Machine learning applied to global scale species distribution models (SDMs)

Alba Fuster-Alonso

afuster@icm.csic.es

Institute of Marine Sciences (ICM) - CSIC, Renewable Marine Resources Department, Barcelona, 08003

Jorge Mestre-Tomás

Institute of Marine Sciences (ICM) - CSIC, Renewable Marine Resources Department, Barcelona, 08003

Jose Carlos Baez

Spanish Institute of Oceanography (IEO) - CSIC, Oceanographic Center of Málaga, Fuengirola, 29640

Maria Grazia Pennino

Spanish Institute of Oceanography (IEO) - CSIC, Oceanographic Center of Madrid, C. Del Corazón de María, 8, 28002, Madrid

Xavier Barber

Operations Research Center, Miguel Hernández University (UMH)

Jose María Bellido

Spanish Institute of Oceanography (IEO) - CSIC, San Pedro del Pinatar, Murcia

David Conesa

Department of Statistics and Operations Research (VaBar), University of Valencia (UV), Valencia

Antonio López-Quílez

Department of Statistics and Operations Research (VaBar), University of Valencia (UV), Valencia

Jeroen Steenbeek

Ecopath International Initiative

Villy Christensen

Institute of the Oceans and Fisheries, University of British Columbia

Marta Coll

Institute of Marine Sciences (ICM) - CSIC, Renewable Marine Resources Department, Barcelona, 08003

Article

Keywords:

Posted Date: May 24th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4411399/v1

Additional Declarations: No competing interests reported.

Machine learning applied to global scale species distribution models (SDMs)

Alba Fuster-Alonso[1,7*], Jorge Mestre-Tomás[1], Jose Carlos Baez[2,3], Maria Grazia Pennino[4], Xavier Barber[5], Jose María Bellido[6], David Conesa[7], Antonio López-Quílez[7], Jeroen Steenbeek[8], Villy Christensen[9,8], and Marta Coll[1,8]

[1]Institute of Marine Sciences (ICM) - CSIC, Renewable Marine Resources Department, Barcelona, 08003, Spain.
[2]Spanish Institute of Oceanography (IEO) - CSIC, Oceanographic Center of Málaga, Fuengirola, 29640, Spain.
[3]Ibero-American Institute for Sustainable Development (IIDS), Autonomous University of Chile, Av. Alemania 1090, Temuco 4810101, Araucanía Region, Chile.
[4]Spanish Institute of Oceanography (IEO) - CSIC, Oceanographic Center of Madrid, C. Del Corazón de María, 8, 28002, Madrid, Spain.
[5]Operations Research Center, Miguel Hernández University (UMH), Spain.
[6]Spanish Institute of Oceanography (IEO) - CSIC, San Pedro del Pinatar, Murcia, Spain.
[7]Department of Statistics and Operations Research (VaBar), University of Valencia (UV), Valencia, Spain.
[8]Ecopath International Initiative (EII), Spain.
[9]Institute of the Oceans and Fisheries, University of British Columbia, Canada.
[*]afuster@icm.csic.es

ABSTRACT

Species Distribution Models (SDMs) have been widely applied in ecology to analyze the historical and future patterns of marine species' distributions. With the increasing impact of climate change in recent decades, understanding potential shifts in species distributions has become a crucial challenge. Research on alterations in spatial and temporal distributions has revealed an increasing focus on developing different statistical approaches for global-scale and long-term forecasts. One promising approach is Bayesian Additive Regression Trees (BART), a non-parametric machine learning tool based on a sum-of-trees model that could be useful for addressing ecological problems. The goal of this study is to apply BART on a global scale and use it to estimate and predict possible present and future habitats of marine species under different climate change scenarios. Here we show an application of BART focused on the functional group of marine turtles, analyzing their historical and future distributions both individually and as a taxonomic group, their relationship with environmental variables, and BART's capacity to predict long-term distributions at global scales. Furthermore, to assess the capabilities of BART, we conduct a simulation study under two distinct scenarios: 1) simulating a hypothetical cosmopolitan species distribution and 2) simulating a hypothetical persistent species distribution. Our results show that BART is a promising approach to predict the potential distribution of our target species, as well as their relationship with key environmental variables, on a global scale.

Introduction

The impact of climate change on marine ecosystems has increasingly been recognized as a global-scale phenomenon, with numerous studies highlighting its effects on a worldwide basis[1–4]. As environmental conditions continue to change, marine species must adapt and potentially shift their distributions to areas with more suitable conditions for their survival and reproduction[1,5–11]. Therefore, understanding the present spatio-temporal distribution of marine species and accurately predicting their future is a critical challenge in the current context of global warming[12,13].

In response to the global impacts of climate change, macro-ecological approaches have gained importance in recent decades[14–21]. These approaches provide a broader perspective by focusing on large-scale patterns and processes, allowing for the prediction of species distribution and abundance at regional and global scales[22]. Therefore, this global perspective is essential for the evaluation of climate change and biodiversity[12,13,23]. The results of large-scale predictions can contribute to the development of effective management strategies with global policy objectives, enabling decision-makers to prioritize conservation efforts, implement sustainable practices, and mitigate the impacts of climate change on marine ecosystems[14,24].

Due to the interest in estimating and predicting the spatial-temporal distribution of marine species, tools such as Species Distribution Models (SDMs) have become fundamental in Ecology[25,26]. SDMs link information about the presence/absence or abundance of species to key environmental drivers to predict where and how a species is likely to be present in unsampled areas

or time periods[27]. This explanatory and predictive capacity makes SDMs valuable for various applications across multiple disciplines, allowing researchers to explore and address a wide range of ecological questions[28].

In the field of Ecology, SDMs have been employed in diverse contexts and have provided insights into species distributions patterns, species-environment relationships, and potential habitat suitability[29–33]. To implement SDMs, researchers can choose from a variety of approaches and software tools that facilitate the inference and forecasting processes[5]. One common form of analysis applied to SDMs involves classification or regression models[34]. These models use statistical algorithms to classify geographic areas into suitable and unsuitable habitats based on environmental conditions, allowing for the identification of areas with high likelihoods of species occurrence[35].

A promising and innovative alternative to traditional SDMs regression tree methods is the Bayesian Additive Regression Trees (BART) approach[36,37]. BART is a non-parametric Bayesian regression approach that builds upon a sum-of-trees model, and is fundamentally an additive model with multivariate components[37]. This methodology offers several notable advantages over conventional SDMs, making it an appealing choice for ecological research.

One of the key advantages of BART over traditional classification tree methods is the incorporation of prior distributions, which helps mitigate the issue of over-fitting commonly associated with regression trees[38]. Therefore, with the use of prior distributions, BART can strike a balance between fitting the data and maintaining flexibility to accurately predict species distribution in unsampled areas or future time periods. This feature enhances the robustness and reliability of the model's predictions[39].

SDMs, such as BART, may be a useful tool for informing other models, such as Marine Ecosystem Models (MEMs)[40]. Mechanistic models usually rely on parameters specifications that are either originated from raw data, estimated from data using statistical methods or elicited from expert input[41]. Some of these inputs pretend to represent the actual native ranges of species, habitat suitability and their functional responses to key environmental variables[42]. But, the uncertainty in these parameters can compromise the precision and validation of MEM results. While there have been efforts to refine these responses to environmental drivers[14], there is still substantial room for improvement to account for the spatial heterogeneity of species and functional groups. For instance,[42] highlight how the integration of SDMs and MEMs, using the outputs of SDMs as inputs in a MEM, can improve the model's results.

Overall, BART has been used in the context of SDMs, but only on a local/regional scale[43–50]. Furthermore, there are very few tools for modeling at a global scale that allow the user to update data or include different drivers. For this reason, our main goal is to apply BART on a global scale for the estimation and prediction of spatial-temporal distributions of different marine species and their relationship with environmental variables. Our hypothesis is that BART may be a powerful approach to predict historical and future scenarios about the distribution of target species and functional groups, as well as their relationship with key environmental variables, on a global scale.

In order to test our hypothesis, we conducted a case study on the functional group of marine turtles and a simulation study to assess the applicability of BART. This group includes all seven existing species of marine turtles, which are distributed very differently in the marine environment[51,52]. Moreover, ongoing research discuss how marine turtles face imminent threats to their survival in the wake of climate change[53–55]. This information, combined with their different distributions patterns, makes marine turtles an ideal functional group for testing the effects of climate change on a global scale. The study that we present here applies the BART method to obtain the native ranges, potential habitat, relations with environmental variables and projections of distribution under different future scenarios of climate change using outputs from Earth System Models (ESM) freely available through the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP)[56,57] and Fish-MIP initiative (`https://fish-mip.github.io/`). Due to the uncertainty related to the predictions of climate change effects, we considered two different ESM set of drivers for our case of study: GFDL-ESM4 and IPSL-CM6A-LR.

Although a real case study can shed light on the predictive capability of a model, the validation associated with such study is contingent upon errors in observations, as we lack knowledge of the true current or future distribution of the species. For this reason, a simulation protocol has also been developed, allowing us to investigate the results of a hypothetical species. Then, through various random samplings of simulations, we can obtain presence and absence data in order to fit the model and assess the model's predictive capacity[58]. In this study we assess two different simulation scenarios: one considers a hypothetical species that is spread over the entire domain (what we call a 'cosmopolitan' species), while the second scenario considers a species that remains permanently in a specific area (we call it a 'persistent' species).

Results

Overview of global BART analysis workflow

Our study is divided into two sections: 1) a case of study, where we present the results obtained from BART for the marine turtle functional group; and 2) the simulation study, where we illustrate the performance of BART in a presence/absence simulation and modeling framework (see Figure S1 of supplementary material).

Therefore, section 1) is based on applying BART on a global scale using the marine turtle functional group as a case of study. Hence, we propose two different models: the native range model and the suitable habitat model. The main difference between these two models is that for the native range, we include the latitude and longitude of observations as covariates in the model, while the suitable habitat model is only based on environmental covariates. The inputs used are georeferenced occurrence data from GBIF and historical, past, and future projections of environmental variables from ISIMIP. The common output of the native range and suitable habitat models is a unified map representing the historical spatial distribution from 1950-2014. Subsequently, using the suitable habitat model, we project the distribution for each year from 1950 to 2100, generating a stack with the spatial distribution for each year. Results are validated using a k-folds cross validation and real observations.

Section 2) aims to assess the capacity of BART to project in space and time the distribution of different species. For this purpose, we simulate two different scenarios of probability according to the behavior of a species: first, a cosmopolitan species, where the species is dispersed over the entire domain, and second, a persistent species, where we observe a concentrated spatial distribution. Then, we perform 30 different random samplings to obtain presence/absence data for fitting 30 different models and predicting using BART. This allows us to obtain error measures of the predicted spatio-temporal distribution with respect to the simulated groundthruth.

Case of study

Species predictions

Here, we present figures depicting the present (1950-2014) and future (2015-2100) predictions of two species and their respective response functions to environmental variables. The two chosen species are the Australian flatback sea turtle *Natator depressus* and the leatherback sea turtle *Dermochelys coriacea*, both with very different spatial distributions. The first species is mainly distributed along the Australian coast, while the second species is widely distributed throughout the world[51,52]. The results for the remaining species can be found in supplementary material (refer to section 2 Species predictions).

Figure 1 shows the historical predicted probability (from 1950 to 2014) in the spatial domain for the two species. It is worth noting that, since we are working in a Bayesian framework, each pixel (1° x 1°) contains a posterior predictive distribution. Hence, we can compute various statistics to represent the species distribution. In this case, the mean of the posterior predictive distribution is shown, along with the uncertainty captured by subtracting the 2.5 % and 97.5 % percentile quantiles. Figure 1 confirms how *Natator depressus* is a species primarily distributed in the Indo-West Pacific, with higher probabilities of presence along the north coast of Australia, while *Dermochelys coriacea* is distributed along the tropical and temperate zones globally, and it has a cosmopolitan distribution excluding polar zones.

On the other hand, it is possible to differentiate between two different models: 1) the native range and 2) the suitable habitat of both species (Figure 1). In Figure 1 it can be observed that the native ranges have a much narrower distribution than the suitable habitat for both species. This is reasonable, considering that the native ranges represent the historical observed distributions of the species and are constrained by adding the coordinates in the model. Furthermore, the results of both models are obtained for each of the Earth System Models (ESMs), resulting in very similar probability maps for both species historically.

Regarding future projections, Figure 2 provides an insight into the predicted probability for the last 10 years of prediction, excluding 2100 (from 2090 to 2099). It should be noted that for future projections, only potential habitats are obtained. Then, similarly to what was mentioned above, the mean of the posterior predictive distribution from 2090 to 2099 is shown in Figure 2 for two different climate scenarios SSP126 and SSP585. Additionally, to observe changes in the distribution over time, the difference between the historical prediction for suitable habitats shown in Figure 1 and the future projection presented in Figure 2 is represented as suitable habitat change (Figure 2).

Specifically, for the species *Natator depressus*, our results project a reduction in potential habitat along the northern coast of Australia and gains in the northern hemisphere 2. This decrease becomes more evident under the SSP585 scenario and with GFDL-ESM4, while the gains are more pronounced with IPSL-CM6-LR. In the case of the species *Dermochelys coriacea*, a loss of potential habitat is observed in the northern hemisphere with GFDL-ESM4, while gains are observed in the south (Figure 2). With IPSL-CM6A-LR, the gains are more pronounced specially in the north, while with GFDL-ESM4 the losses are more pronounced in the Atlantic Ocean. Moreover, for the SSP126 scenario, the losses and gains appear to be less distinct than for the SSP585 scenario in both species (Figure 2).

In Figure 3 we can observe, for the two mentioned species, the contribution of environmental variables and the nonlinear relationships of the response variable with each of the environmental variables. For both species, the two variables that contribute the most to the model are bathymetry and sea surface temperature (Figure 3). However, it is worth noting that all variables in the model have a similar contribution (Figure 3). In the nonlinear relationships with the response variable, it can be seen that for bathymetry, both species have their optima at low bathymetric values (Figure 3). Whereas for SST, the behavior is sigmoidal, their distribution increases until reaching a maximum and then starts to decrease (Figure 3).

In the same way, the native ranges and suitable habitats for other five species are presented in section 2 Species predictions

(Figures S2, S3 and S4) of the supplementary material. Notably, *Caretta caretta* appears to have a distribution concentrated in sub-tropical latitudes and along the west coast of Africa (Figure S2 in the supplementary material). *Lepidochelys olivacea* seems to have a more southern distribution compared to *Caretta caretta*, along with *Eretmochelys imbricata* and *Chelonia mydas* (Figure S2, S3 and S4 of the supplementary material). In contrast, *Lepidochelys kempii* exhibits a more confined distribution primarily in the Atlantic Ocean and Europe.

Furthermore, in section 2, Figures S4, S5, and S6 of the supplementary material, we can find information on the suitable habitat changes for these five species, as well as their contributions and relationships with environmental variables (Figure S7, S8, and S9 in the supplementary material). Most species seem to experience significant losses around the tropical zones, except for *L. kempii*, where losses are focused on the Atlantic North Coast. The spatial results obtained from both ESMs seem to agree in general terms, with the SSP585 climate scenario showing much more pronounced losses and gains for all species.

In terms of changes over time, we have projected the suitable or potential habitats of all species from 1950 to 2100. Figure 5 illustrates the suitable habitat's mean probability for each projected year, enabling us to observe how the mean probability of finding a specific species globally changes over time. Depending on the species, we observe an increase or decrease in the mean probability of suitable habitat. Specifically, for *C. caretta*, and *L. olivacea*, the temporal trend appears to decrease, indicating less suitable habitat will be available for them in the future. *L. kempii* presents a more stable temporal trend, with a smooth decrease for both GFDL-ESM4 climate change scenarios and IPSL-CM6-CR SSP126 scenario results, and a sigmoid function for the IPSL-CM6-LR SSP585 scenario. Conversely, for *E. imbricata* and *C. mydas*, the potential habitat seems to increase in both ESM configurations, especially for the SSP585 scenario. With respect to *Natator depressus*, there is an increase during the historic period, but it seems to stabilize for the future scenarios under GFDL and IPSL 126. Moreover, it's worth mentioning that for *D. coriacea*, we obtain different results depending on the ESM. GFDL-ESM4 shows a decrease in the temporal trend, while IPSL-CM6-LR shows an increase, especially for the SSP585 scenario (Figure 5).

Furthermore, in Table 1, we have quantified the percentage of change between the historic suitable habitat (1950-2014) and the last ten years of the future suitable habitat (from 2090 to 2099). The percentages displayed in Table 1 show that depending on the species, the available suitable habitat will either increase or decrease. It's worth noting that the percentages obtained for some species, such as *D. coriacea* or *L. kempii*, vary depending on the Earth System Model (ESM) and the climate change scenario employed in the model.

Functional group predictions

Results of the whole functional group obtained for the seven marine turtle species considered in this study are presented (Figure 4) in terms of the native ranges and potential habitats for all marine turtles. Both outcomes are obtained by calculating the median of all native ranges and suitable habitats for each species (Figure 4).

We can observe how the highest probabilities for native ranges of the marine turtles functional group are located on the east coast of the USA and the northern coast of Australia. Meanwhile, for suitable habitats, it appears that coastal areas generally have optimal environmental conditions for the functional group. Similarly, for both models (native ranges and suitable habitats), the lowest probabilities are found around the poles. Likewise, it is worth mentioning that for both GFDL-ESM4 and IPSL-CM6-LR configurations, the width of the intervals is not too large, especially for the native ranges (Figure 4).

On the other hand, it is possible to observe the median of the year 2090-2099 for the future projections of potential habitats based on the two climate scenarios (SSP126 and SSP585). For both scenarios, we observe a significant loss of potential habitat on the east coast of North America (Figure 4). In contrast, the southern hemisphere appears to acquire optimal conditions for marine turtles over time. Furthermore, the losses and gains are more pronounced in the SSP585 climate scenario compared to SSP126 (Figure 4). Based on the results from both Earth System Models (ESMs), IPSL-CM6-LR exhibits more pronounced changes in suitable habitat in the Northern Hemisphere compared to GFDL-ESM4.

Finally, in Figure 5 and Table 1, we observe an overall increase in the mean probability of suitable habitat of marine turtles, especially for the SSP585 scenario of IPSL-CM6-LR, where the percentage of change is 43.98 % higher in the last ten years compared to the historical suitable habitat. With GFDL-ESM4, the changes in the probability are less pronounced, with a decrease of -0.80 % for the conservative climate scenario (SSP126) and an increase of 3.5 % for the pessimistic climate scenario (SSP585).

Validation

In Table 2, we present the results of various metrics obtained through cross-validation. Table 2 displays the median values of four different statistics calculated for each Earth System Model (ESM) and species. Results show that across all metrics, the model performance is notably strong, with most values approaching 1. While there is a slight performance difference favoring GFDL-ESM4 ESM over IPSL-CM6A-LR, this discrepancy is small.

On the other hand, to validate our future projections, we compared maps projected for the period 2015 to 2023 with actual observations obtained from GBIF during the same time frame, since the hindcasting period of our models spans from 1950

to 2014. This comparison allowed us to calculate the sensitivity, which measures the proportion of true positives correctly identified, for each climate scenario and ESM.

Our results indicate a notable level of sensitivity, with GFDL-ESM4 standing out as the ESM that provides the highest sensitivity values, reaching 0.77 and 0.79 for the SSP126 and SSP585 scenarios, respectively. Additionally, the IPSL-CM6A-LR model also demonstrates strong performance, achieving sensitivity values of 0.75 for both SSP126 and SSP585 climate scenarios.

Simulation

Regarding the simulation framework, both simulation scenarios of the occurrence data of a hypothetical cosmopolitan and persistent species were developed by considering several effects: 1) a spatial-temporal effect, 2) a bathymetric effect, 3) a temperature effect, and 4) a temporal trend. Therefore, a series of parameters were set for all the terms in equations 2 and 3.

Cosmopolitan species

Our hypothetical field of study is a regular grid (10x10) over a 20-year time window. We started simulating the correlated spatial effect by setting the values of the range $r = 3.5$, variance $\sigma = 1$, and temporal correlation $\rho = 0.7$. Then, we constructed a bathymetry effect (constant over time) in a range from 0 to 800 meters using the following formula $100 \cdot \log(xy+1)$, where x and y are the coordinates, reflecting that the closer to the axis, the lower the bathymetry. As the response function of species to bathymetry is usually non-linear, we set up a polynomial of degree two to achieve a quadratic relationship ($\beta_{1X_1(s)} = -1.5$, and $\beta_{2X_1(s)} = -1.1$) between them. To simulate temperature, we considered only the y axis, using the formula $\sqrt{y+1}+10$. Additionally, since temperature is a dynamic variable, we added 0.5 for each year to represent an increase over time. Concerning the temporal trend, we simulated a vector of values from an autoregressive model of order 1 (AR1), where $\rho_t = 0.7$.

Then, all the terms of the predictor in Equation 2 were summed and transformed to the probability scale (ranging from 0 to 1) using the inverse of the logit link function. Therefore, we obtained the probability of presence across space and time. In Figure 6 a), we provide an example of a simulated probability field, illustrating how our hypothetical species exhibits a cosmopolitan behavior across the simulated study field and time. After simulating the probabilities, we performed 50 samplings to obtain the presence/absence data for fitting the BART model (1). Then we simulate from a Bernouilli distribution to distinguish between presence and absence according to the simulated probability. It is noteworthy that, for each sampling, we selected a total of 50 observations.

Once we had the presence/absence data, we were able to fit the models and then make predictions across both spatial and temporal dimensions. In the repository we can observe all predictions across space and time for each replica. Note that we excluded the data from years 18, 19, and 20 during the fitting process. This exclusion allowed for the subsequent projection of the entire distribution into the future.

Figure 6 c) presents the median, along with the 0.025 and 0.975 quantiles, of the validation measures (sensitivity, specificity, and accuracy) across all replications. Examining sensitivity and specificity, we observe that at the beginning of the study period, sensitivity is notably high, while specificity is lower, particularly in the second year. This pattern shifts around year 7, where specificity increases, leading to a decrease in sensitivity. However, when we assess accuracy, it consistently remains close to one throughout the entire period, except for a slight dip in year 2.

Persistent species

Regarding the simulation results for the persistent species scenario, it essentially follows the same process as the cosmopolitan species scenario but it is based on Equation 3. Specifically, we adjusted the values of several parameters. The range and variance of the spatial-temporal effect were 5.6 and 1, respectively, with an autoregressive coefficient of $\rho_U = 0.1$ for the spatial-temporal effect and $\rho_t = 0.7$ for the temporal trend. Concerning bathymetry and temperature, the formulation was the same as mentioned above, but the values for the fixed coefficients were $\beta_{1X_1(s)} = 7.5$ and $\beta_{2X_2(s,t)} = -0.8$, respectively.

After simulating all the terms, we transformed the predictions into the probability range using the inverse of the logit link function. This process enables us to generate probability maps. In Figure 6 b), we present the simulated probability distribution, demonstrating clear persistence across space. Utilizing these maps, we performed a random sampling of presence/absence data, employing the same methodology as utilized in the cosmopolitan species simulation. Then, we used these presence/absence data to fit the BART models and perform posterior predictions over space and time. All predictions are display at the repository for each replication (note that years 18, 19, and 20 were excluded from the fitting, making the entire map a prediction).

Finally, Figure 6 d) presents the median, along with the 0.025 and 0.975 quantiles, of the validation measures (sensitivity, specificity, and accuracy) across all replications. In this scenario, all measures were high, especially for those years where the prediction was limited to unsampled locations and not in time. It is worth mentioning the decline in all measures, particularly sensitivity, in the last years, which are the ones used for projecting the entire distribution. However, the accuracy remained quite high even for the last years, consistently exceeding 0.75.

Discussion

Climate change stands as a significant threat to many marine species[59]. Among them, marine turtles are commonly considered susceptible to the impacts of climate change due to the important role of temperature in their life cycle[59,60]. In 2009, the IUCN Red List categorized most of the marine turtles species as vulnerable, endangered, or critically endangered[61–67]. Therefore, preserving marine turtle populations under new climate change conditions demands global actions to reduce its impact and bolster turtle resilience[59].

Our study successfully tested the capability of global SDMs to investigate the global distribution of marine turtles and project future potential habitats under different scenarios of climate change. In fact, our results highlight the heterogeneity distribution of such a diverse group and shows the divergence in future projections according to specific species. While some species are likely to face important challenges in the future and may experience declines in available suitable habitats, others are projected to expand their potential habitats. These results highlight the need to tailored management actions to specific species and regions.

Therefore, based on the future suitable habitats obtained in our research, future studies could focus on analyzing the ability of different species of marine turtles to reach this new potential habitat[68]. It's important to consider that even if a habitat is suitable in terms of environmental conditions, factors such as proximity or other non-environmental drivers may prevent the species from reaching that new potential habitat[69–71].

Model results also confirm the important role of sea temperature to drive species distributions, and specifically the distribution of marine turtles. The increase in sea surface temperatures due to global warming can have various impacts on marine turtles, affecting their habitats, food sources, reproductive patterns, and overall survival[72]. For example, rising sea levels and increased temperatures can lead to the loss of nesting beaches for marine turtles. Coastal erosion can destroy nesting sites, making it challenging for turtles to find suitable areas to lay their eggs[73]. In the past decade, there has been an observed rise in sporadic nesting occurrences of sea turtles[74], notably linked to unusual increases in Sea Surface Temperature (SST)[75]. This phenomenon raises significant concerns regarding the conservation and management of marine turtle populations, prompting an in-depth exploration of the implications and potential causes behind these irregular nesting events. The correlation between sporadic nesting and atypical SST elevations suggests a plausible connection between marine turtle behavior and environmental fluctuations. Elevated SST levels could potentially influence the nesting behavior of sea turtles, prompting shifts in their traditional nesting patterns and leading to sporadic nesting events in atypical locations. Another important mechanisms linking sea warming and marine turtles distributions is sex determination since elevated temperatures can cause a bias in sex ratios of populations due to alterations in ecological sex determination (for example[76]). The temperature during the incubation period of turtle eggs determines the sex of the hatchlings, and higher temperatures can skew the sex ratio, leading to an imbalance in male and female populations.

In the current context of conservation and management of marine turtle, another important threat to these species today is bycatch in fishing gears[77,78]. Our results can be used to identify current hotspots of presence of marine turtles and be used to minimize fishing practices in those areas with higher risk of by catch. Consequently, an expansion of suitable areas for marine turtles to specific areas should be done minimizing the risk of interactions with fishing gear. This brings to light the intricate balance between conservation efforts and the unintended consequences that may arise from increased suitable habitats intersecting with fishing activities.

Overall, the forecasts models such as the ones presented in the current study could help to inform conservation efforts of marine turtles, and to minimize incidental capture in fishing gear, potentially through the establishment of protected marine areas. In fact,[79] proposed the use of Regional Marine Turtle Management Units (RMUs) as a framework for prioritizing conservation across multiple scales of sea turtles. However, this RMU overview could be completely change due to climate change. While expanding suitable areas for marine turtles is crucial for their conservation, it necessitates a comprehensive understanding of the intricate interplay between habitat availability, fishing activities, and the broader ecosystem dynamics. Integrating these complexities into conservation models and strategies is imperative to ensure the long-term survival of marine turtle populations.

While we acknowledge that BART is a useful tool for solving ecological issues, our study has some limitations, too. One main concern is the uncertainty linked to the data we used. We relied on the GBIF database, which may have a large amount of uncertainty within its observations. Despite that, we have followed standard procedures to clean and improve the data quality. Similarly, environmental drivers could involve significant uncertainty, particularly in future projections. To partially account for some of the uncertainly, we utilized two different Earth System Model (ESM) outputs, ensuring that we do not rely only on a single set of drivers. Indeed, for some species of marine turtles GFDL-ESM4 and IPSL-CM6-LR lead to different results in terms of future potential habitats. This raises the need of considering the uncertainty related to ESMs when we use environmental drivers as inputs, which has been already observed in previous studies.[80].

Another limitation is on how we generated pseudo-absences. Since we lack absence data, we had to create pseudo-absences. We have tried to make this in a way that does not heavily impact the results, using random generation and equal amounts

of absences and presences. Furthermore, to address these concerns, we conducted a simulation study to better assess the performance of BART. This helped us to have a more reliable understanding of the tool's capabilities, particularly in combination with a rigorous case study.

Despite the valuable utility of SDMs in estimating distribution changes over time, there is an ongoing need to enhance these models[81]. Combining complementary models can yield better results, providing a more comprehensive understanding of species behavior[42]. For example, it is important to note that our predictions do not account for changes in ecological relationships, such as prey-predator dynamics or other crucial factors like fisheries mortality. Changes in sea temperatures can alter the distribution and abundance of marine turtle prey, such as jellyfish, crustaceans, and sea grasses. This can impact the feeding habits of turtles and affect their growth and health but our results can only capture this implicitly. In addition, marine turtles rely on coral reefs for food and shelter. Increased sea temperatures can lead to coral bleaching events, which reduce the quality and availability of habitat for turtles and their prey. This is the case of *Eretmochelys imbricata*, which exhibits strong associations with coral reef ecosystems, feeding predominantly on sponges[82]. Hence, their distribution might be more closely linked to food availability than to other environmental factors. This underscores the complexity of factors governing the distribution and habitat preferences of marine turtles, suggesting that conservation strategies should consider specific dietary needs and habitat dependencies of individual species. Hence, it's relevant to integrate models, such as SDM and MEMs, to account for these additional relationships[42,83]. As such, our findings have significant potential value for parameterizing MEMs in order to improve the overall accuracy of predicted spatial-temporal species distributions of marine species, such as marine turtles, globally.

Due to the potential use of global SDMs, it is crucial to continue developing tools that enable us to assess the past, present, and future status of marine species, such as marine turtles[84]. In this context, the results obtained in this study highlight the capability of machine learning models like BART to predict changes in the current and future habitats of marine species, making these models a valuable approach for assessing management and conservation efforts[36]. Our study shows how BART can be a reliable tool for predicting both current and future habitats of marine turtles on a global scale. We anticipate significant developments in both current and future applications of global SDMs approaches.

Methods

This section includes the methodological details of our case of study and the simulation study. The entire analysis was conducted using `RStudio` software[85] and all code is available in a GitHub repository code.

Case study

The focus of this study is to estimate and predict the probability of presence over space and time for the marine turtles functional group (refer to Table S1 of supplementary material) for biological information about the species). In order to achieve our goal, a series of steps were carried out. First, presence data of each marine turtle species and environmental variables potentially driving their distribution were extracted and cleaned. Then, the BART model was implemented using the collected data of individual species. Last, the different results were validated and compared.

Extraction and cleaning of the data

Presence-only data of a species are one of the most widely used datasets in the context of SDMs due to their accessibility at different scales[86–88]. For our study, which aims to predict using a global perspective, we obtained data from the Global Biodiversity Information Facility (GBIF) using the `rgbif` package in R[89,90]. All the DOIs with the downloaded raw data for each species are available in the supplementary material section 1 Marine turtles information and study workflow.

The presence data for the seven species of marine turtles currently occurring in the marine environment were processed by eliminating repeated and terrestrial locations. We excluded terrestrial locations because we were only interested in predicting distribution in the oceans. However, it's worth mentioning that female marine turtles spend part of their life cycle on land. We also employed the `CoordinateCleaner` package in `R` to remove presences with significant uncertainty[91]. BART requires both presence and absence data to operate correctly. Due to the lack of available absence data for statistical modeling using a Bernouilli distribution, we randomly generated pseudo-absences equal to the number of presences for each species[92].

Furthermore, we incorporated global spatial time series of varying environmental conditions obtained from The Inter-Sectoral Impact Model Intercomparison Project (ISIMIP)[56,57] and Fish-MIP initiative (`https://fish-mip.github.io/`). We drove our model using outputs from two different Earth System Models (ESMs) of the Coupled Model Intercomparison Project Phase 6 (CMIP6): GFDL-ESM4 and IPSL-CM6A-LR[93]. These models were built under prescribed scenarios for historic (1950s to 2014) and future (2015 to 2100) time periods[93]. Moreover, for both ESMs, we used two different Shared Socio-economic Pathway (SSP) climate scenarios: a more conservative one, SSP126, and a more pesimistic one, SSP585.

Among the various ESM variables available under ISIMIP, we selected SST (Sea Surface Temperature in degree Celsius), SSS (Sea Surface Salinity in PSU), LPHY (mole content of diatoms), O2 (mole concentration of dissolved molecular oxygen),

DPHY (mole content of diazotrophs), and SPHY (mole content of picophytoplankton). It is worth noting that the last two variables were only available for GFDL-ESM4. Additionally, we included bathymetry as a static variable for all analysis. To prepare the variables for predictions, we standardized all the environmental variables. We carry out standardization by calculating the mean and standard deviation of the historical data, and then subtracting that mean and dividing by the standard deviation each historical and future layer. However, for obtaining the functional responses, we utilized the non-standardized environmental variables to get the response curve in the real scale.

Modeling approach: BART

Regarding statistical modeling, BART models are based on a sum of regression trees.[38] provide an illustration of the formulation and representation of a single tree model, offering a comprehensive insight into the formulation underlying these models. Essentially, regression trees are algorithms meant for modeling and prediction in machine learning[94,95]. The formulation of a regression tree g could be defined in terms of two components: (1) T a set of decision rules and nodes, and (2) $M = \mu_1, ..., \mu_b$ a set of parameter values associated to each terminal node of T. Then, $g(X;T,M)$ is the function that assigns a value to the b parameters ($M = \mu_1, ..., \mu_b$) according to the covariates (X) added to the tree model.

The main problem with regression trees is that they tend to overfit, as they can split the space until they get one parameter per datum[38]. This overfitting may considerably bias predictions. To address this problem, approaches such as BART have been developed. Through the ensemble of decision trees and regularization using *a priori* distributions in the Bayesian context, BART methods reduce the overfitting without performing a cross validation for model parametrization[36,37]. In our study, we adopted the default prior distribution for BART, as the literature praises its strong performance with default parameters[38].

In order to model the presences/pseudo-absences data, the statistical model applied in this work was as follows:

$$\begin{aligned} Y_i &\sim Ber(\pi_i), \quad i = 1, ..., n, \\ \phi^{-1}(\pi_i) &= \sum_{j}^{m} g_j(\mathbf{X}; T_j, M_j), \end{aligned} \tag{1}$$

where Y_i represents the presence/pseudo-absence of species for observation i; π_i is the parameter of interest linked to the predictor by a link function; ϕ^{-1} denotes the standard normal cdf (probit link function); g_j is the $j-$th ($j = 1, \ldots, m$) tree of the form $g_j(\mathbf{X}; T_j, M_j)$, where m is the total number of trees, X is a vector of multiple covariates, T_j represents a binary tree structure consisting of a set of interior node decision rules and a set of terminal nodes, and $M_j = \{\mu_j, ..., \mu_{jb}\}$ denotes a set of parameter values associated with each of the b_j terminal nodes of T_j.

Furthermore, a differentiation was established between two types of models: 1) native ranges, which refer to the areas where the species is known to have occurred historically and it is likely currently present; and 2) suitable habitats, which are understood as potential habitats where conditions are suitable for the target species. The reason for this differentiation is that certain areas may be considered potential habitats, but due to other factors such as geographic barriers or physical distances, the species has never been observed or is not present in those areas. Therefore, the main difference when modeling these two distributions is that for suitable habitats (2) the $\mathbf{X}$ vector of covariates only includes environmental variables, while for native ranges (1) the $\mathbf{X}$ vector of covariates also incorporates the coordinates of historical observations to account for realistic or plausible spatial variability in the model.

After inferring the model parameters, space and time predictions were carried out for the historical period (1950-2014) and for future projections (2015-2100) using two different ESMs (GFDL-ESM4 and IPSL-CM6A-LR) and climate change scenarios (SSP126 and SSP585). Hence, we generated the historical (1950-2014) and future (2015-2100) projections by year using the suitable habitat model. Consequently, the predictions provide insights into the future areas where environmental conditions will be optimal for the seven marine turtle species. In contrast, we generated two different aggregated historical distributions in space: one using the native range model and the other using the suitable habitat model. For these aggregated historical spatial distributions, we employed the mean of the environmental variables (see Figure S1 in the supplementary material).

Validation and comparison of predictions

For the validation of the models, we calculate several measures, distinguishing between two types of validations: an internal validation using a k-folds cross-validation method, and an external validation using new species distributions from GBIF that were not included in the models. Therefore, for internal validation, we applied the k-folds method to assess the performance of our model in the historical period. For external validation, we calculated these measures by comparing future projections of several years with actual observations that were not used in the fitting process.

For the internal validation, we divided the data in k = 10 subsets to test BART's predictive capacity. Therefore, we obtained a total of 10 different replicas. To analyze the results, we calculated error measures such as sensitivity, specificity, accuracy and F_1 score (see section 3 Error measures of supplementary material). All measures calculated are based on the estimations of

true positives, true negatives, false positives, and false negatives[96]. Furthermore, as our forecasting extended from 2015 to 2100, we were able to compare the model predictions from 2015 to 2023 with observed data from GBIF to evaluate the model's performance in projecting the distribution. We compared the observed data with the predicted probability values and calculated error metrics such as sensitivity. These metrics are essential when dealing with presence and absence data. Sensitivity evaluates the model's ability to correctly identify true positives (actual presences), specificity assesses its ability to correctly identify true negatives (actual absences), and accuracy measures the overall correctness of the model's predictions (see Section X of supplementary material).

Finally, we compared the historical predictions (1950-2014) of each species with the last 10 years (2090-2099), excluding 2100, to assess potential future habitat changes. We exclude the last year of the series due to the potential bias in the ESMs models for this final year of projection. To quantify these changes, we calculated the difference between the predicted historical distribution and the projections for the last ten years. This allowed us to estimate the extent of potential habitat change based on future climate change scenarios. Likewise, we extracted the mean probability for each projected year from 1950 to 2100, allowing us to assess changes over time in the mean probability of potential suitable habitat for each species and for the entire functional group.

Simulation

To strengthen the validity of our study, we conducted also a simulation designed to corroborate the predictive capabilities of our BART model across both spatial and temporal dimensions. This simulation involved two specific scenarios: 1) simulating a cosmopolitan species dispersed across the entire domain and 2) simulating a persistent species with consistent spatial and temporal patterns. Through this process, we aimed to provide further evidence supporting the reliability of our BART model in accurately predicting species distribution dynamics over space and time.

Simulation allowed us to replicate the behavior of a random variable in both space and time under controlled conditions, such as the probability of being presence of a species population. Therefore, the first consideration in simulation is understanding the factors influencing our variable of interest and developing a model that accounts for its nature. Typically, we lack information about the entire population and work with a sample instead. In such cases, we propose a model and make inferences about its parameters to obtain representative insights into the population. However, when simulating the entire population, we have knowledge of the parameters, enabling us to assess the accuracy of our model estimates[34].

For a more detail explanation and figures of the simulation process refer to the following vignettes.

Spatio-temporal occurrence simulation scenarios

The probability of the presence of a given target species is commonly influenced by various external factors (e.g., environmental, anthropogenic, etc.) as well as spatially structured biological processes (e.g., predation, competition, etc.). Moreover,[97] argue that all species, in one way or another, exhibit spatial structure. However, considering all the factors that affect the probability distribution of a target species in the modeling is practically impossible. For this reason, we have simplified the reality of our response variables taking into account two environmental variables (temperature and bathymetry) as essential drivers to explain distributions, a temporal dependence over the years, and a spatial-temporal effect related to species movement and dispersal. This selection was made considering that temperature and bathymetry typically play a key role in the spatial and temporal distribution of marine species[98]. Additionally,[99] discuss how incorporating a spatial effect can enhance prediction accuracy and mitigate the impact of variables not considered in the modeling. Hence, the simulation models for the different scenarios (cosmopolitan and persistent species) are formulated as follows:

1. **Cosmopolitan species**

$$\begin{aligned} Y(s,t) &\sim \text{Bernoulli}(\pi(s,t)), \\ \text{logit}(\pi(s,t)) &= \beta_0 + f_1(t) + f_2(\text{X}_1(s)) + \beta_1 X_2(s,t) + U(s,t), \end{aligned} \tag{2}$$

   where, the response $Y(s,t)$ represents the occurrence (presence/absence) of the cosmopolitan species at time t in the location s following a Bernoulli distribution with parameter $\pi(s,t)$; $\pi(s,t)$ is linked to the predictor by the logit link function; β_0 is the intercept; $f_1(t)$ stands for the temporal trend in the year t; $f_2(\cdot)$ is a deterministic function for the bathymetry ($X_1(s)$); and β_1 is the parameter associated to the temperature ($X_2(s,t)$). Lastly, $U(s,t)$ refers to the spatio-temporal structure.

2. **Persistent species**

$$\begin{aligned} Z(s,t) &\sim \text{Bernoulli}(\pi(s,t)), \\ \text{logit}(\pi(s,t)) &= \beta_0 + f_1(t) + \beta_1 X_1(s) + \beta_2 X_2(s,t) + U(s,t), \end{aligned} \tag{3}$$

where the response $Z(s,t)$ represents now the occurrence (presence/absence) of the persistent species at time t in the location s following a Bernoulli distribution with parameter $\pi(s,t)$; β_1 is a fixed effect for the bathymetry $X_1(s)$; and β_2 is the parameter associated to the temperature $X_2(s,t)$; and the remaining terms are those in 2.

With the model structure determined to start simulating the occurrence data of both scenarios ($Y(s,t)$ and $Z(s,t)$), some explanation is warranted to describe how to perform these simulations, in particular, how to deal with each one of the terms included in the predictors in (2) and (3). First, the spatio-temporal structure is simulated as a Gaussian Markov Random Field (GMRFs) correlated with an autoregressive $AR(1)$ model with parameter of autocorrelation ρ_{sp}[100]. Secondly, we simulate species-specific depth preferences. Particularly, for the bathymetry covariate, a range between 0-800 meters was simulated, with a non-linear effect for the cosmopolitan species scenario $f(X_1(s))$ and a linear effect for the persistent species scenario $\beta_1 X_1(s)$. Last, for the temporal trend $f(t)$, changes in the probability values over time are included by simulating a vector of values from an autoregressive model of order 1 with parameter of autocorrelation ρ_t.

Once the predictor terms have been obtained, the occurrence of both species ($Y(s,t)$ and $Z(s,t)$) has to be determined by using a Bernouilli distribution. Then, once we have obtained the simulated presence/absence data that will be fitted with the BART model (Equation 1), we need to perform several random samplings of each simulation. In this case, we conducted 50 samplings for each simulation scenario, allowing us to replicate the simulation and ensure the robustness of the analysis. For model validation, we calculated three commonly used measures already mentioned: sensitivity, specificity, and accuracy. To achieve this, we compared the estimated values (whether they indicate presence or absence) with the actual simulated presence or absence data. This process allowed us to determine how effectively our model assigns the correct status of presence or absence in relation to the simulated ground truth.

References

1. Worm, B. & Lotze, H. K. Marine biodiversity and climate change. In *Climate change*, 445–464 (Elsevier, 2021).
2. Gosling, S. N. *et al.* A review of recent developments in climate change science. part ii: The global-scale impacts of climate change. *Prog. Phys. Geogr.* **35**, 443–464 (2011).
3. Sippel, S., Meinshausen, N., Fischer, E. M., Székely, E. & Knutti, R. Climate change now detectable from any single day of weather at global scale. *Nat. climate change* **10**, 35–41 (2020).
4. Hodapp, D. *et al.* Climate change disrupts core habitats of marine species. *Glob. Chang. Biol.* (2023).
5. Miller, J. Species distribution modeling. *Geogr. Compass* **4**, 490–509 (2010).
6. Van der Putten, W. H., Macel, M. & Visser, M. E. Predicting species distribution and abundance responses to climate change: why it is essential to include biotic interactions across trophic levels. *Philos. Transactions Royal Soc. B: Biol. Sci.* **365**, 2025–2034 (2010).
7. Peterson, A. T. *et al.* Future projections for mexican faunas under global climate change scenarios. *Nature* **416**, 626–629 (2002).
8. Rosenzweig, C. *et al.* Attributing physical and biological impacts to anthropogenic climate change. *Nature* **453**, 353–357 (2008).
9. Cheung, W. W. *et al.* Projecting global marine biodiversity impacts under climate change scenarios. *Fish fisheries* **10**, 235–251 (2009).
10. Miller, D. D., Ota, Y., Sumaila, U. R., Cisneros-Montemayor, A. M. & Cheung, W. W. Adaptation strategies to climate change in marine systems. *Glob. Chang. Biol.* **24**, e1–e14 (2018).
11. Gordó-Vilaseca, C., Stephenson, F., Coll, M., Lavin, C. & Costello, M. J. Three decades of increasing fish biodiversity across the northeast atlantic and the arctic ocean. *Proc. Natl. Acad. Sci.* **120**, e2120869120 (2023).
12. Allyn, A. J. *et al.* Comparing and synthesizing quantitative distribution models and qualitative vulnerability assessments to project marine species distributions under climate change. *PLoS One* **15**, e0231595 (2020).
13. Martínez, M., González-Aravena, M., Held, C. & Abele, D. A molecular perspective on the invasibility of the southern ocean benthos: The impact of hypoxia and temperature on gene expression in south american and antarctic aequiyoldia bivalves. *Front. Physiol.* **14**, 1083240 (2023).
14. Coll, M. *et al.* Advancing global ecological modeling capabilities to simulate future trajectories of change in marine ecosystems. *Front. Mar. Sci.* **7**, 567877 (2020).
15. Oyinlola, M. A., Reygondeau, G., Wabnitz, C. C., Troell, M. & Cheung, W. W. Global estimation of areas with suitable environmental conditions for mariculture species. *PLoS One* **13**, e0191086 (2018).

16. Pompa, S., Ehrlich, P. R. & Ceballos, G. Global distribution and conservation of marine mammals. *Proc. Natl. Acad. Sci.* **108**, 13600–13605 (2011).

17. Ringler, T. *et al.* A multi-resolution approach to global ocean modeling. *Ocean. Model.* **69**, 211–232 (2013).

18. Ready, J. *et al.* Predicting the distributions of marine organisms at the global scale. *Ecol. Model.* **221**, 467–478 (2010).

19. Tittensor, D. P. *et al.* Global patterns and predictors of marine biodiversity across taxa. *Nature* **466**, 1098–1101 (2010).

20. Tittensor, D. P. *et al.* Next-generation ensemble projections reveal higher climate risks for marine ecosystems. *Nat. Clim. Chang.* **11**, 973–981 (2021).

21. Lotze, H. K. *et al.* Global ensemble projections reveal trophic amplification of ocean biomass declines with climate change. *Proc. Natl. Acad. Sci.* **116**, 12907–12912 (2019).

22. Kerr, J. T., Kharouba, H. M. & Currie, D. J. The macroecological contribution to global change solutions. *science* **316**, 1581–1584 (2007).

23. Muluneh, M. G. Impact of climate change on biodiversity and food security: a global perspective—a review article. *Agric. & Food Secur.* **10**, 1–25 (2021).

24. Hoegh-Guldberg, O., Northrop, E. & Lubchenco, J. The ocean is key to achieving climate and societal goals. *Science* **365**, 1372–1374 (2019).

25. Merow, C. *et al.* What do we gain from simplicity versus complexity in species distribution models? *Ecography* **37**, 1267–1281 (2014).

26. Tyberghein, L. *et al.* Bio-oracle: a global environmental dataset for marine species distribution modelling. *Glob. ecology biogeography* **21**, 272–281 (2012).

27. Martínez-Minaya, J., Cameletti, M., Conesa, D. & Pennino, M. G. Species distribution modeling: a statistical review with focus in spatio-temporal issues. *Stoch. environmental research risk assessment* **32**, 3227–3244 (2018).

28. Elith, J. & Leathwick, J. R. Species distribution models: ecological explanation and prediction across space and time. *Annu. review ecology, evolution, systematics* **40**, 677–697 (2009).

29. Martínez-Bello, D., López-Quílez, A. & Prieto, A. T. Spatiotemporal modeling of relative risk of dengue disease in colombia. *Stoch. environmental research risk assessment* **32**, 1587–1601 (2018).

30. Izquierdo, F., Menezes, R., Wise, L., Teles-Machado, A. & Garrido, S. Bayesian spatio-temporal cpue standardization: Case study of european sardine (sardina pilchardus) along the western coast of portugal. *Fish. Manag. Ecol.* **29**, 670–680 (2022).

31. Goetz, S. J., Sun, M., Zolkos, S., Hansen, A. & Dubayah, R. The relative importance of climate and vegetation properties on patterns of north american breeding bird species richness. *Environ. Res. Lett.* **9**, 034013 (2014).

32. Arntzen, J. W. A two-species distribution model for parapatric newts, with inferences on their history of spatial replacement. *Biol. J. Linnean Soc.* **138**, 75–88 (2023).

33. Charbonnel, A. *et al.* Developing species distribution models for critically endangered species using participatory data: The european sturgeon marine habitat suitability. *Estuarine, Coast. Shelf Sci.* **280**, 108136 (2023).

34. Guisan, A. & Zimmermann, N. E. Predictive habitat distribution models in ecology. *Ecol. modelling* **135**, 147–186 (2000).

35. Edwards Jr, T. C., Cutler, D. R., Zimmermann, N. E., Geiser, L. & Moisen, G. G. Effects of sample survey design on the accuracy of classification tree models in species distribution models. *ecological modelling* **199**, 132–141 (2006).

36. Carlson, C. J. embarcadero: Species distribution modelling with bayesian additive regression trees in r. *Methods Ecol. Evol.* **11**, 850–858 (2020).

37. Chipman, H. A., George, E. I. & McCulloch, R. E. BART: Bayesian additive regression trees. *The Annals Appl. Stat.* **4**, 266 – 298, DOI: 10.1214/09-AOAS285 (2010).

38. Martin, O. A., Kumar, R. & Lao, J. *Bayesian modeling and computation in python* (CRC Press, 2021).

39. Hill, J., Linero, A. & Murray, J. Bayesian additive regression trees: A review and look forward. *Annu. Rev. Stat. Its Appl.* **7**, 251–278 (2020).

40. Robson, B. J. *et al.* Towards evidence-based parameter values and priors for aquatic ecosystem modelling. *Environ. modelling & software* **100**, 74–81 (2018).

41. Steenbeek, J. *et al.* Making spatial-temporal marine ecosystem modelling better–a perspective. *Environ. Model. & Softw.* **145**, 105209 (2021).

42. Coll, M., Pennino, M. G., Steenbeek, J., Solé, J. & Bellido, J. M. Predicting marine species distributions: complementarity of food-web and bayesian hierarchical modelling approaches. *Ecol. Model.* **405**, 86–101 (2019).

43. Dansereau, G., Legendre, P. & Poisot, T. Evaluating ecological uniqueness over broad spatial extents using species distribution modelling. *Oikos* **2022**, e09063 (2022).

44. Poursanidis, D. *et al.* Uncertainty in marine species distribution modelling: Trying to locate invasion hotspots for pterois miles in the eastern mediterranean sea. *J. Mar. Sci. Eng.* **10**, 729 (2022).

45. Konowalik, K. & Nosol, A. Evaluation metrics and validation of presence-only species distribution models based on distributional maps with varying coverage. *Sci. Reports* **11**, 1–15 (2021).

46. Sparapani, R. *et al.* Novel electrocardiographic criteria for the diagnosis of left ventricular hypertrophy derived with bayesian additive regression trees: the multi-ethnic study of atherosclerosis. *Circulation* **138**, A10908–A10908 (2018).

47. Yen, J. D., Thomson, J. R., Vesk, P. A. & Mac Nally, R. To what are woodland birds responding? inference on relative importance of in-site habitat variables using several ensemble habitat modelling techniques. *Ecography* **34**, 946–954 (2011).

48. Thompson, M. S., Couce, E., Schratzberger, M. & Lynam, C. P. Climate change affects the distribution of diversity across marine food webs. *Glob. Chang. Biol.* (2023).

49. Ahmadi, K. *et al.* Modeling tree species richness patterns and their environmental drivers across hyrcanian mountain forests. *Ecol. Informatics* **77**, 102226 (2023).

50. Costa, E. F., Encarnação, J., Teodósio, M. A. & Morais, P. Aquatic species shows asymmetric distribution range shifts in native and non-native areas. *Front. Mar. Sci.* **10**, 1158206 (2023).

51. Lutz, P. L., Musick, J. A. & Wyneken, J. *The biology of sea turtles, Volume II*, vol. 2 (CRC press, 2002).

52. Bowen, B. W. & Karl, S. Population genetics and phylogeography of sea turtles. *Mol. ecology* **16**, 4886–4907 (2007).

53. Maurer, A. S. *et al.* Population viability of sea turtles in the context of global warming. *BioScience* **71**, 790–804 (2021).

54. Mazaris, A. D. *et al.* Priorities for mediterranean marine turtle conservation and management in the face of climate change. *J. Environ. Manag.* **339**, 117805 (2023).

55. Jensen, M. P. *et al.* Environmental warming and feminization of one of the largest sea turtle populations in the world. *Curr. Biol.* **28**, 154–159 (2018).

56. Hempel, S., Frieler, K., Warszawski, L., Schewe, J. & Piontek, F. A trend-preserving bias correction–the isi-mip approach. *Earth Syst. Dyn.* **4**, 219–236 (2013).

57. Warszawski, L. *et al.* The inter-sectoral impact model intercomparison project (isi–mip): project framework. *Proc. Natl. Acad. Sci.* **111**, 3228–3232 (2014).

58. Reese, G. C., Wilson, K. R., Hoeting, J. A. & Flather, C. H. Factors affecting species distribution predictions: a simulation modeling experiment. *Ecol. Appl.* **15**, 554–564 (2005).

59. Poloczanska, E. S., Limpus, C. J. & Hays, G. C. Vulnerability of marine turtles to climate change. *Adv. marine biology* **56**, 151–211 (2009).

60. Hawkes, L. A., Broderick, A. C., Godfrey, M. H. & Godley, B. J. Climate change and marine turtles. *Endangered Species Res.* **7**, 137–154 (2009).

61. IUCN. The iucn red list of threatened species. version 2023-1. https://www.iucnredlist.org (2023). Accessed on 12 Februeary 2024.

62. Red List Standards & Petitions Subcommittee. Natator depressus (errata version published in 2022). *The IUCN Red List Threat. Species* **1996**, e.T14363A210612474 (1996). Accessed on 12 February 2024.

63. Wallace, B., Tiwari, M. & Girondot, M. Dermochelys coriacea. *The IUCN Red List Threat. Species* **2013**, e.T6494A43526147, DOI: 10.2305/IUCN.UK.2013-2.RLTS.T6494A43526147.en (2013). Accessed on 12 February 2024.

64. Casale, P. & Tucker, A. Caretta caretta (amended version of 2015 assessment). *The IUCN Red List Threat. Species* **2017**, e.T3897A119333622, DOI: 10.2305/IUCN.UK.2017-2.RLTS.T3897A119333622.en (2017). Accessed on 12 February 2024.

65. Abreu-Grobois, A. & Plotkin, P. I. S. M. T. S. G. Lepidochelys olivacea. *The IUCN Red List Threat. Species* **2008**, e.T11534A3292503, DOI: 10.2305/IUCN.UK.2008.RLTS.T11534A3292503.en (2008). Accessed on 12 February 2024.

66. Wibbels, T. & Bevan, E. Lepidochelys kempii (errata version published in 2019). *The IUCN Red List Threat. Species* **2019**, e.T11533A155057916, DOI: 10.2305/IUCN.UK.2019-2.RLTS.T11533A155057916.en (2019). Accessed on 12 February 2024.

67. Mortimer, J. & Donnelly, M. I. S. M. T. S. G. Eretmochelys imbricata. *The IUCN Red List Threat. Species* **2008**, e.T8005A12881238, DOI: 10.2305/IUCN.UK.2008.RLTS.T8005A12881238.en (2008). Accessed on 12 February 2024.

68. de Sousa Miranda, L. *et al.* Combining connectivity and species distribution modeling to define conservation and restoration priorities for multiple species: a case study in the eastern amazon. *Biol. Conserv.* **257**, 109148 (2021).

69. Vigo, M. *et al.* Dynamic marine spatial planning for conservation and fisheries benefits. *Fish Fish.* (2024).

70. Afán, I., Giménez, J., Forero, M. G. & Ramírez, F. An adaptive method for identifying marine areas of high conservation priority. *Conserv. biology* **32**, 1436–1447 (2018).

71. Afán, I., Chiaradia, A., Forero, M. G., Dann, P. & Ramírez, F. A novel spatio-temporal scale based on ocean currents unravels environmental drivers of reproductive timing in a marine predator. *Proc. Royal Soc. B: Biol. Sci.* **282**, 20150721 (2015).

72. Coles, W. & Musick, J. A. Satellite sea surface temperature analysis and correlation with sea turtle distribution off north carolina. *Copeia* **2000**, 551–554 (2000).

73. Maneja, R. H. *et al.* Multidecadal analysis of beach loss at the major offshore sea turtle nesting islands in the northern arabian gulf. *Ecol. Indic.* **121**, 107146 (2021).

74. Carreras, C. *et al.* Sporadic nesting reveals long distance colonisation in the philopatric loggerhead sea turtle (caretta caretta). *Sci. reports* **8**, 1435 (2018).

75. Báez, J. C. *et al.* Primer registro de nidificación de tortuga boba (caretta caretta) en el mar de alborán: significado biológico e implicaciones del manejo en la conservación. *Boletín Asociación Herpetológica Española* **31**, 157–162 (2020).

76. Patrício, A. R., Hawkes, L. A., Monsinjon, J. R., Godley, B. J. & Fuentes, M. M. Climate change and marine turtles: Recent advances and future directions. *Endangered Species Res.* **44**, 363–395 (2021).

77. Wallace, B. P. *et al.* Global patterns of marine turtle bycatch. *Conserv. letters* **3**, 131–142 (2010).

78. Camiñas, J. A. *et al.* Tuna regional fisheries management organizations and the conservation of sea turtles: a reply to godley et al. *Oryx* **55**, 12–12 (2021).

79. Wallace, B. P. *et al.* Marine turtle regional management units 2.0: an updated framework for conservation and research of wide-ranging megafauna species. *Endangered Species Res.* **52**, 209–223 (2023).

80. Liu, H., Song, Z., Wang, X. & Misra, V. An ocean perspective on cmip6 climate model evaluations (2022).

81. Howard, C., Stephens, P. A., Pearce-Higgins, J. W., Gregory, R. D. & Willis, S. G. Improving species distribution models: the value of data on abundance. *Methods Ecol. Evol.* **5**, 506–513 (2014).

82. Gaos, A. R. *et al.* Shifting the life-history paradigm: discovery of novel habitat use by hawksbill turtles. *Biol. Lett.* **8**, 54–56 (2012).

83. Redfern, J. *et al.* Techniques for cetacean–habitat modeling. *Mar. Ecol. Prog. Ser.* **310**, 271–295 (2006).

84. Hamann, M. *et al.* Global research priorities for sea turtles: informing management and conservation in the 21st century. *Endangered species research* **11**, 245–269 (2010).

85. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2022).

86. Gomes, V. H. *et al.* Species distribution modelling: Contrasting presence-only models with plot abundance data. *Sci. reports* **8**, 1003 (2018).

87. Naimi, B., Skidmore, A. K., Groen, T. A. & Hamm, N. A. Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling. *J. Biogeogr.* **38**, 1497–1509 (2011).

88. Morera-Pujol, V. *et al.* Bayesian species distribution models integrate presence-only and presence–absence data to predict deer distribution and relative abundance. *Ecography* **2023**, e06451 (2023).

89. Chamberlain, S. *et al. rgbif: Interface to the Global Biodiversity Information Facility API* (2023). R package version 3.7.5.

90. Chamberlain, S. & Boettiger, C. R python, and ruby clients for gbif species occurrence data. *PeerJ PrePrints* (2017).

91. Zizka, A. *et al.* Coordinatecleaner: standardized cleaning of occurrence records from biological collection databases. *Methods Ecol. Evol.* –7, DOI: 10.1111/2041-210X.13152 (2019). R package version 2.0-20.

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

93. Lange, S. & Büchner, M. Isimip3b bias-adjusted atmospheric climate input data (v1.1.). *ISIMIP Repositoy* (2021).

94. Loh, W.-Y. Logistic regression tree analysis. In *Springer handbook of engineering statistics*, 593–604 (Springer, 2023).

95. Loh, W.-Y. Classification and regression trees. *Wiley interdisciplinary reviews: data mining knowledge discovery* **1**, 14–23 (2011).

96. Karakaya, J. Evaluation of binary diagnostic tests accuracy for medical researches. *Turkish J. Biochem.* **46**, 103–113 (2020).

97. Hoyle, S. D. *et al.* Catch per unit effort modelling for stock assessment: A summary of good practices. *Fish. Res.* **269**, 106860 (2024).

98. Worm, B. & Tittensor, D. P. *A theory of global biodiversity (MPB-60)* (Princeton University Press, 2018).

99. Paradinas, I., Illian, J. & Smout, S. Understanding spatial effects in species distribution models. *Authorea Prepr.* (2022).

100. Krainski, E. *et al. Advanced spatial modeling with stochastic partial differential equations using R and INLA* (Chapman and Hall/CRC, 2018).

Acknowledgements (not compulsory)

This study is a contribution to the project ProOceans (Ministerio de Ciencia e Innovación, Proyectos de I + D + I (RETOS PID2020-118097RB-I00)). AFA received funding from the Spanish project ProOceans through an FPI grant (Ministerio de Ciencia e Innovación, Proyectos de I + D + I (RETOSPID2020-118097RB-I00)). AFA, MC, JMT and JS acknowledge funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 817578 (TRIATLAS project) and Nº 869300 (FutureMares). XB, DC and ALQ acknowledge support by the grant PID2022-136455NB-I00, funded by Ministerio de Ciencia, Innovación y Universidades of Spain (MCIN/AEI/10.13039/501100011033/FEDER, UE) and the European Regional Development Fund. DC also acknowledges funding provided by the Conselleria de Educación, Universidades y Empleo de la Generalitat Valenciana via the project CIAICO/2022/165. Authors also acknowledge institutional support of the 'Severo Ochoa Center of Excellence' accreditation (CEX2019-000928-S). This project was supported by the Catalan government through the iMARES research group of quality at the Institute of Marine Sciences (ICM-CSIC) in Barcelona.

Author contributions statement

A.F.A., D.C., M.G.P, X.B., V.C., A.L.Q., J.S., J.M.B., and M.C. conceptualized the work. A.F.A. developed and implemented the model approach, performed the analysis, and generated visualizations. J.M.T. contributed to implementing the model approach. M.C. envisioned the study and supervised the work. J.C.B. developed the biological part of the discussion. A.F.A. drafted the manuscript. All authors reviewed the manuscript the final manuscript.

Data availability

The datasets generated and/or analysed during the current study are available in the Global Biodiversity Information Facility (GBIF) repository, https://www.gbif.org/es/. DOIs are available in the supplementary material.

Additional information

The authors have no conflict of interest to declare.

Species	GFDL-ESM4		IPSL-CM6A-LR	
	SSP126	SSP585	SSP126	SSP585
Natator depressus	1.81	2.21	12.88	47.14
Dermochelys coriacea	-2.62	-6.97	10.53	57.12
Caretta caretta	-4.83	-1.97	-14.72	-23.25
Lepidochelys olivacea	-24.31	-44.51	-1.45	-29.35
Chelonia mydas	11.70	40.11	9.17	41.63
Lepidochelys kempii	-2.85	4.45	36.68	-9.31
Eretmochelys imbricata	21.38	45.59	11.74	62.18
Functional group	-0.80	3.51	26.58	43.98

Table 1. Percentage (%) of increase or decrease of the suitable habitat's mean probability between the historical suitable habitat (1950-2014) and the last ten years of the future suitable habitat's projections (2089-2099).

GFDL-ESM4				
Species	***Sensitivity***	***Specificity***	***Accuracy***	***F_1 score***
Natator depressus	0.98	0.98	0.98	0.98
Dermochelys coriacea	0.80	0.84	0.82	0.82
Caretta caretta	0.94	0.90	0.92	0.92
Lepidochelys olivacea	0.92	0.93	0.92	0.92
Chelonia mydas	0.93	0.92	0.92	0.92
Lepidochelys kempii	0.97	0.98	0.98	0.98
Eretmochelys imbricata	0.95	0.94	0.95	0.95
IPSL-CM6A-LR				
Natator depressus	0.97	0.97	0.97	0.97
Dermochelys coriacea	0.73	0.83	0.78	0.77
Caretta caretta	0.90	0.88	0.90	0.90
Lepidochelys olivacea	0.90	0.91	0.90	0.90
Chelonia mydas	0.91	0.89	0.90	0.90
Lepidochelys kempii	0.96	0.97	0.96	0.96
Eretmochelys imbricata	0.93	0.94	0.93	0.93

Table 2. Different error measures for each species and ESM results (GFDL-ESM4 and IPSL-CM6A-LR). We have calculated sensitivity, specificity, accuracy, and F_1 score.

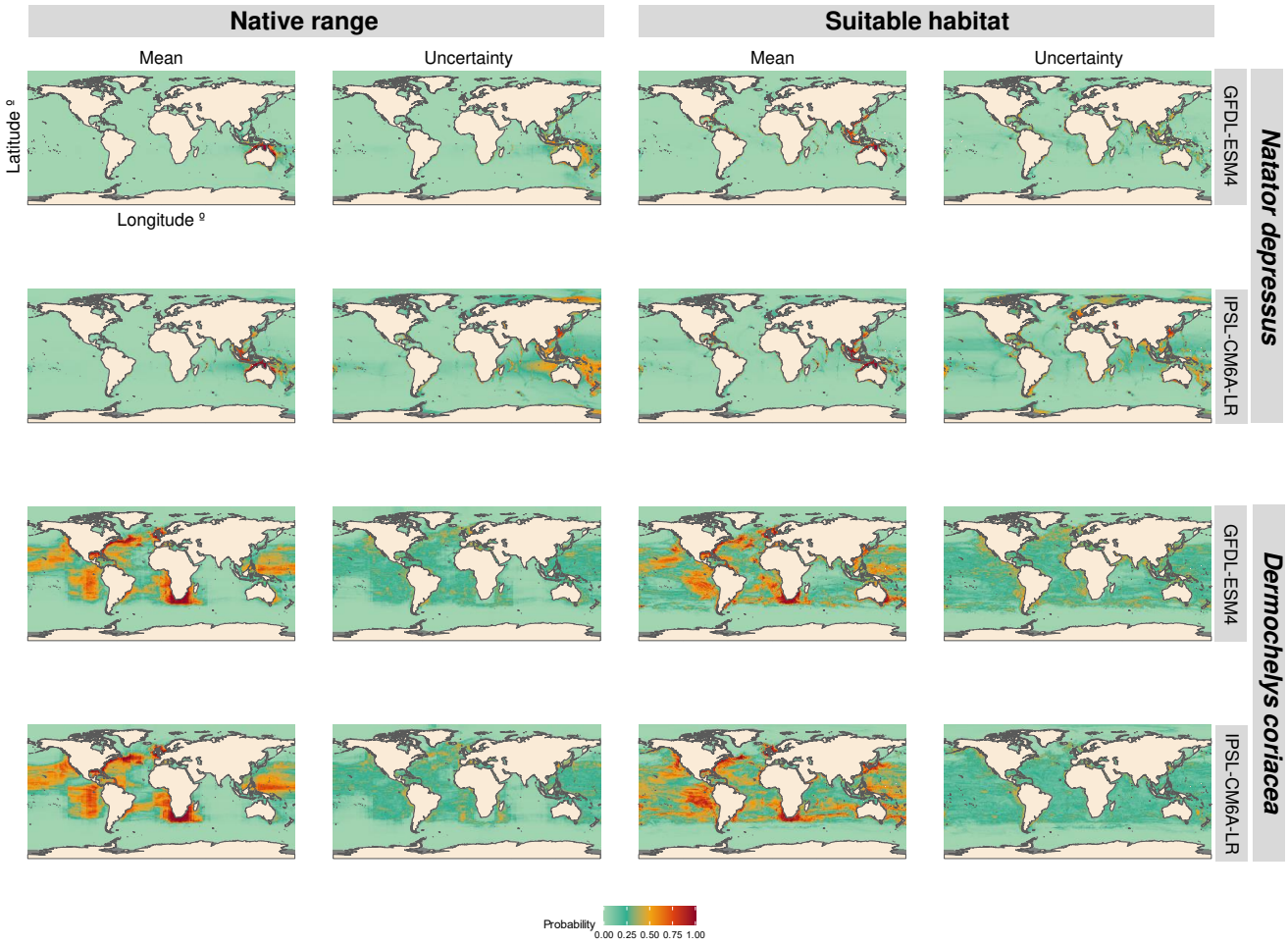

Figure 1. Maps depict the probability of presence for two species from 1950 to 2014, *Natator depressus* and *Dermochelys coriacea*. The first and second columns illustrate the native ranges (current distribution), while the third and fourth columns portray the suitable or potential habitats. The first and third rows correspond to the results for the GFDL-ESM4 model, while the second and fourth rows depict the results of IPSL-CM6ALR. We are presenting the mean posterior predictive distribution for both species, accompanied by uncertainty represented as the subtraction of quantiles 0.025 and 0.975.

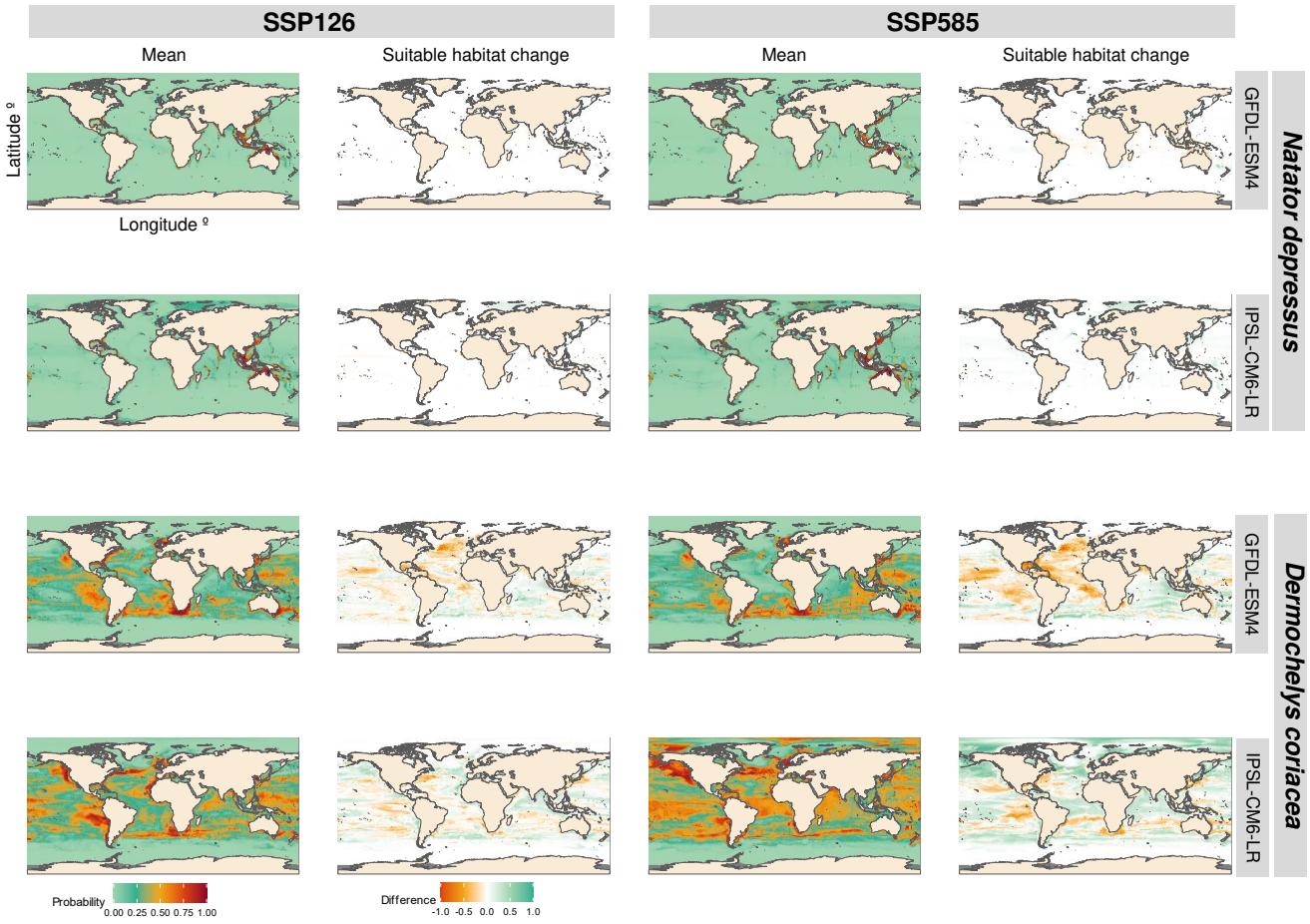

Figure 2. Maps representing the mean probability of presence from 2089 to 2099 for *Natator depressus* and *Dermochelys coriacea*, along with the difference between the historical suitable habitat (Figure 1) and the projections for the last 10 years (2089-2099). We have calculated the difference for both climate change scenarios, SSP126 and SSP585, and also for both Earth System Models (GFDL-ESM4 and IPSL-CM6A-LR).

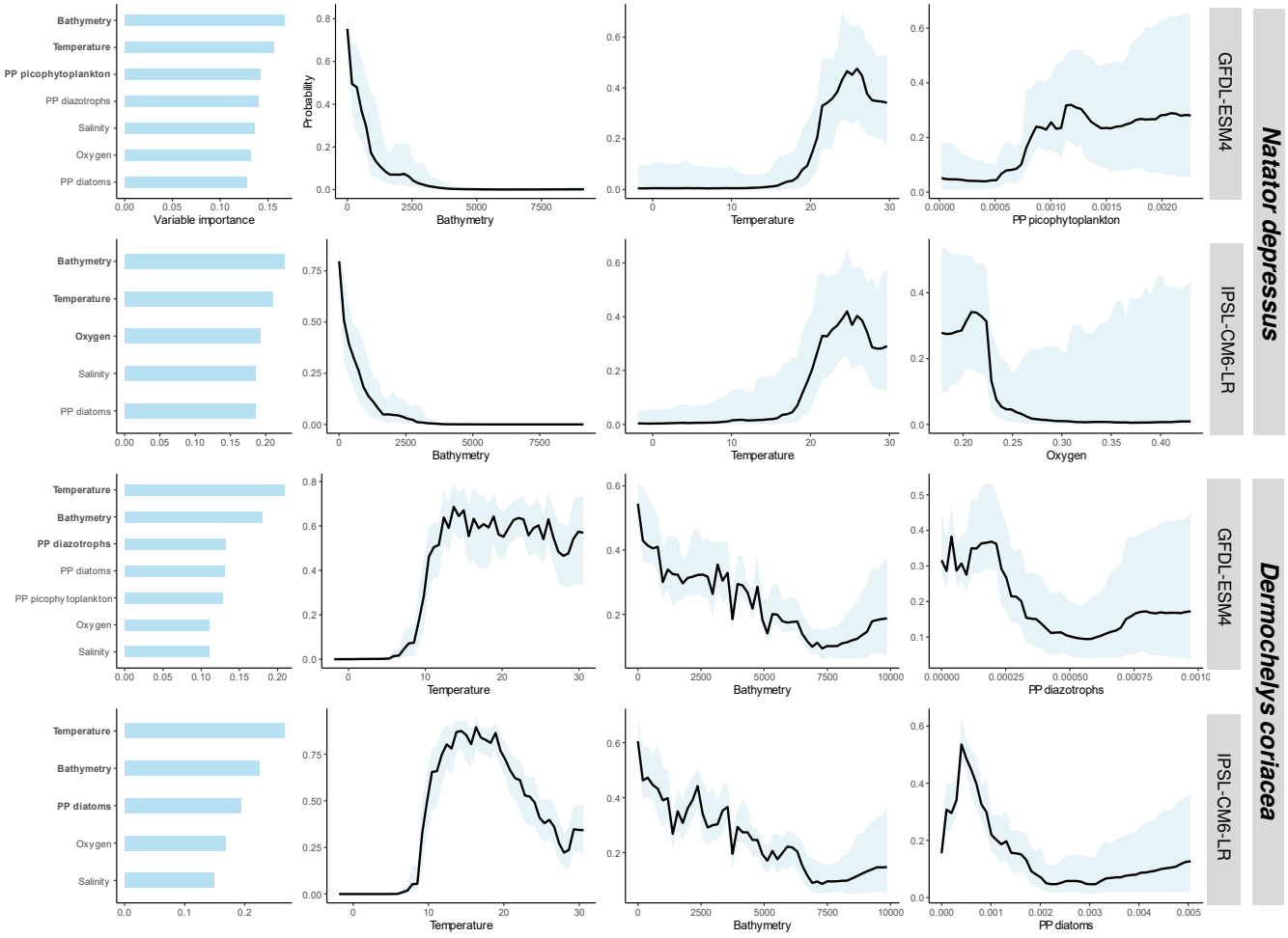

Figure 3. Contributions of all the variables to the model for both ESMs are presented. We also provide the additive relation for the variables that have contributed the most to the model. These additive relations represent the probability of being present at some point along the x-axis.

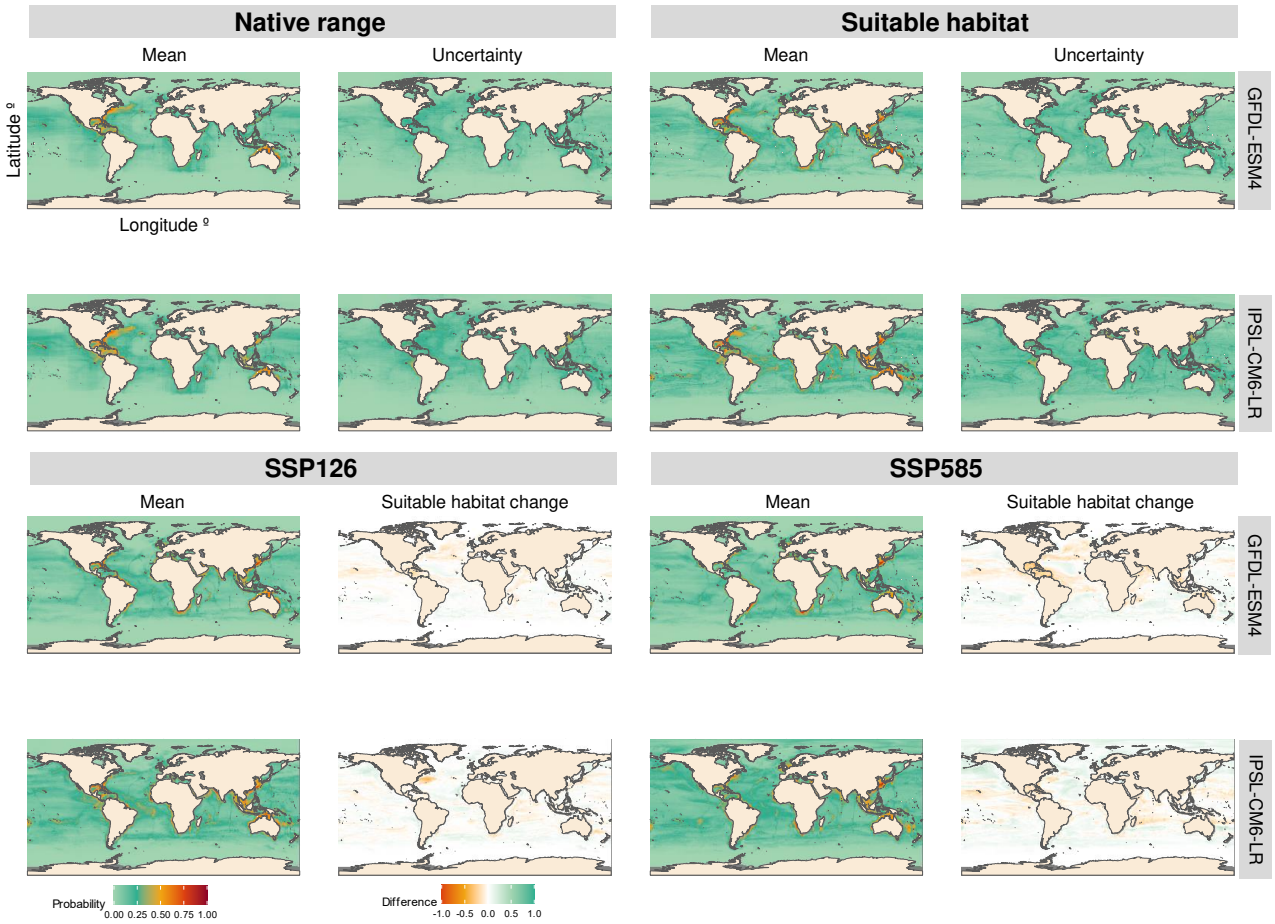

Figure 4. Functional group results of the native ranges and suitable habitats (1950-2014) and future suitable habitats (2015-2100) are provided. Rows one and two represent the spatial predictions for the current distribution, while the third and fourth rows depict the predictions for the last ten years of projections (2089-2099), including the difference between the projections and the current suitable habitat. All of these are represented for both climate scenarios and ESMs.

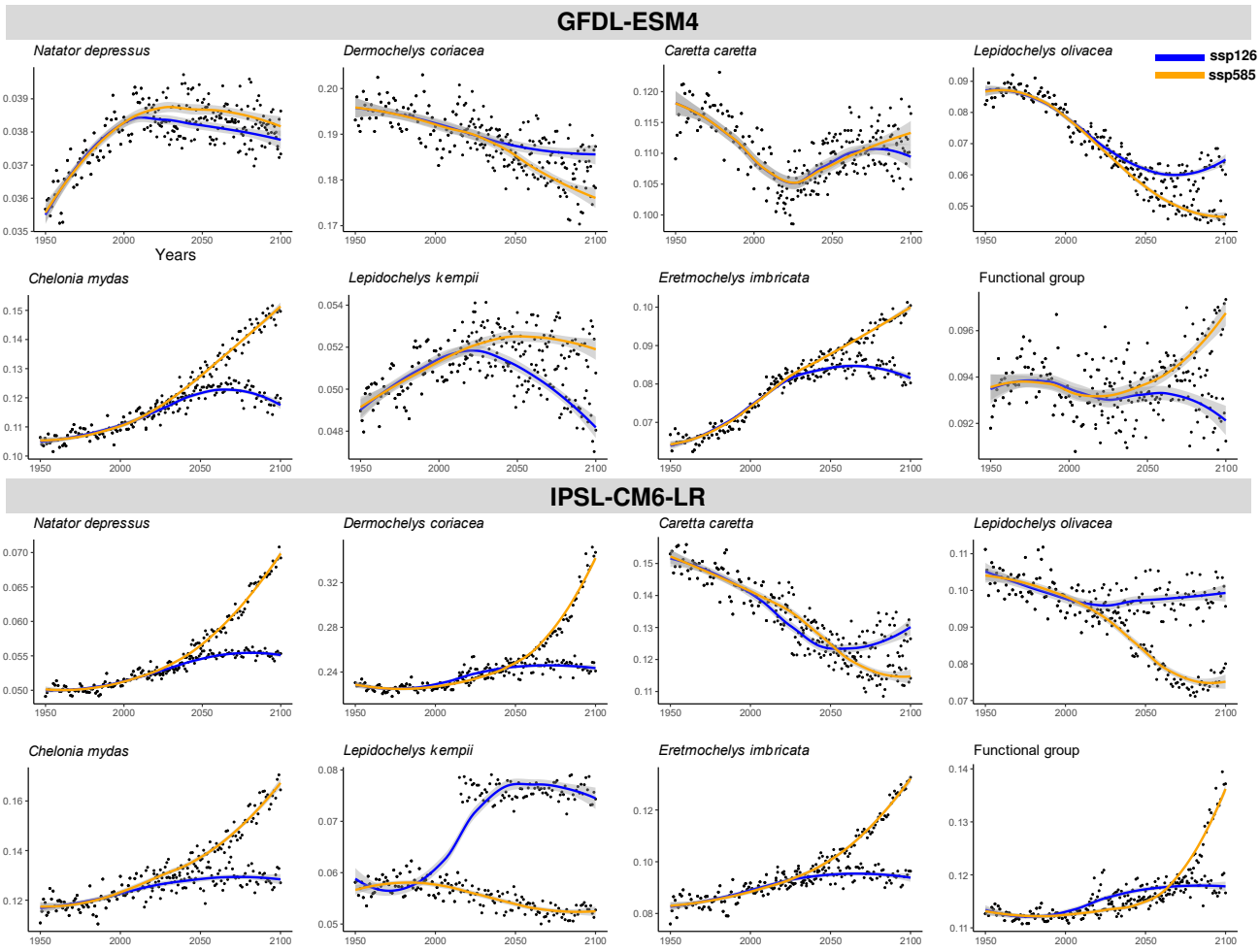

Figure 5. Changes over time in the mean probability of suitable habitat. The x-axis represents the years from 1950 to 2100, while the y-axis represents the mean probability for each year of the projected suitable habitat. The orange line represents the climate scenario SSP585, and the blue line represents the SSP126 climate scenario. Dots represent the mean probability calculated for each year.

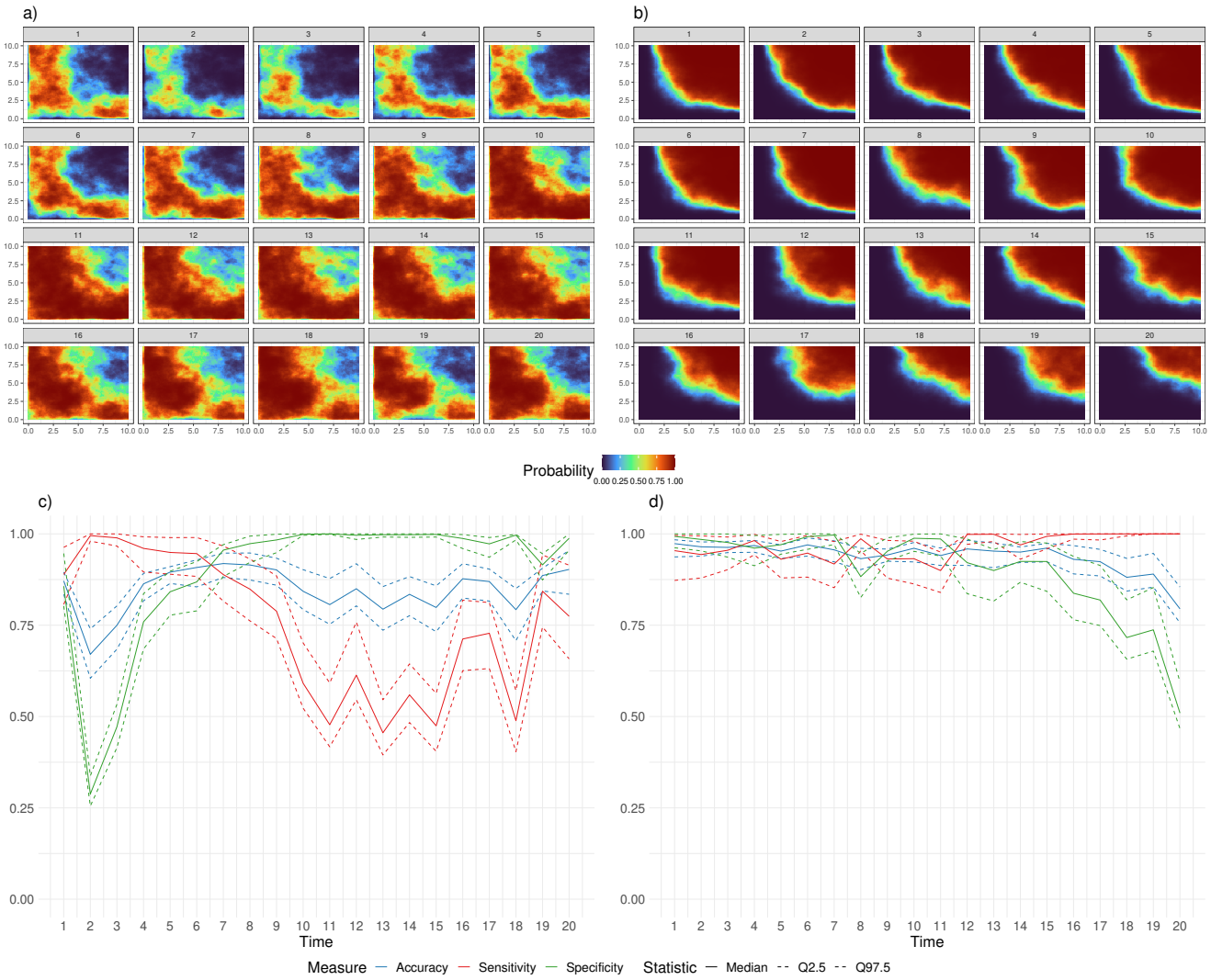

Figure 6. a) and b) are the simulation of the probability distribution in space and time for a cosmopolitan and persistent species scenarios respectively. The time window is 20 years, and we can observe changes over space and time. c) and d) are the results of sensitivity, specificity, and accuracy for the cosmopolitan and persistent scenarios respectively. We have calculated the mean and quantiles (0.025 and 0.975) over the 50 replications conducted.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementarymaterialFusterAlonsoScientificreports.pdf