# 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

## SUPPLEMENTARY MATERIAL

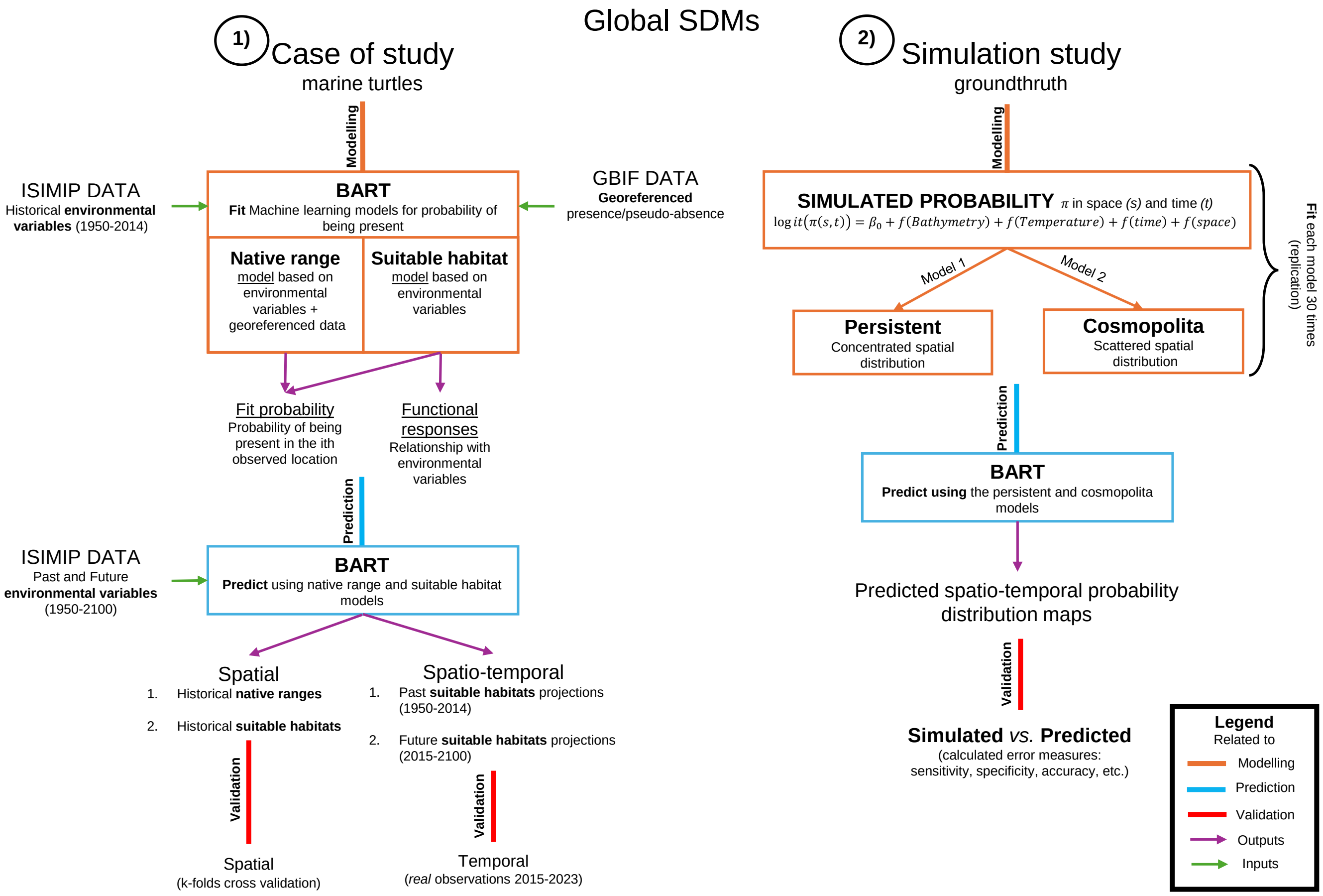
## 1 Marine turtles information and study workflow

| *Species* | *IUCN Red List* | *Climate Zone* | *Distribution* |
|---|---|---|---|
| *Natator depressus* | Data Deficient | Tropical | Indo-West Pacific |
| *Dermochelys coriacea* | Vulnerable | Tropical | Circumglobal |
| *Caretta caretta* | Vulnerable | Tropical | Circumglobal |
| *Lepidochelys olivacea* | Vulnerable | Tropical | Indo-Pacific and Atlantic Ocean |
| *Chelonia mydas* | Endangered | Tropical | Circumglobal |
| *Lepidochelys kempii* | Critically Endangered | Tropical | Atlantic Ocean and Mediterranean |
| *Eretmochelys imbricata* | Critically Endangered | Tropical | Circumglobal |

**Table S1.** Marine turtle species information from IUCN Red List and SeaLifeBase, including IUCN Red List status, climate zone, and distribution.

| ***Species*** | ***DOI*** |
|---|---|
| *Natator depressus* | https://doi.org/10.15468/dl.wbweak |
| *Dermochelys coriacea* | https://doi.org/10.15468/dl.4ub6fn |
| *Caretta caretta* | https://doi.org/10.15468/dl.bvgx97 |
| *Lepidochelys olivacea* | https://doi.org/10.15468/dl.bmhp5d |
| *Chelonia mydas* | https://doi.org/10.15468/dl.e3757n |
| *Lepidochelys kempii* | https://doi.org/10.15468/dl.5gzu4c |
| *Eretmochelys imbricata* | https://doi.org/10.15468/dl.8uw52a |

**Table S2.** Digital Object Identifier (DOI) from GBIF for each marine turtle species.



Global SDMs
1) Case of study
marine turtles
Modelling
ISIMIP DATA
Historical environmental variables (1950-2014)
BART
Fit Machine learning models for probability of being present
GBIF DATA
Georeferenced presence/pseudo-absence
Native range
model based on environmental variables + georeferenced data
Suitable habitat
model based on environmental variables
Fit probability
Probability of being present in the ith observed location
Functional responses
Relationship with environmental variables
Prediction
ISIMIP DATA
Past and Future environmental variables (1950-2100)
BART
Predict using native range and suitable habitat models
Spatial
1. Historical native ranges
2. Historical suitable habitats
Spatio-temporal
1. Past suitable habitats projections (1950-2014)
2. Future suitable habitats projections (2015-2100)
Validation
Validation
Spatial
(k-folds cross validation)
Temporal
(real observations 2015-2023)
2) Simulation study
groundthruth
Modelling
SIMULATED PROBABILITY π in space (s) and time (t)
$\log it(\pi(s,t)) = \beta_0 + f(Bathymetry) + f(Temperature) + f(time) + f(space)$
Fit each model 30 times (replication)
Model 1
Model 2
Persistent
Concentrated spatial distribution
Cosmopolita
Scattered spatial distribution
Prediction
BART
Predict using the persistent and cosmopolita models
Predicted spatio-temporal probability distribution maps
Validation
Simulated vs. Predicted
(calculated error measures: sensitivity, specificity, accuracy, etc.)
Legend
Related to
Modelling
Prediction
Validation
Outputs
Inputs

**Figure S1.** Study workflow: 1) represents the workflow follow for the marine turtles case of study and 2) represents the simulation study. Orange color refers to the modelling process, blue color refers to prediction and red color refers to validation. Green arrows represent the inputs use in the modelling and purple arrows represent the model and prediction outputs.

## 2 Species predictions

For the species spatio-temporal predictions, we have two different outputs. Figures 1, 2 and 3 refer to the spatial current predictions (native ranges and suitable habitats) done in this work for 5 species (1950-2014). While Figures 3, 4, and 5 refer to the suitable habitats obtain for those 5 species using future projections of environmental variables (2015-2100).

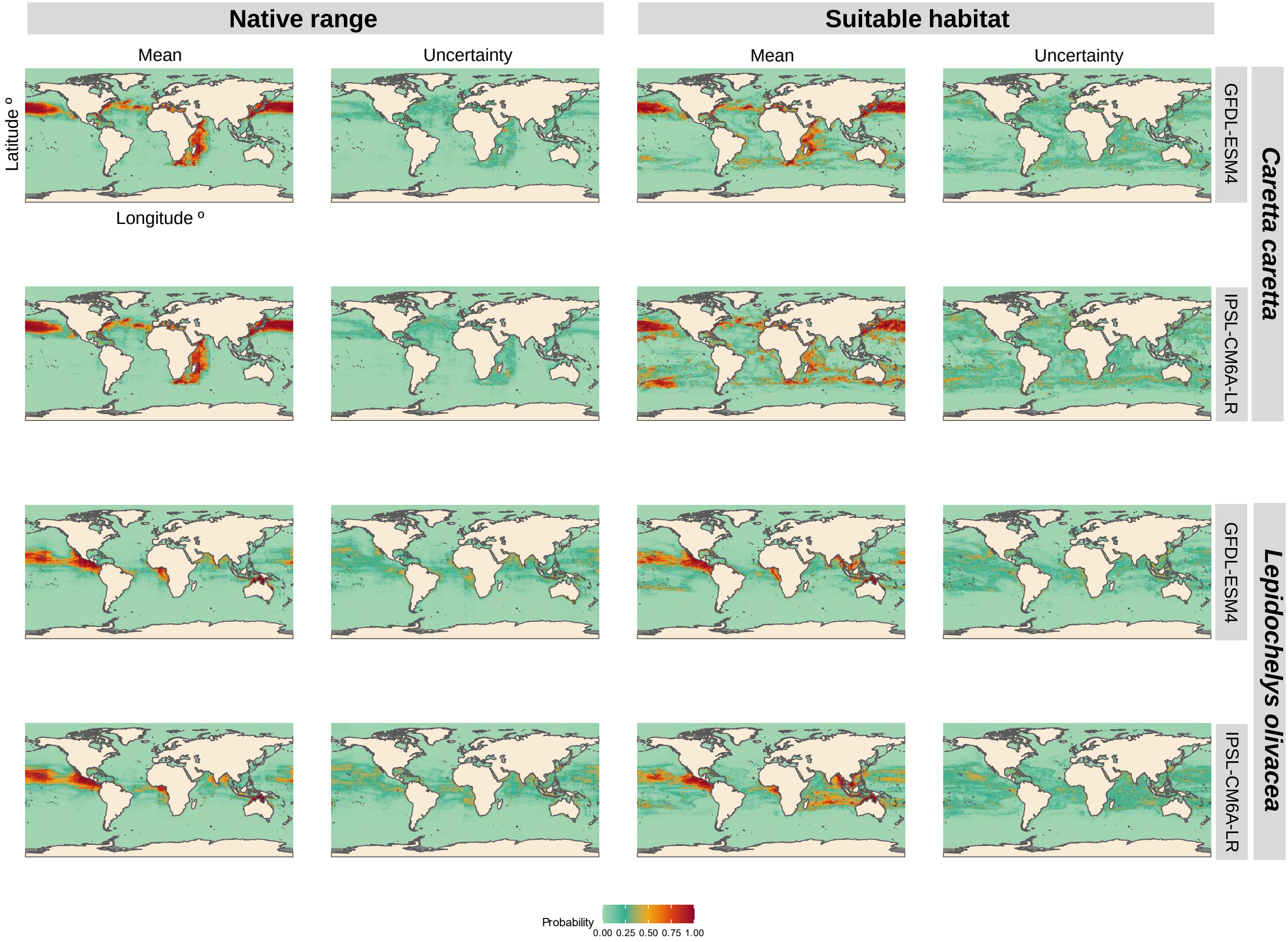


**Figure S2.** Maps depict the probability of presence for two species from 1950 to 2014, *Caretta caretta* and *Lepidochelys olivacea*. The first and second columns illustrate the native ranges (current distribution), while the third and fourth columns portray the suitable or potential habitats. The 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.

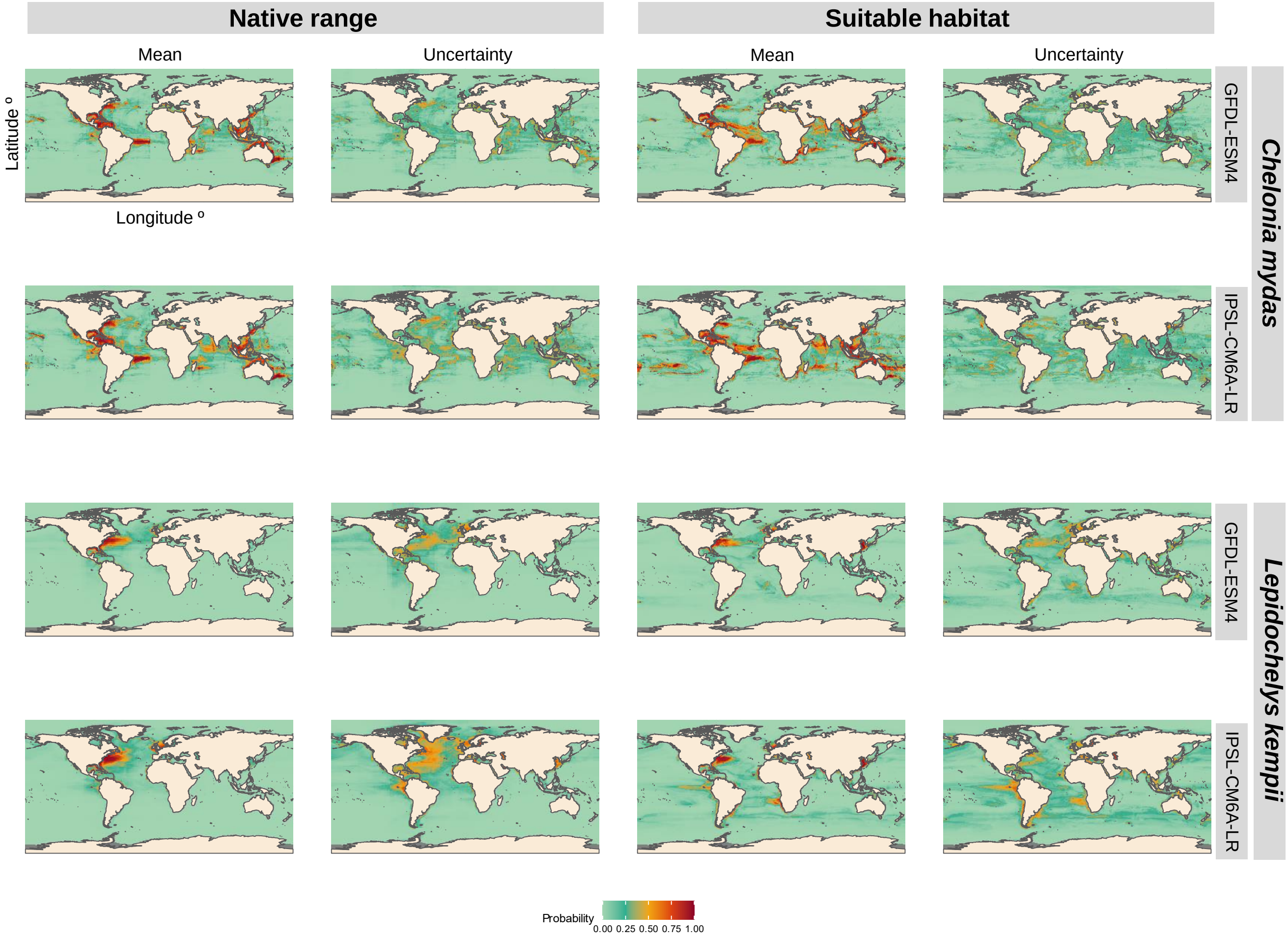


**Figure S3.** Maps depict the probability of presence for two species from 1950 to 2014, *Chelonia mydas* and *Lepidochelys kempii*. The first and second columns illustrate the native ranges (current distribution), while the third and fourth columns portray the suitable or potential habitats. The first 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.

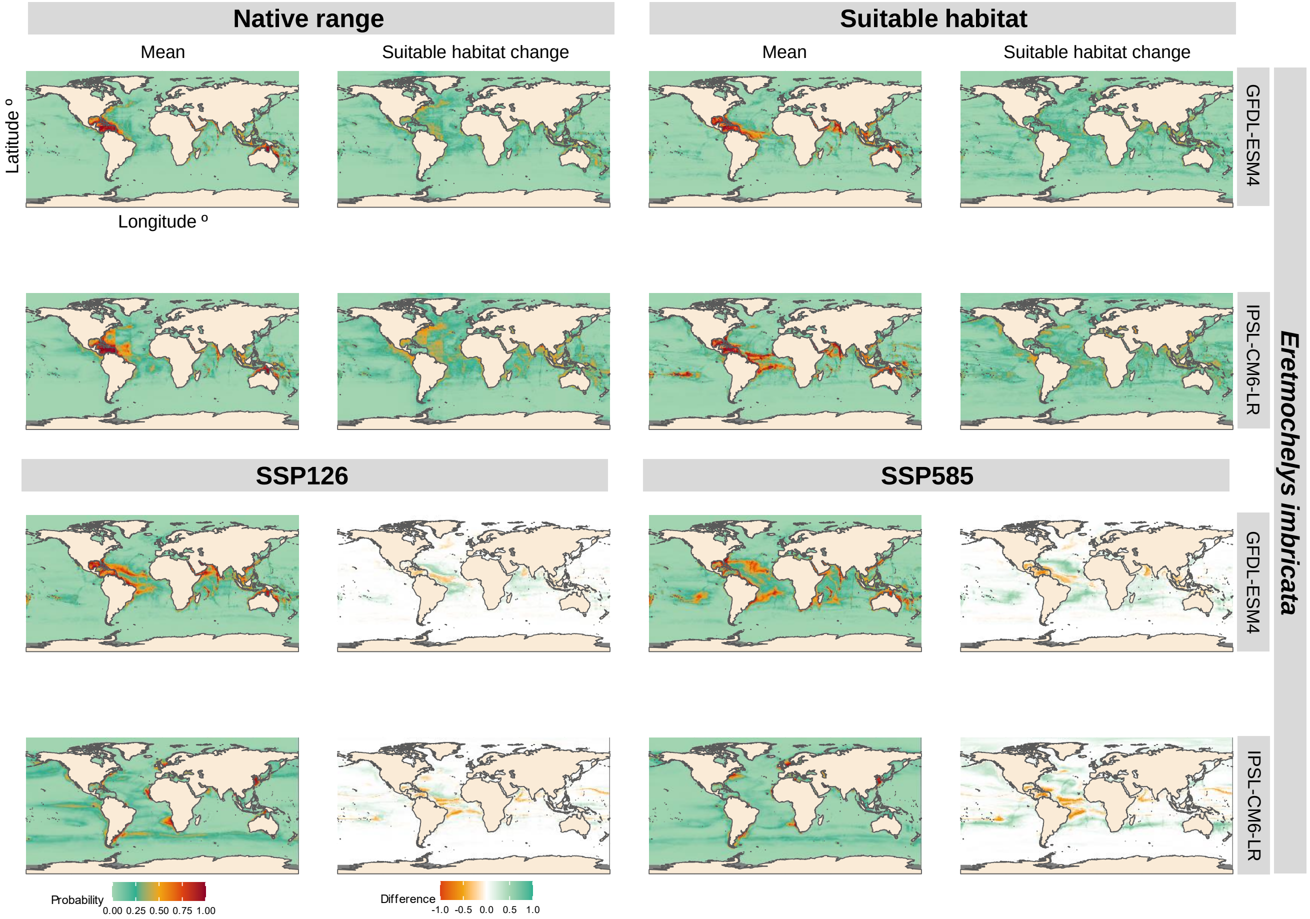


**Figure S4.** Maps depict the probability of presence for one species from 1950 to 2014 of *Eretmochelys imbricata*. The first and second columns and rows illustrate the native ranges, while the third and fourth columns and the first and second rows portray the suitable or potential habitats. We are presenting the mean posterior predictive distribution, accompanied by uncertainty represented as the subtraction of quantiles 0.025 and 0.975. Rows thrid and fourth represent the maps with the mean probability of presence from 2089 to 2099, along with the difference between the historical suitable habitat and the projections for the last 10 years (2089-2099).

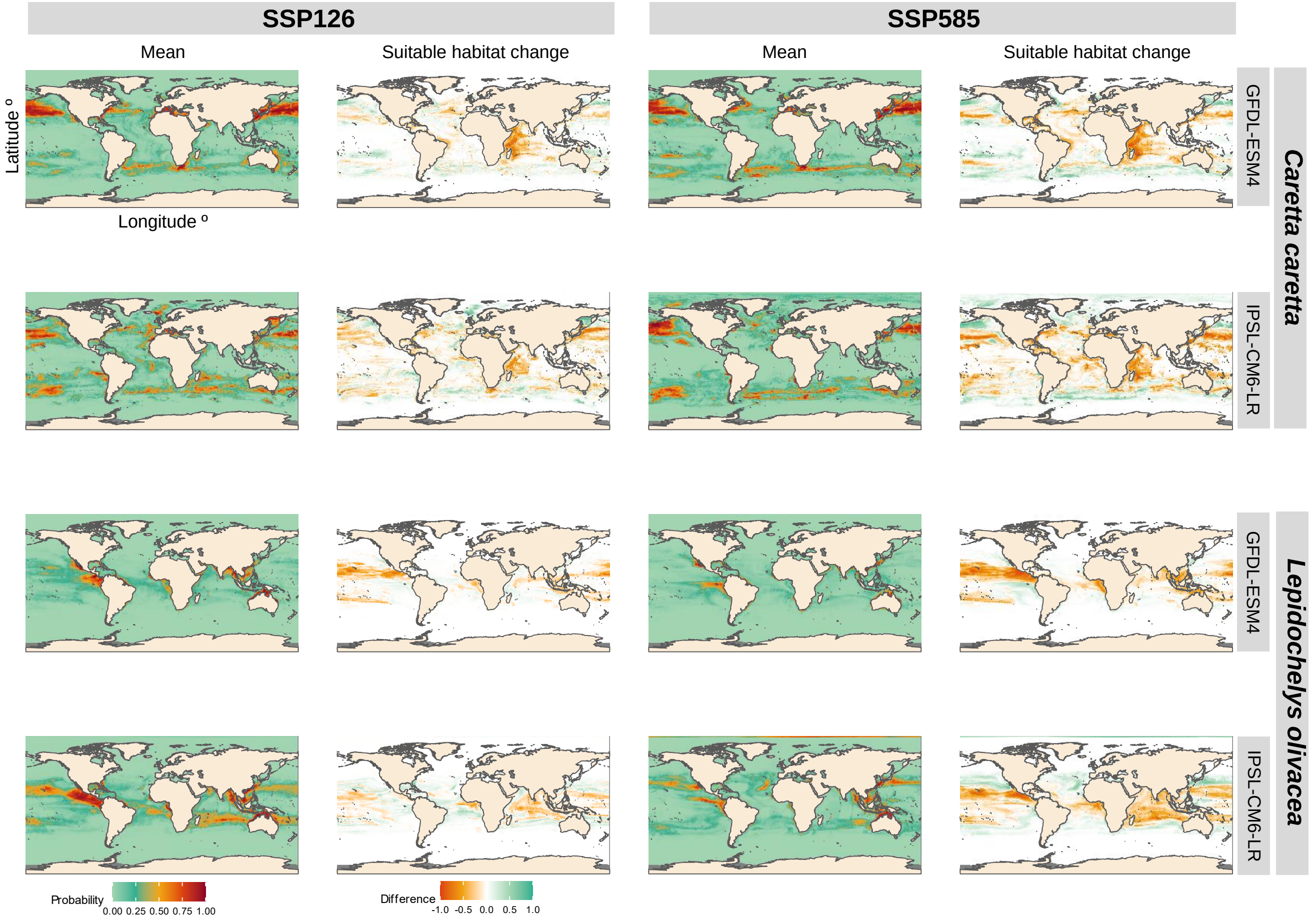


**Figure S5.** Maps representing the mean probability of presence from 2089 to 2099 for *Caretta caretta* and *Lepidochelys olivacea*, along with the difference between the historical suitable habitat and the projections for the last 10 years (2089-2099). We have calculated the difference for both climate change scenarios, ssp126 and ssp585, and also for both Earth System Models (GFDL-ESM4 and IPSL-CM6A-LR).

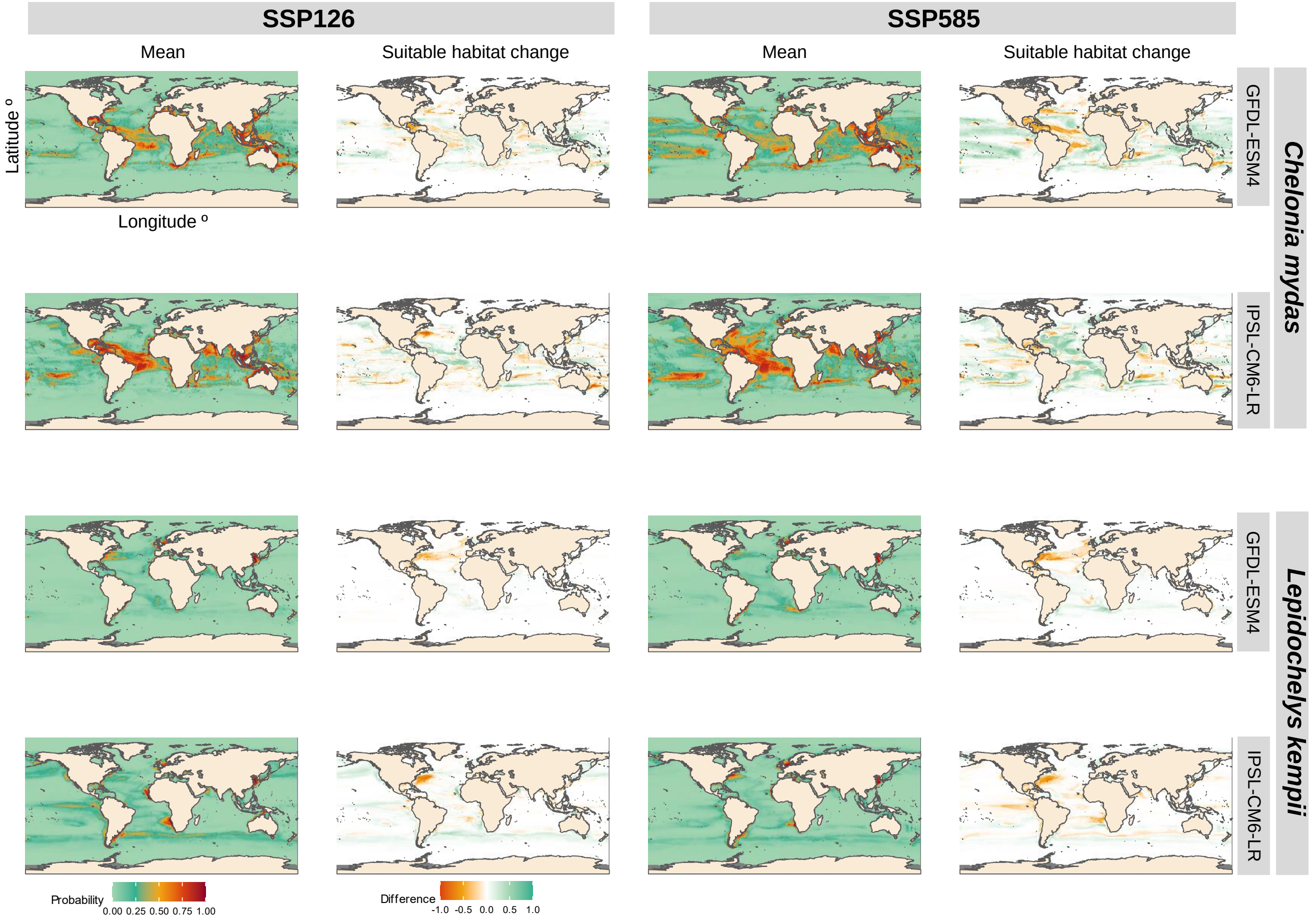


**Figure S6.** Maps representing the mean probability of presence from 2089 to 2099 for *Chelonia mydas* and *Lepidochelys kempii*, along with the difference between the historical suitable habitat and the projections for the last 10 years (2089-2099). We have calculated the difference for both climate change scenarios, ssp126 and ssp585, and also for both Earth System Models (GFDL-ESM4 and IPSL-CM6A-LR).

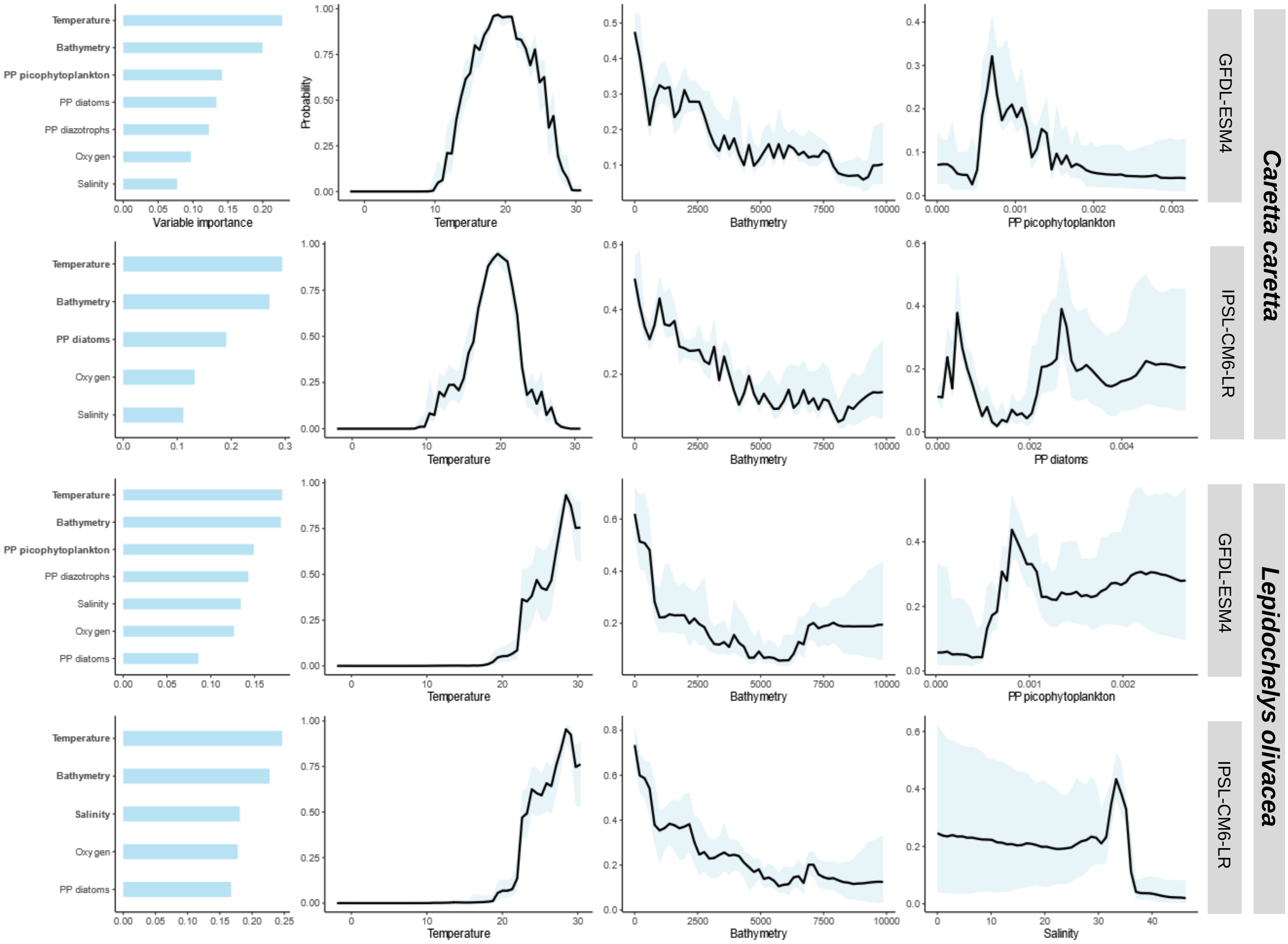


**Figure S7.** Results for the species *Caretta caretta* and *Lepidochelys olivacea*. The first column represents the contributions of all the variables to the model for both ESMs. We also provide the additive relation for the variables that have contributed the most to the model. These additive relations represent the probability of being present at some point along the x-axis.

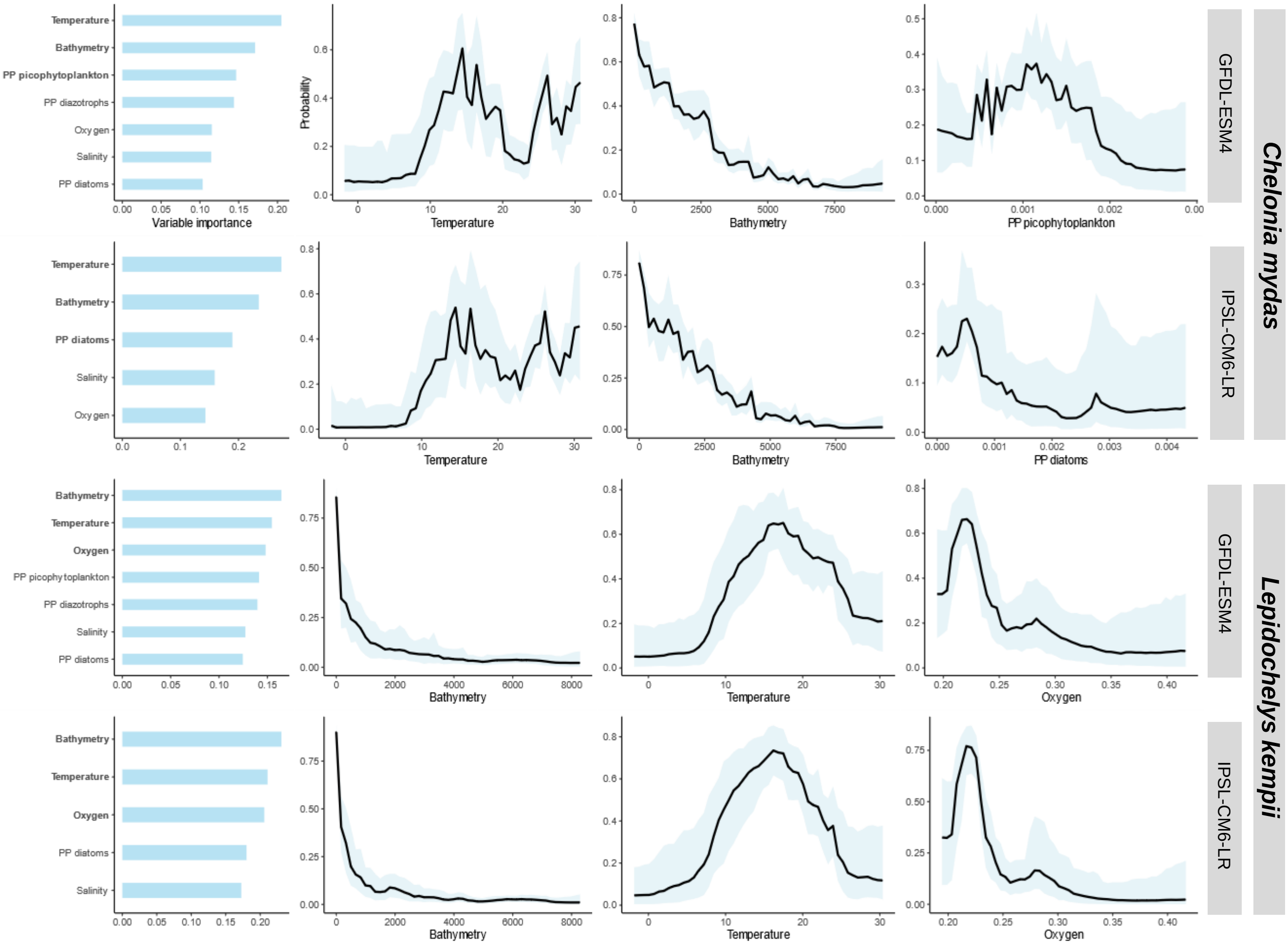


**Figure S8.** Results for the species *Chelonia mydas* and *Lepidochelys kempii*. The first column represents the contributions of all the variables to the model for both ESMs. We also provide the additive relation for the variables that have contributed the most to the model. These additive relations represent the probability of being present at some point along the x-axis.

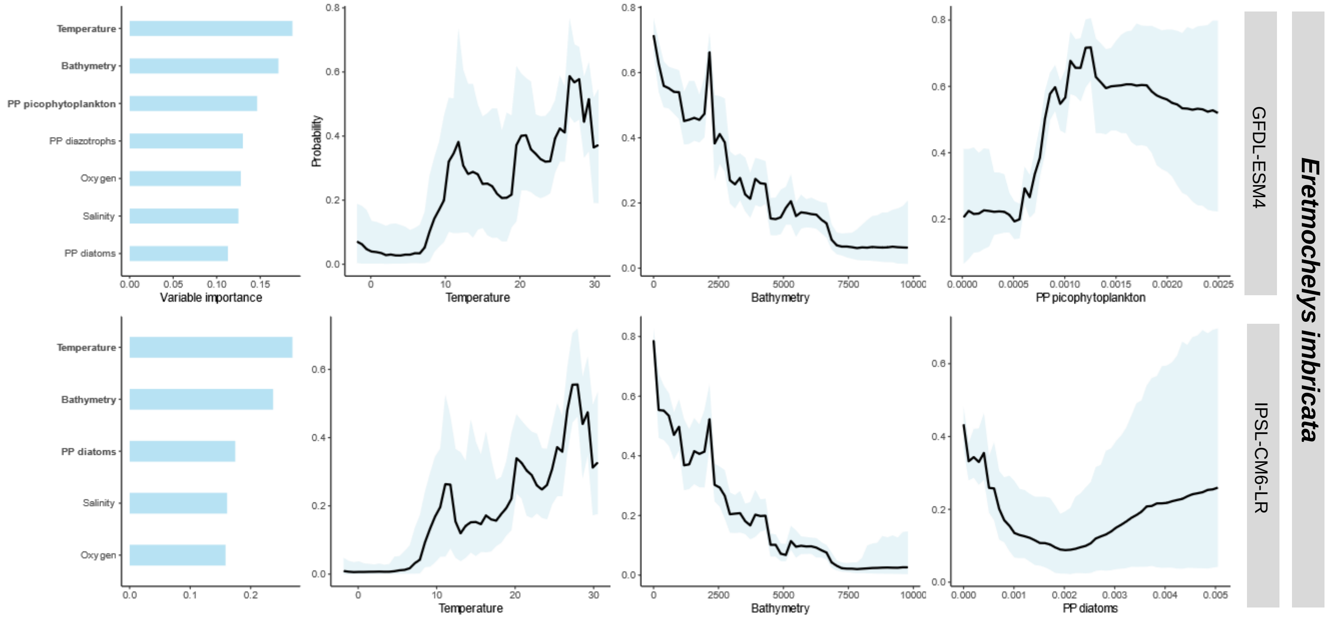


**Figure S9.** Results for the species *Eretmochelys imbricata*. The first column represents the contributions of all the variables to the model for both ESMs. We also provide the additive relation for the variables that have contributed the most to the model. These additive relations represent the probability of being present at some point along the x-axis.

## 3 Error measures

In this section, we present all the measures used in our study to compare the estimations of different regression models with respect to the simulated biomass of a fish stock.

### Specificity:

$$\mathrm{SPC} = \frac{TN}{N},$$

where *SPC* is the specificity or selectivity, $TN$ are the true negatives, and $N$ are the total negatives.

### Sensitivity

$$\mathrm{SEN} = \frac{TP}{P},$$

where *SEN* is the sensitivity or probability of detection, $TP$ are the true positives, and $T$ are the total positives.

### Accuracy:

$$\mathrm{ACC} = \frac{TP + TN}{P + N},$$

where *ACC* is the accuracy and the remain terms are those specify in the previous equations.

### $\mathrm{F}_1$ score

$$\mathrm{F}_1 \text{ score} = \frac{2 \times TP}{2 \times TP + FP + FN},$$

where $FP$ are false positives, $FN$ are false negatives, and the remaining terms are those specified in the previous equations.