk areas of *Siganus luridus* in the Greek Seas.

**How to cite this article:** Mestre-Tomás, J., Fuster-Alonso, A., Bellido, J. M., & Coll, M. (2026). GLOSSA: A user-friendly R Shiny application for Bayesian machine learning analysis of marine species distribution. *Methods in Ecology and Evolution*, 17, 505–517. <https://doi.org/10.1111/2041-210x.70167>