Preprint for review (2023), 1–26

ARTICLE

Predicting species distributions in the open oceans with convolutional neural networks.

Gaétan Morand[*,1], Alexis Joly[2], Tristan Rouyer[1], Titouan Lorieul[2], and Julien Barde[1]

[1]UMR Marbec, IRD, Univ Montpellier, CNRS, Ifremer
[2]LIRMM, INRIA, Univ. Montpellier

*Corresponding author: gaetan.morand@ird.fr

(Preprint published on bioRxiv and submitted to PCI Ecology for peer-review August 11th, 2023)

Abstract

As biodiversity plummets due to anthropogenic disturbances, the conservation of oceanic species is made harder by limited knowledge of their distributions and migrations. Indeed, tracking species distributions in the open ocean is particularly challenging due to scarce observations, and the complex and variable nature of the ocean system. In this study, we propose a new method that leverages deep learning, specifically convolutional neural networks (CNNs), to capture spatial features of environmental variables. This novelty eliminates the need to predefine these features before modelling and creates opportunities to discover unexpected correlations. Our aim is to present the results of the first trial of this method in the open oceans, discuss limitations, and provide feedback for future improvements or adjustments.

In this case study, we considered 38 taxa which include pelagic fishes, elasmobranchs, marine mammals, as well as marine turtles and birds. We trained a model to make probability predictions from the environmental conditions at any specific point in space and time, using species occurrence data from the Global Biodiversity Information Facility (GBIF) and environmental data from various sources. These variables included sea surface temperature, chlorophyll concentration, salinity, and fifteen others.

During the testing phase, the model was applied to environmental data at locations where species occurrences were recorded. The model accurately predicted the observed taxon as the most likely taxon in 69% of cases and included the observed taxon among the top three most likely predictions in 89% of cases. These findings show the adequacy of deep learning for species distribution modelling in the open ocean and demonstrate the relevance of CNNs for prospective modelling of the impacts of future ocean conditions on oceanic species.

Additionally, this black box model was then analysed with explicability tools to understand which variables had an influence on the model's predictions. While variable importance was species-dependent, we identified finite-size Lyapunov exponents (FSLEs), sea surface temperature, pH, bathymetry and salinity as the most influential variables, in that order. These insights can prove valuable for future species-specific movement ecology studies.

Keywords: deep learning; megafauna; open oceans; pelagic species; species distribution models

1. Introduction

1.1 Background

The open ocean is a vast and complex ecosystem that covers over 70% of the Earth's surface, yet it remains one of the least understood and studied ecosystems on our planet [1]. It plays a critical role in regulating the Earth's climate, nutrient cycles, and biogeochemical cycles (including carbon sequestration), making it a vital component of all life on Earth [2].

However, the ocean is facing a range of human-induced threats, including over-fishing, pollution, and climate change [3, 4, 5]. These threats can have serious consequences for marine biodiversity, and therefore negatively impact the livelihoods of millions of people who depend on the oceans for their food or income [6].

To focus on solving the most pressing challenges, it is essential to understand how marine life is distributed within open oceans. Species distribution models can provide valuable insights into where different species are likely to be found, what environmental factors are driving their distribution, and the influence of temporal variations [7]. By developing accurate and reliable models, we can identify areas that are most at risk and develop effective conservation strategies to protect these ecosystems.

Furthermore, changes in the Earth's climate are already affecting ocean conditions, namely warming waters, ocean acidification, and sea level rise, among others [8]. This makes it even more urgent to understand the link between environmental variables and species distributions, to be able to predict how marine biodiversity may respond to these changes. This information is critical for informing decision-making and management efforts to ensure the long-term sustainability of marine ecosystems and the services they provide to society.

Therefore, studying species distribution in open oceans is essential for advancing our understanding of these complex ecosystems and for developing effective conservation and management strategies to protect them.

1.2 Existing methods for predicting species distributions

There is a wide variety of Species Distribution Models (SDMs) discussed in literature [9]. This is generally done through modelling a species-specific *environmental niche*: the area where environmental conditions are favourable to the species [10]. Predictors are chosen empirically to best fit the observed species' occurrences with specific environmental conditions.

Usually, SDMs use environmental data at the exact location where the prediction is computed, which is insufficient to represent the full nature of the environmental seascape around animals. Even averaging data over a buffer area does not fully represent the environmental conditions, as it cannot convey bathymetry features such as seamounts or trenches. The same applies to other variables, which spatial structure may be more important than punctual or average values: algal blooms, temperature fronts, eddies, etc. Yet these spatial structures are essential to understanding species distributions [11, 12].

A solution to this shortcoming is to include the values of these environmental data in the neighbourhood of species occurrences, but the number of variables then becomes much larger than the number of observations. This is unfit for statistical models and requires a feature extraction step to summarize input data as fewer significant variables. This work may be made manually, which enables the model to take advantage of scientists' expert knowledge. This is how some spatial features are added into SDMs [13], but it limits the performance of the model to the scope of existing knowledge and prevents the discovery of previously unknown influential factors.

Furthermore, SDMs rarely take into account the high temporal variability of environmental data [14], which seriously hinders the prediction of highly mobile species distributions.

This calls for new methods to fully take into account the complex spatial structures of environmental seascapes and their influence on species distributions.

1.3 Potential benefits of using deep learning for modelling marine species distribution

Convolutional neural networks (CNNs) were invented for image recognition, so they have embedded feature extractors that are designed to detect multiple levels of details using convolution layers [15]. With image classification, one can identify the following levels, from most precise to coarser:

1. Values of specific pixels
2. Value of a small group of pixels: textures, edges
3. Association of several groups of pixels: shapes, geometric features
4. Association of several shapes: objects, animals, plants
5. Average and extreme values on the whole image: brightness/tint

This is especially useful with environmental data raster layers (from satellite observations or models) as it enables the model to detect the same various levels of details on environmental variables. Here are some examples of the same levels of detail, applied to environmental variables :

1. Values at the point of occurrence
2. Homogeneity of the variable in the neighbourhood of occurrence: temperature fronts, slopes
3. Geographic features: bays, underwater canyons, river plumes
4. Complex shapes: current structures, cyclones
5. Average and extreme values over the buffer zones

The use of CNNs to model species distributions was successfully developed for terrestrial plants [16]. The CNN architecture proved especially useful to capture spatial features, as well as to transfer knowledge from better-known species to lesser-known species [17]. While these studies were mostly based on satellite imagery (Sentinel-2), we cannot rely on this information in the open oceans. Another significant difference is the high temporal variability of the oceanic seascape. This is why we present an adaptation of their work that relies only on environmental data.

1.4 Objectives of the study

Through the present study, we explore and report the possibilities that deep-learning-based SDMs offer in the open oceans. We first give a detailed overview of the data that was used to build our model. Then we show the results that we obtained, including performance metrics and distribution maps. Finally, we show an example of how this method may be used to simulate a change in future ocean conditions and how probabilities predictions respond. We also point out the limits that we have found with our methodology choices and suggest ways to improve the results' quality in the future.

2. Methods

The main step of our process is to build a model to link environmental data to species presence. To achieve this, we used freely available occurrence data from the Global Biodiversity Information Facility (GBIF) [18], and downloaded environmental data in a buffer around each of their locations, at the date of their occurrence.

After training, this provided us with a model which takes environmental data as input and outputs a vector of observation probabilities (one for each taxon). It is important to note that as training data

is presence-only, we cannot predict abundance or any absolute measure of presence. The predictions are observation probabilities, relative to the 38 taxa that are the subject of this study. The full process is summarized in Figure 1 and is explained in detail in this section.

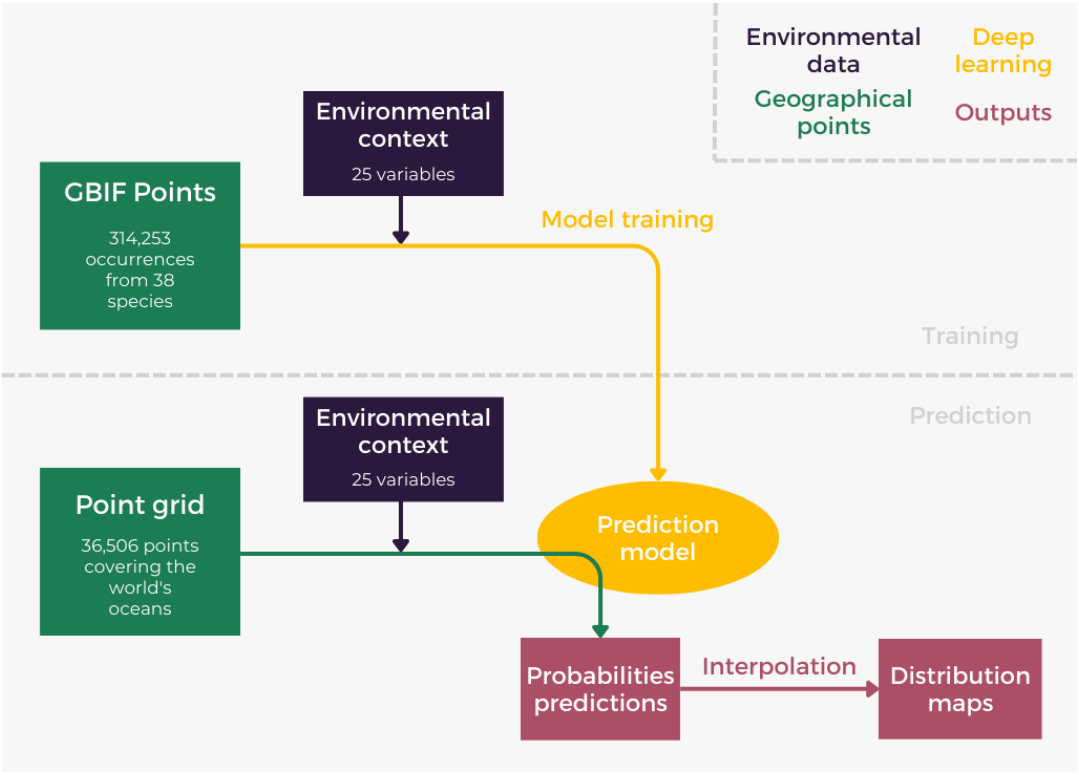


Figure 1. Summary view of the analysis process: model training in the top half and predictions in the bottom half.

2.1 Description of the occurrence data

Thirty-eight marine species or genera were selected for the proof of concept that is described in the present article. They include large pelagics, elasmobranchs, turtles, sea mammals, and two species of marine birds (*cf.* table 1). These taxa may be replaced or complemented with others in the future.

Presence data for these taxa were downloaded from the GBIF. This database contains species occurrences that are free to access and download, which is essential to reproducibility. Some flawed data is unavoidably present in the database, but small errors in the geographical coordinates are not a problem as the oceanographic landscape that we consider has limited precision due to environmental data resolution. We removed points located on the continents, and no other filtering was conducted on geographical precision. Furthermore, convolutional neural networks are known to be robust against occasional labelling mistakes [19].

Digital object identifiers (DOIs) for the download of each species are available in table 1. When there were more than 10,000 occurrences of a taxon, a random sample of 10,000 occurrences was selected. In all cases, GBIF identifiers of occurrences that were actually used are available in the training data set CSV files (*id* column).

This added up to 314,253 occurrences for all taxa, depicted in figure 2.

Table 1. Species that were included in the study, coloured by taxonomic class. The last column is the digital object identifier (DOI) of downloaded archives.

English name	Taxonomic name	DOI
Yellowfin tuna	*Thunnus albacares*	10.15468/dl.gr2wbb
Longfin tuna	*Thunnus alalunga*	10.15468/dl.aqjv3y
Atlantic bluefin tuna	*Thunnus thynnus*	10.15468/dl.nnyeyb
Southern bluefin tuna	*Thunnus maccoyii*	10.15468/dl.tw97qj
Bigeye tuna	*Thunnus obesus*	10.15468/dl.c96qpp
Skipjack tuna	*Katsuwonus pelamis*	10.15468/dl.6y2zzm
Frigate Tuna	*Auxis thazard*	10.15468/dl.kfm6kq
Sailfish	*Istiophorus*	10.15468/dl.f48dug
Black marlin	*Istiompax indica*	10.15468/dl.b5acky
Blue marlin	*Makaira*	10.15468/dl.sygtaw
Swordfish	*Xiphias gladius*	10.15468/dl.hazqd2
Dolphinfish	*Coryphaena*	10.15468/dl.q67bqt
Humphead wrasse	*Cheilinus undulatus*	10.15468/dl.9g76hq
Oceanic Whitetip	*Carcharhinus longimanus*	10.15468/dl.b5ws4q
Whitetip	*Carcharhinus albimarginatus*	10.15468/dl.vpc772
Silk shark	*Carcharhinus falciformis*	10.15468/dl.vg4rwh
Sandbar shark	*Carcharhinus plumbeus*	10.15468/dl.7fczpa
Grey reef shark	*Carcharhinus amblyrhynchos*	10.15468/dl.ccqyws
Mako shark	*Isurus oxyrinchus*	10.15468/dl.h5akxk
Blue shark	*Prionace glauca*	10.15468/dl.zqkssk
Devil ray	*Mobula mobular*	10.15468/dl.p4e2sx
Reef manta	*Mobula alfredi*	10.15468/dl.bkjkgu
Eagle ray	*Myliobatis*	10.15468/dl.3u3v7k
Humpback whale	*Megaptera novaeangliae*	10.15468/dl.yzg4n3
Fin whale	*Balaenoptera physalus*	10.15468/dl.r9kaq8
Blue whale	*Balaenoptera musculus*	10.15468/dl.28f7xd
Bottlenose	*Tursiops*	10.15468/dl.bec9p4
Spinner dolphin	*Stenella longirostris*	10.15468/dl.xz5eds
Common dolphin	*Delphinus delphis*	10.15468/dl.u5be7v
Sperm whale	*Physeter macrocephalus*	10.15468/dl.7pf4ue
Harbour porpoise	*Phocoena phocoena*	10.15468/dl.afr2fn
Southern right whale	*Eubalaena australis*	10.15468/dl.e3hdkj
Green turtle	*Chelonia mydas*	10.15468/dl.6gs9rp
Loggerhead	*Caretta caretta*	10.15468/dl.dmb6ds
Hawksbill turtle	*Eretmochelys imbricata*	10.15468/dl.e6w44w
Emperor penguin	*Aptenodytes forsteri*	10.15468/dl.s5unhs
Wedge-tailed shearwater	*Puffinus pacificus*	10.15468/dl.vyztue
Acropora coral	*Acropora*	10.15468/dl.vg752f

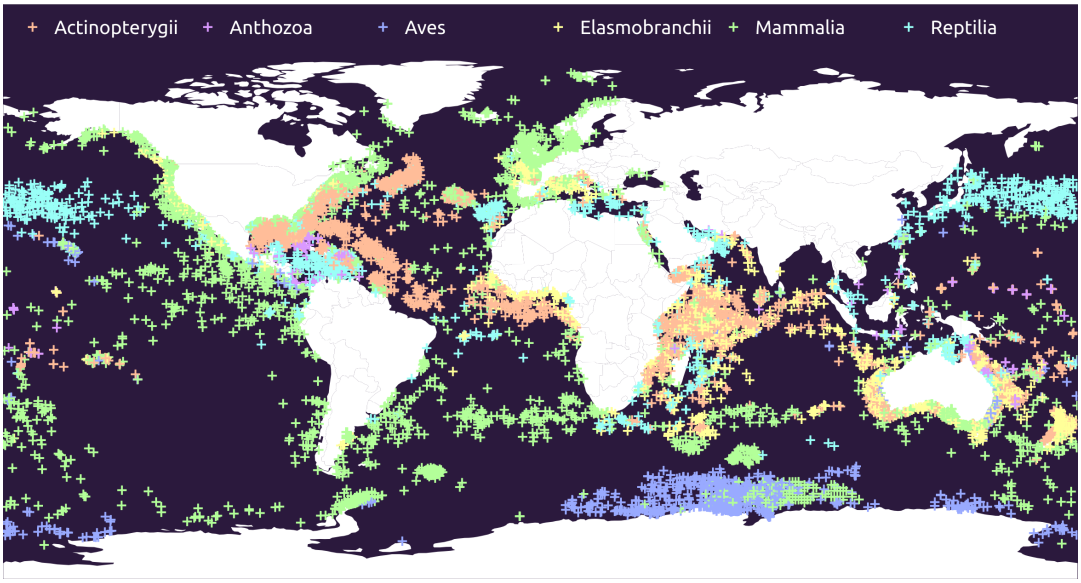


Figure 2. Random sample (10%) of the training data set, coloured by taxonomic class.

2.2 Description of the environmental data used as inputs

Eighteen environmental variables were considered, some from satellite observations and others from models. Three of them contain two components: surface wind, geostrophic current and finite-size Lyapunov exponents (FSLEs). Temperature and chlorophyll values were also included 15 and 5 days before the occurrences, as it was previously demonstrated that marine animals may have a delayed response to some variables, especially temperature [20]. Finally, four geographical variables were added (cf. section 2.3.2). This amounts to a total of 29 layers of data, shown in table 2.

2.3 Data preparation

2.3.1 Enrichment

Environmental data were downloaded in a buffer around the occurrences using the GeoEnrich python package, which was developed for this purpose and made available to other researchers for a wide range of uses in the GitHub IRDG2OI/geoenrich repository [21]. A spatial buffer of 115km was used, to include at least one data point from the least precise data (2° resolution). This is consistent with values of daily potential movement for fast animals that may travel up to 120 km per day [22, 23].

These data arrays with various resolutions (minimum 1×1 for wave height, maximum 241×241 for bathymetry) also have various horizontal dimensions due to the longitude contraction closer to the poles. They were all interpolated (up-scaled or down-scaled depending on the initial resolution) to fit the same 32×32 grid centred around the occurrence.

2.3.2 Ocean basin and hemisphere

An initial goal of the study was to produce a geography-agnostic model, which means that two points with the same oceanic conditions, wherever they are, should yield the same predictions. But this is ecologically wrong for one main reason: there are natural barriers that animals cannot cross, namely continents and for some species, the warm waters around the equator. Therefore we added four binary variables: three for the main oceanic basins and one for the hemisphere.

The world's oceans were split into three main basins: the Atlantic, Indian and Pacific oceans. Very few of the occurrences were located in the Arctic Ocean: they were assigned the closest of these ocean basins. Occurrences from the Southern Ocean were more numerous and are not separated from these three oceans by any physical barrier, so they could be assigned to the closest one.

It is important to note that the Ocean basin and the hemisphere are the only geographical information provided to the model. This is by design to avoid learning the observation bias that is present in the training data.

Table 2. 29 layers used as input data

Variable	Source	Source type
Bathymetry	GEBCO [24]	Observations
Salinity	Copernicus [25]	Observations
Wave Height	Copernicus [26]	Observations
Surface wind (u)	CCMP [27]	Observations
Surface wind (v)	CCMP [27]	Observations
Oxygen	Copernicus [28, 29]	Models
pH	Copernicus [28, 29]	Models
FSLEs (strength)	Aviso [30]	Observations
FSLEs (orientation)	Aviso [30]	Observations
Geostrophic Current (u)	Copernicus [31, 32]	Observations
Geostrophic Current (v)	Copernicus [31, 32]	Observations
Eddy kinetic energy	Calculated	
Chlorophyll	OCCI [33]	Observations
Sea surface temperature	MUR [34]	Observations
Mixed layer thickness	Copernicus [25]	Observations
Diatoms	Copernicus [35]	Observations
Dinophytes	Copernicus [35]	Observations
Haptophytes	Copernicus [35]	Observations
Green algae	Copernicus [35]	Observations
Prochlorophytes	Copernicus [35]	Observations
Prokaryotes	Copernicus [35]	Observations
Chlorophyll (D-15)	OCCI [33]	Observations
Sea surface temperature (D-15)	MUR [34]	Observations
Sea surface temperature (D-5)	MUR [34]	Observations
Chlorophyll (D-5)	OCCI [33]	Observations
Atlantic Ocean	Calculated	
Indian Ocean	Calculated	
Pacific Ocean	Calculated	
North hemisphere	Calculated	

2.3.3 Feature scaling

All data were scaled to the $[0, 1]$ interval, and saved into a data cube (32×32 geographical pixels $\times 29$ layers). Outliers (highest and lowest 1% of the values of the training data set) were clipped and the scaling factors were saved to be reapplied to any subsequent input data.

There are some missing data because of natural phenomena such as clouds, or because the occurrences were out of the data set time range. In that case, we used the median value of the variable over the tile. If data was missing over the whole tile, we used the median value over the whole data set instead.

Figure 3 shows an example of all the data that are included in the data cube used for training, with the feature scaling reversed in order to show the real values. The figure does not show the four binary geographical variables.

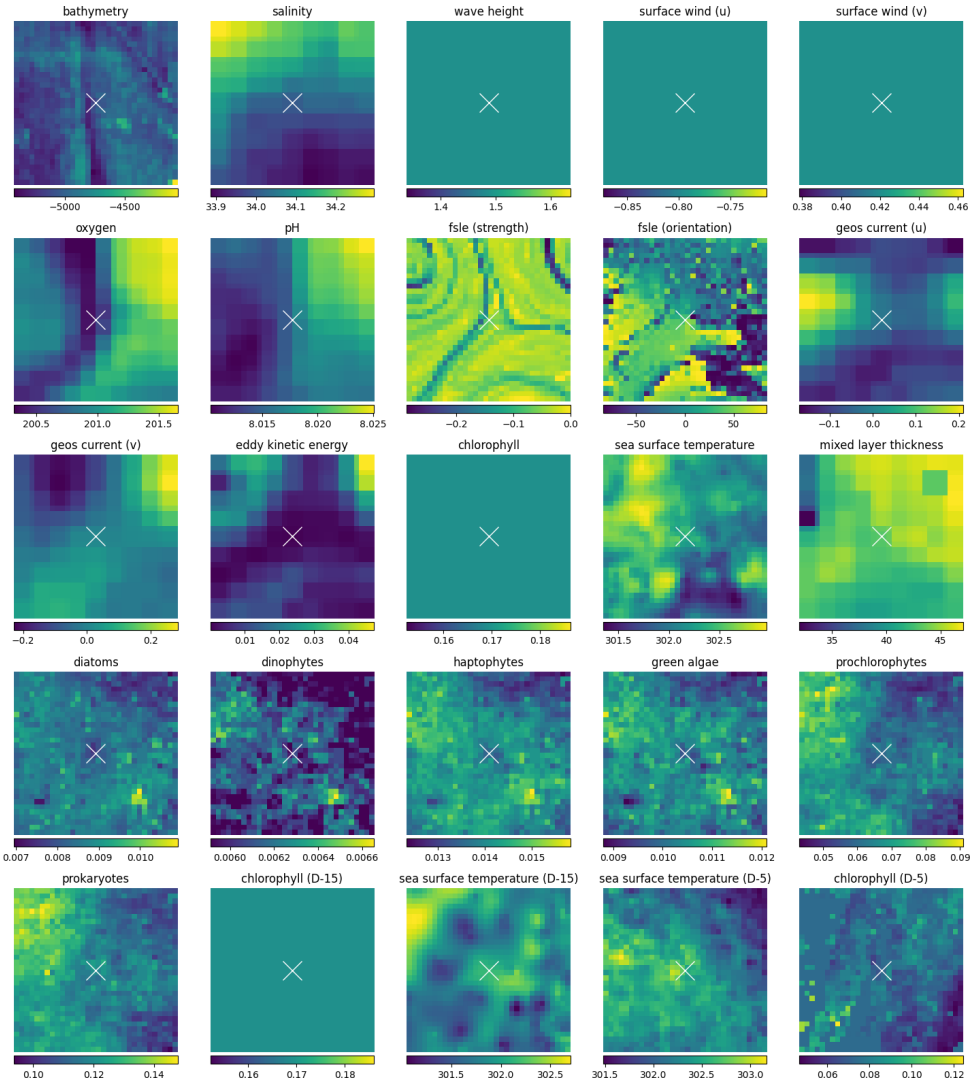


Figure 3. Environmental variables around the point of coordinates -14.389°S, 78.918°E on March 20th, 2021.

2.4 Training the model

The modelling technique that we describe in this study was developed for plant species distributions [17]. We used the *Malpolon* framework [36] after some adaptations to our use case. It was built on top of PyTorch [37] and PyTorch Lightning [38] frameworks.

The *Malpolon* framework implements a convolutional network with the *resnet50* feature extractor [15]. We adapted it to use 29 inputs channels, and 38 numerical outputs converted to relative probabilities by a Softmax function. It was trained in two sessions by minimizing the Binary Cross Entropy loss: one with a .1 learning rate, and another one with a .01 learning rate to fine-tune the weights.

The target probabilities for training were set using one-hot encoding, *ie.* a one for the observed species and zeroes for all others. This follows the principle of assumed negatives [39]: we assume that only the observed species is present at the point of observation. This equates to considering the pseudo-absences of the other species. This is obviously wrong, but as we work with a limited number of species in an extensive area and period of time, chances are slim that the model receives contradictory information.

2.5 Evaluation metrics and performance assessment

Training data were randomly split into three sets: training (60%), validation (20%) and test (20%). The validation set was used to assess and improve performance during the training phase, while the test set was used after training to compute the final performance of the model, on data that it had never seen before.

The accuracy of the final version of the model was 69%, which means that in 69% of cases, the most likely taxon according to the model was the same as the one that was actually observed. See table 3 for more complete accuracy results.

Table 3. Probability that the observed taxon is among the Top N predictions of the model, for 11 values of N

Top N	Probability
1	69.15%
2	83.19%
3	89.13%
4	92.83%
5	95.13%
6	96.63%
7	97.48%
8	97.99%
9	98.43%
10	98.75%
38	100.00%

A confusion matrix was computed on the test data set and is shown in Figure 4. It shows that some taxa were easily identified by the model (the top two being *Aptenodytes forsteri* and *Mobula alfredi*). Others were harder to predict, the worst two being *Istiompax indica* and *Carcharhinus longimanus*.

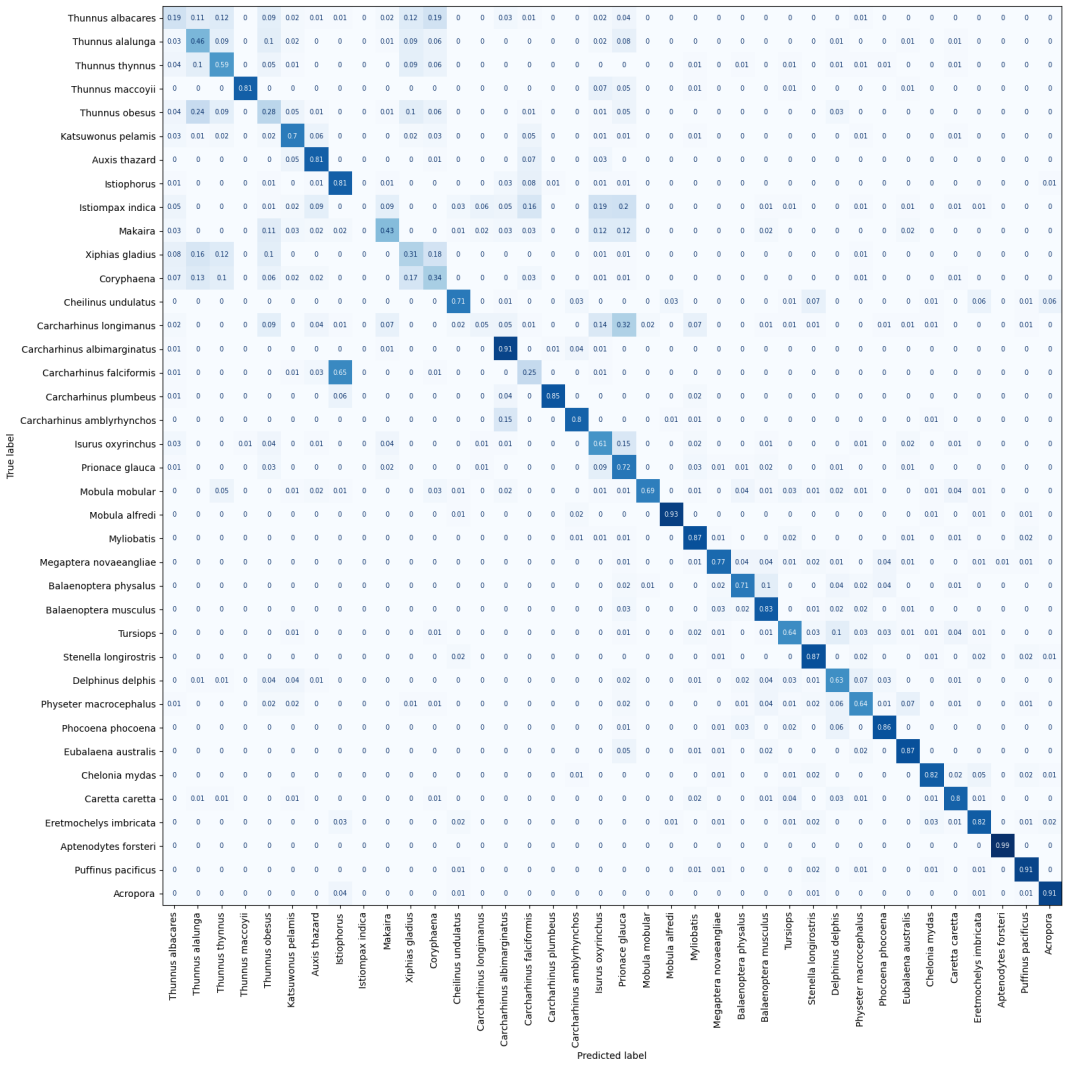


Figure 4. The confusion matrix shows the top predictions of the model on the test data set, for each observed taxon.

2.6 From probability predictions to distribution maps

After training the model, we used it on new data to generate distribution maps. As the environmental data download phase can be quite slow, we had to choose to focus either on time extent or spatial extent, but not both at the same time. This is why we chose to compute two different outputs:

- Global species distribution maps at four dates in 2021.
- Regional species distribution maps for the Southwestern Indian Ocean at 53 dates in 2021.

It is important to note that the model may be used at any date in any area; the only limitation is the availability of environmental data, and the time required to download them.

Two grids covering both areas were generated. They comprised 36,506 points for the global oceans and 3,001 for the Southwestern Indian Ocean. Environmental data were downloaded for each of these points and run through the model, which led to 38 relative probability predictions for each of

these points, at each requested date. For each taxon, these probabilities were interpolated over the whole area to generate rasterized distribution maps.

It is worth noting that since we are working with relative probabilities (because of the softmax layer), the absolute values have little purpose. Therefore no scale is provided for all distribution maps: they should be interpreted relatively to one another, across species, time or space.

2.7 Influence of variables

To study the influence of variables, a new model was trained after removing chlorophyll and sea surface temperature at D-5 and D-15, as well as Eddy Kinetic Energy. Indeed, these layers are highly correlated with chlorophyll and sea surface temperature on the day of occurrence, and geostrophic current respectively.

It is worth noting that this model has almost the same accuracy (69.08%) as the previously described one (69.15%), which shows that the 5 variables that were removed have very little influence on the classification.

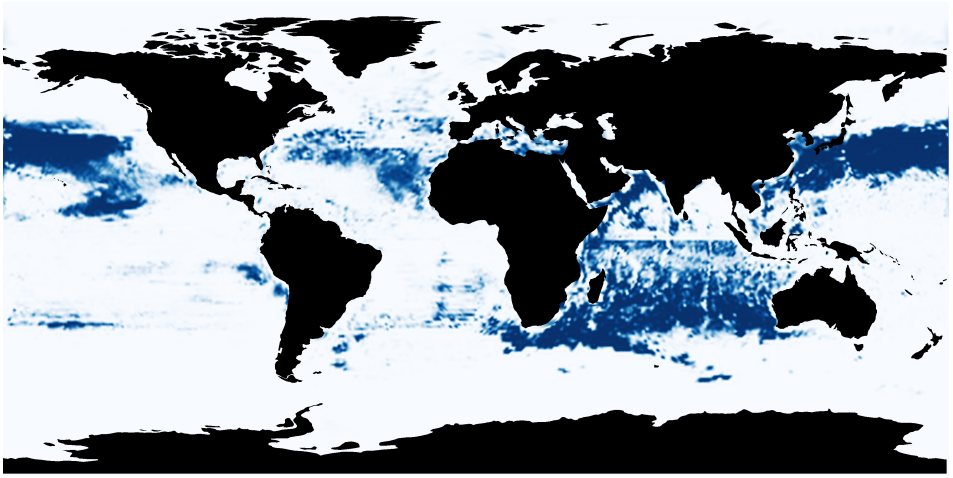
Afterwards, the most determining variables were calculated using the integrated gradients method for a random sample (N=1,000) of the points on the world grid [40], using the Captum python package [41]. They were then aggregated over the geographical area (sum of absolute values) and these values were grouped by taxon (the top prediction).

3. Results

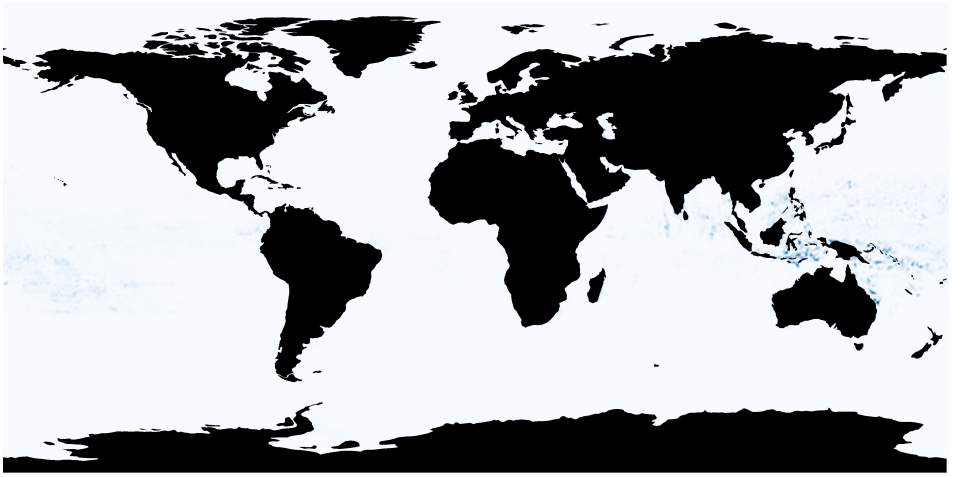
3.1 Presentation of the species distributions maps

3.1.1 Global oceans

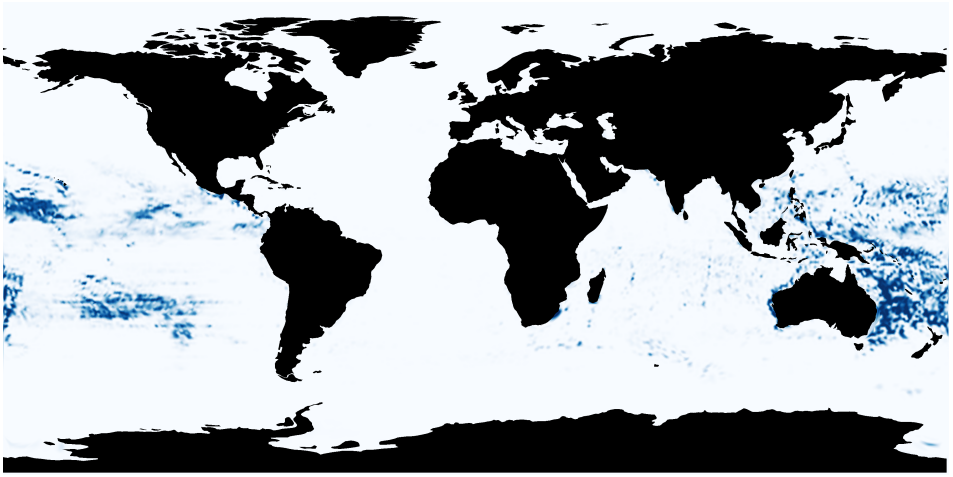
Distribution maps were calculated on four dates, all in 2021, corresponding to both solstices and both equinoxes, for the thirty-eight taxa. Figure 5 shows these maps for three species on the spring equinox. All 152 distribution maps are available online [42].



(a) *Caretta caretta*



(b) *Mobula alfredi*



(c) *Puffinus pacificus*

Figure 5. Examples of distribution maps on March 20th, 2021.

3.1.2 Southwestern Indian Ocean

In this case, distribution maps were calculated each week of 2021, for the 38 taxa. To make visualization easier, they were exported as animated GIFs that are available online [43]. An example for *Prionace glauca* is shown in Figure 6.

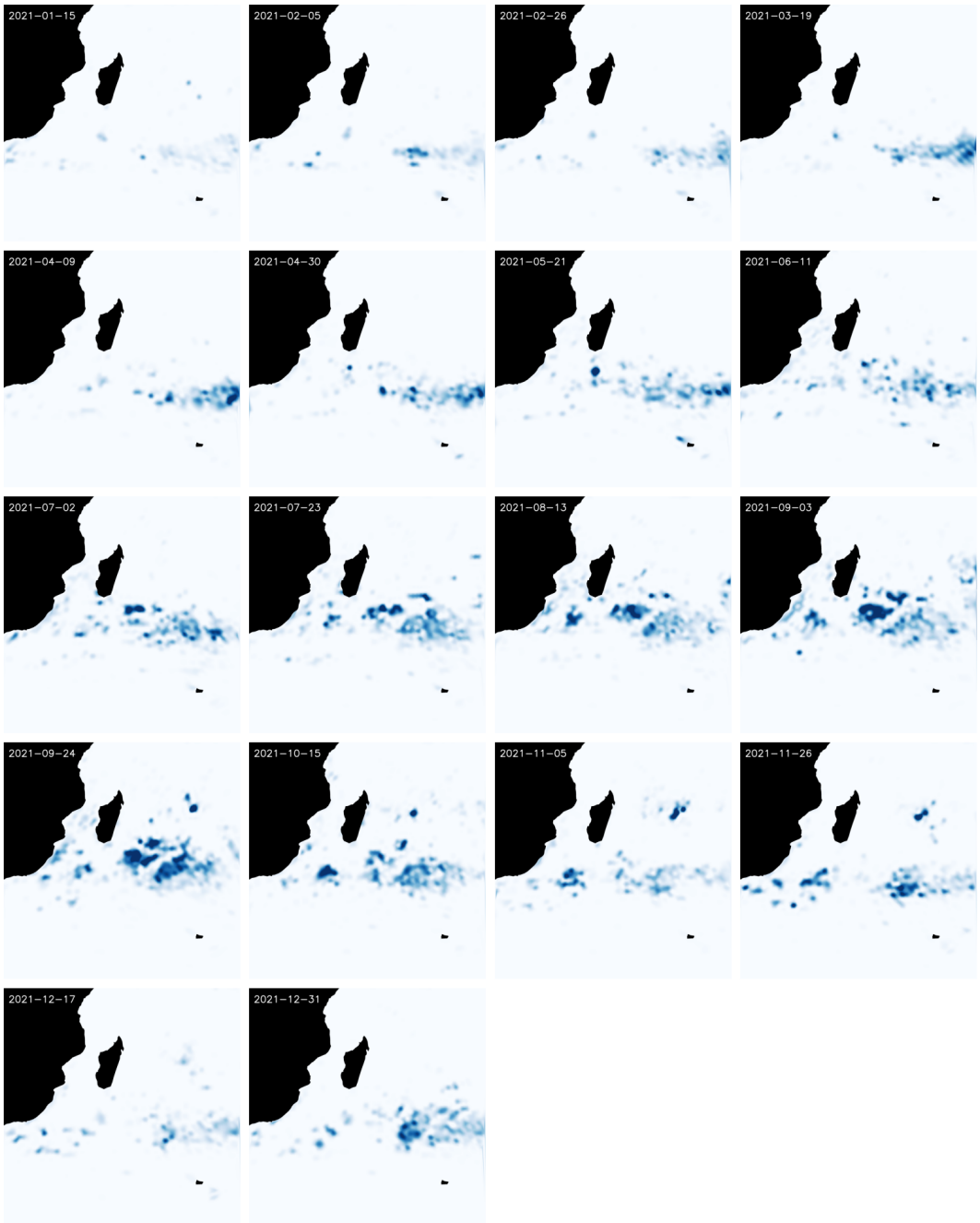
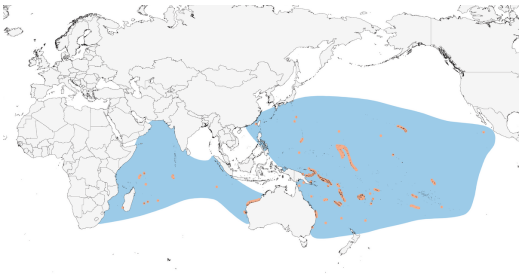


Figure 6. Distribution maps for *Prionace glauca* every three weeks of 2021. The contrast was increased to improve visibility: colours from this figure should not be compared to other figures.

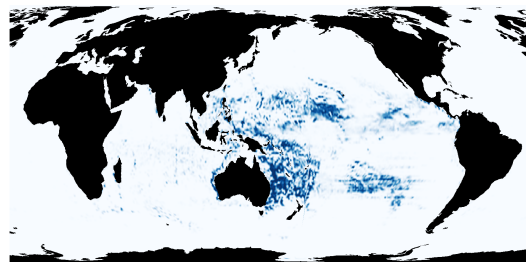
3.2 Comparison of predicted distribution maps to established maps

Validation of the distribution maps is challenging because existing distribution maps rarely take into account temporal variability, except sometimes broad seasonal variations. Yet the maps that we produce are highly dependent on time, see Figure 6 for instance.

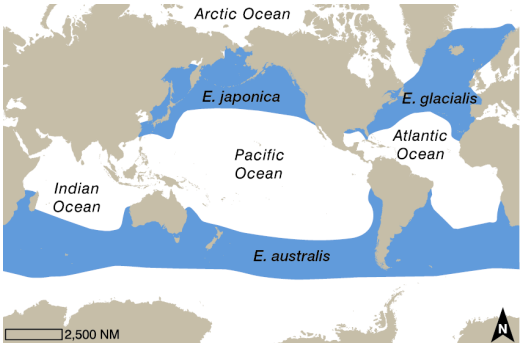
We compared some of our distribution maps to established ones, to check for inaccuracies. See Figure 7 for a few examples.



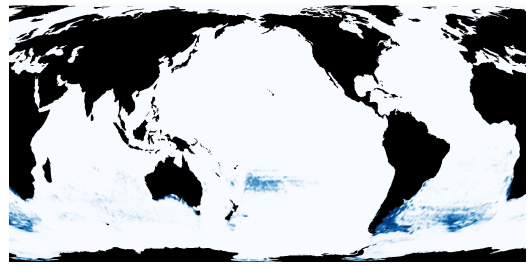
(a) *Puffinus pacificus* (established map [44])



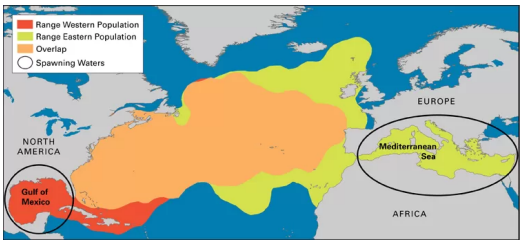
(b) *Puffinus pacificus* (predicted)



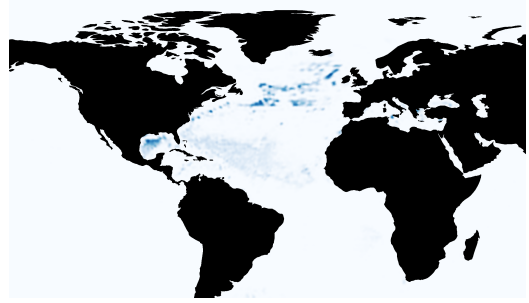
(c) *Eubalaena australis* (established map [45])



(d) *Eubalaena australis* (predicted)



(e) *Thunnus thynnus* (established map [46])



(f) *Thunnus thynnus* (predicted)

Figure 7. Comparison between established distribution maps and deep-learning generated maps

Puffinus pacificus The prediction map 7a is coherent with the established one 7b, although there is a significant difference between the Indian and Pacific oceans. Since established maps are usually binary (presence/absence), the difference we see between the two oceans cannot be (in)validated. It

is possible that the Pacific Ocean is more suitable to this species than the Indian Ocean, or there may be an under-representation of occurrences in our data set.

Eubalaena australis The predicted distribution 7d fits within the known geographical range of *Eubalaena australis* 7c, and there is a strong disparity of the prediction density within this area. Again, no assumption can be made on the plausibility of the predictions, as this heterogeneity may be caused by temporal variation, or it may not fit reality. Another possibility is that the established geographical range is not fully used by the species.

Thunnus thynnus The predicted range for *Thunnus thynnus* 7f is within the established range 7e, but it does not include all of it. Specifically, the Mediterranean Sea and the Bay of Biscay are excluded, even though there is a major population living in these areas [47]. After checking our input data, this shortcoming can be explained by the under-representation of this population in the occurrences used for training. This will be discussed further in section 4.2.2.

3.3 Analysis of determining variables

Over the predictions for the 2021 Global use case, the most influential variables were FSLEs (strength), sea surface temperature, pH, bathymetry, salinity, and FSLEs (orientation), in this order. See table 4 for a full accounting of variable influence.

Figure 8 shows the median integrated gradient for each taxon. Note that *Istiompax indica* is not present in this chart as it was never predicted to be the most likely taxon.

3.4 Effect of a 2°C increase in sea surface temperature

Predictions were computed after adding 2°C to sea surface temperature, leaving all other variables unchanged. In the context of climate change, this is a tentative projection but it is theoretical, as there are significant and complex correlations between future changes in various environmental variables. Our model does not provide abundance results, but relative probabilities indicate which taxa benefit or suffer from this change.

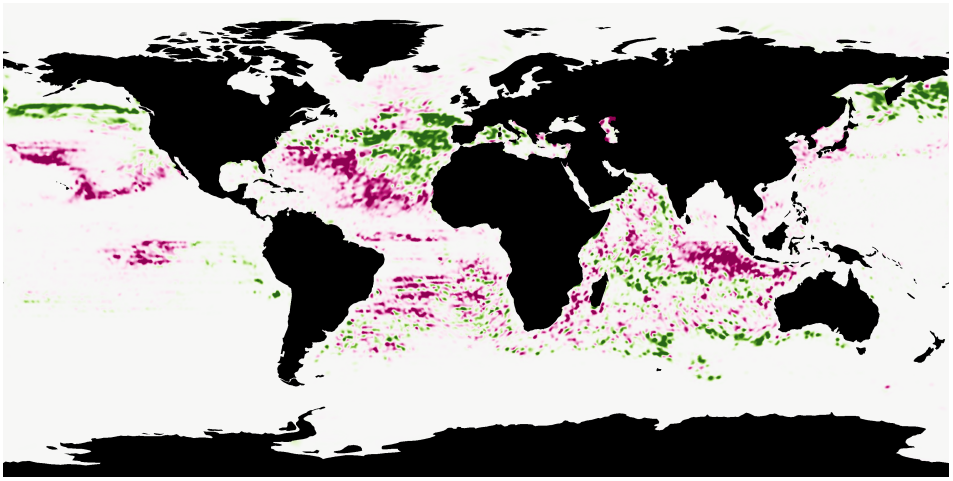
There is also a change in presence areas, illustrated in Figure 9. For *Caretta caretta* and, to a smaller extent, *Eubalaena australis*, there is an expected poleward shift: away from the equator and towards cooler water, as projected previously [48]. The evolution of the *Katsuwonus pelamis* range is harder to interpret: this species may not be that sensitive to warmer water because it already lives in the equatorial band, so the change may be caused by the interaction between temperature and other variables. In reality, a temperature rise would cause a decrease in oxygen levels (not modelled here), which might drive the *Katsuwonus pelamis* population away from the equator [49].

Evolution maps were calculated by subtracting the +2°C projection from the corresponding 2021 predictions, for each geographical point. This process creates new raster layers with a -1 to 1 range reflecting the evolution of predicted probabilities.

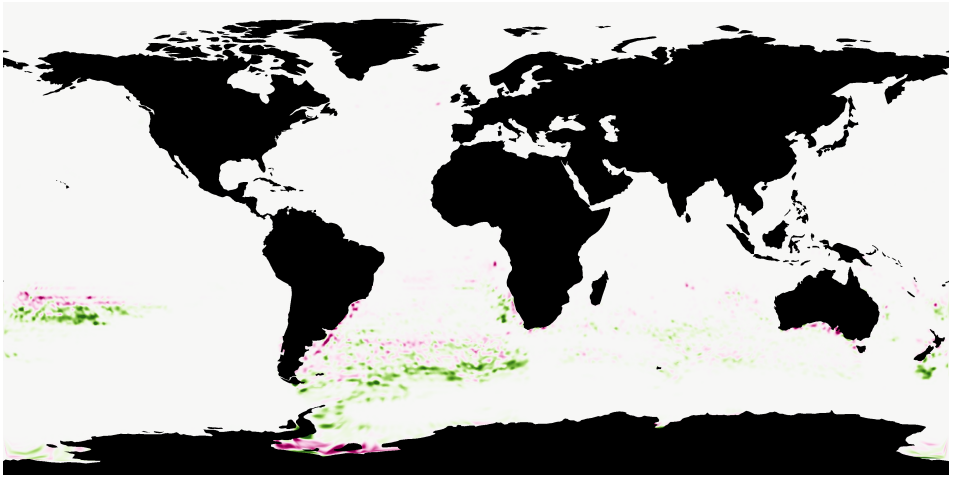
Table 4. Statistics of the influence of variables on a sample of 1,000 predictions ($\times 1,000$, sorted by median).

Variable	mean	std	min	25%	50%	75%	max
FSLEs (strength)	1.08	0.8	0.12	0.53	0.8	1.36	3.87
SST	0.77	0.71	0	0.24	0.56	1.1	3.66
pH	0.64	0.62	0	0.25	0.41	0.8	3.35
Bathymetry	0.67	0.76	0	0.17	0.39	0.85	5.25
Salinity	0.6	0.53	0	0.24	0.38	0.75	2.83
FSLEs (orientation)	0.53	0.42	0.04	0.24	0.37	0.64	2.62
Geos. current (u)	0.4	0.34	0	0.18	0.28	0.49	2.15
Geos. current (v)	0.39	0.31	0.05	0.18	0.28	0.49	1.56
Surface wind (v)	0.37	0.28	0.04	0.17	0.26	0.45	1.36
Oxygen	0.42	0.45	0	0.06	0.24	0.66	3.04
Surface wind (u)	0.33	0.26	0.03	0.15	0.23	0.4	1.21
Pacific Ocean	0.41	0.56	0	0	0.23	0.63	3.11
Wave height	0.18	0.14	0.02	0.08	0.13	0.22	0.7
Mixed layer thickness	0.26	0.37	0	0.05	0.12	0.34	2.83
Prochlorophytes	0.07	0.16	0	0.02	0.03	0.08	2.9
Green algae	0.08	0.15	0	0.01	0.03	0.08	2.23
Haptophytes	0.07	0.14	0	0.01	0.03	0.07	2.27
Prokaryotes	0.03	0.05	0	0.01	0.02	0.03	1.15
Chlorophyll	0.03	0.04	0	0.01	0.01	0.03	0.53
Dinophytes	0.04	0.08	0	0	0.01	0.03	0.83
Diatoms	0.03	0.07	0	0	0.01	0.03	0.72
Atlantic Ocean	0.19	0.43	0	0	0	0.16	2.6
Indian Ocean	0.25	0.6	0	0	0	0	3.28
North hemisphere	0.42	0.7	0	0	0	0.56	3.66

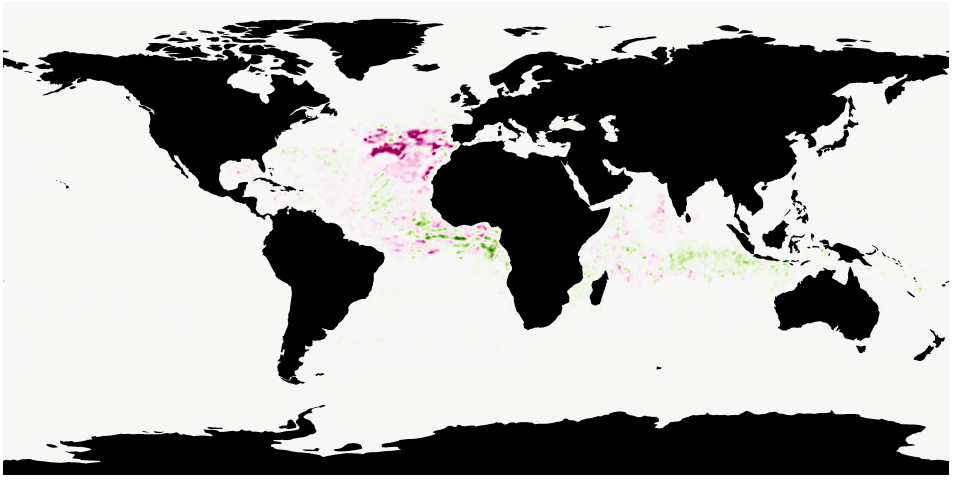
Figure 8. Variables that had the most influence on the determination of each taxon presence



(a) *Caretta caretta*



(b) *Eubalaena australis*



(c) *Katsuwonus pelamis*

Figure 9. Examples of evolution maps after adding +2°C. Green areas are more suitable with 2°C more, while pink areas are less suitable. Maps for all 38 species on all 4 dates are available online. [50]

4. Discussion

4.1 Ecological interpretation of the results, implications for offshore species distributions

The variables that were identified are coherent with past research. Specifically, FSLEs were identified as a particularly important movement predictor for top marine predators [51]. Sea surface temperature was also expected to be an important predictor, as it is the most frequently used descriptor in marine SDMs [14] and was identified as the most relevant factor in an SDM review [52].

This study also demonstrates a high sensitivity to temporal variations in environmental conditions, as shown in figure 6. This highlights the need for distribution models of fast-moving species to consider these variations, instead of relying only on averaged values..

There are some surprises in influential variables: bathymetry was not a good predictor of *Acropora* coral distribution, which is contradictory with their need for light. A possible explanation is that the model may have used other variables as a proxy for low depths.

4.2 Benefits and limitations of using deep learning for SDMs in the open ocean

This method holds promise in helping researchers uncover new correlations between the oceanic conditions and species distributions: implicit feature extraction allows the use of more numerous and more complex features. But there is a drawback: determining features have to be studied afterwards, and it is not straightforward. In this study, we showed the variables that had the most influence on average. This needs to be complemented by a deeper study of the nature of the determining features.

We noted three main limitations of our method, namely performance metrics, biases in the input data, and some undetected patterns.

4.2.1 Accuracy

The accuracy of the model could still be improved, depending on the ecological feasibility. Indeed, if some species are frequently seen together, there is no way for the model to discriminate between the two. In that case, this uncertainty will show as a .5 mistake rate even though it is the correct result. A way to improve the final accuracy score would be to group species by traits. But this would remove the possibility of studying differences between similar species.

For example, the confusion matrix in figure 4 shows that *Xiphias gladius* and *Coryphaena* are often predicted instead of each other. This may be the result of these two taxa having similar habitats, and the low resulting score does not necessarily mean that the predictions are wrong.

Accuracy is not an ideal metric for this use case, but the scarcity of training data is very limiting in that aspect. This is why we provided a Top-N score in table 3 for a more complete performance assessment.

4.2.2 Observer bias

Most observation data in the open ocean come from fishing vessels, which target certain species. This causes observations to mostly include target species or frequently associated species. Furthermore, fishing boats tend to target some areas based on outputs of fishing guidance models so it creates an artificial correlation between the parameters used in these models and the presence of animals.

The strength of deep learning in this context is that it makes no assumption when there is no data: it replicates the results from similar well-known areas. This partly compensates for sampling effort heterogeneity. But this only works when there is a homogeneous population, unlike *Thunnus thynnus* which has two separate stocks (West and East Atlantic). This explains why the model failed

to extrapolate from West Atlantic data and to predict high probabilities in the Mediterranean Sea and the Bay of Biscay.

Finally, some data come from scientific tracking of individual animals, so these individuals may be over-represented in our data and reflect their preferences rather than the general tendency of their species. The large amount of occurrences that we used help tackle this bias.

These biases would be better tackled with more available data, which is a serious issue in the open oceans. Little data is produced relative to the size of the oceans, and a large part of this data is not shared publicly. More data is key to better models and more trustworthy distribution maps.

4.2.3 Undetected patterns

Detection of seasonal migrations is incomplete. For instance, we should see the *Megaptera novaeangliae* distribution spreading north during the southern winter. The model also did not catch the *Thunnus thynnus* seasonal spawning in summer in the Mediterranean Sea.

4.3 Suggestions to further improve the modelling methods

4.3.1 Occurrence data

As a first experiment, occurrence data were selected randomly for this study. Even though the aim should not be to have a perfect fit between observation data and model predictions, sourcing training data from many different providers rather than randomly would help reduce observer bias.

4.3.2 Environmental data

It has been suggested to include 3D environmental data, as most variables vary with depth and occurrence data are not limited to the surface [53]. Such data could easily be included in the input data with no change to our method. Additional data may be beneficial, in particular the distance to the nearest coastline or level of anthropisation.

4.3.3 Other use cases

The present method could be used at different scales, in particular in coastal areas. This would require a significant change in the input variables, as the resolution of globally available environmental data is a limiting factor. They could be replaced by satellite or drone images, as well as locally available (more precise) environmental data.

5. Conclusion

5.1 Main findings and their significance

The present method provides a way to estimate species distribution at all dates and all areas, provided environmental data is available. While this makes it difficult to judge accuracy (there is often no reference data), it provides a baseline that can be calculated for any species (that have enough existing observations). Researchers working on terrestrial plants have also shown that such models may be used to infer species distribution for rare species, by extrapolating results from co-occurring species [17].

5.2 Implications for management and conservation of offshore species

We hope this method will be developed further and used on other endangered species, together with existing methods and field observation. The technique that we presented would be especially useful in the hands of scientists who are experts in the life cycle of specific species. It would help them

increase scientific knowledge of their distributions, which is essential for decision-makers to target areas of interest for conservation.

5.3 Recommendations for future research and potential applications

While the accuracy of our distribution maps is difficult to assess, there is exceptional room for improvement and further research. All the blocks in Figure 1 can be modified, either to adapt the process to a different use case or to try to improve the quality of the results. Here are some examples of potential changes:

- To study other species, the initial choice of species can be changed, for example, to focus on sedentary species or a specific area.
- To improve accuracy, the occurrence data may be selected in other ways that are not random.
- To investigate the influence of other variables, they may be added to the variable set.
- To study the long-term effect of environmental conditions, some variables may be included with a longer time lag such as months or years.

The results we presented in this article are a small part of what can be achieved with this model. Many other scientific questions can be investigated both with the model we provide (already trained) or with other models trained with the same method.

Acknowledgement

This study would not have been possible without Sylvain Poulain and his support in setting up docker images and GPU calculation environments.

We thank Hervé Demarcq for sharing his expertise on environmental data products, and Emmanuel Chassot for his invaluable insights into pelagic fishes and species distribution models.

Funding statement

This study was conducted as part of the G2OI project, co-financed by the European Union, the Reunion region, and the French Republic.

Open data statement

Code

The code that was used to prepare the data, train the model, and export the outputs is available on GitHub in the IRDG2OI/deep-sdm-oceans repository and on Zenodo [54].

Input data

The input data include the CSV file describing the geographical points, the standardized numpy arrays of corresponding environmental data, and the standardization factors. They are available on Zenodo [55] for each use case:

- Training data (includes train+validation+test)
- Prediction data for the world at 4 dates
- Prediction data for the Western Indian Ocean at 53 dates

Modelling

We provide the model checkpoint and configuration file [56], so researchers can make predictions with the presently described model.

We also provide the code that was used for training so researchers can adapt it to their needs and retrain a new model [54]. It consists of Python files based on a custom version of Malpolon.

Results

The distribution maps were uploaded to Zenodo for easy visualisation, in three repositories:

- Global predictions as PNGs and GeoTIFFs [42]
- Western Indian Ocean predictions as GIFs and GeoTIFFs [43]
- +2°C hypothesis as PNGs and GeoTIFFs [50]

References

[1] Thomas J. Webb, Edward Vanden Berghe and Ron O'Dor. 'Biodiversity's Big Wet Secret: The Global Distribution of Marine Biological Records Reveals Chronic Under-Exploration of the Deep Pelagic Ocean'. In: *PLOS ONE* 5.8 (2nd Aug. 2010), e10223. DOI: 10.1371/journal.pone.0010223.

[2] Andrew R. Thurber, Andrew K. Sweetman, Bhavani E. Narayanaswamy, Daniel O. B. Jones, Jeroen Ingels and R. L. Hansman. 'Ecosystem Function and Services Provided by the Deep Sea'. In: *Biogeosciences (Online)* 11.14 (2014), pp. 3941–3963.

[3] Jeremy B. C. Jackson et al. 'Historical Overfishing and the Recent Collapse of Coastal Ecosystems'. In: *Science (New York, N.Y.)* 293.5530 (27th July 2001), pp. 629–637. DOI: 10.1126/science.1059199.

[4] José Vinicio Macías-Zamora. 'Chapter 19 - Ocean Pollution'. In: *Waste*. Ed. by Trevor M. Letcher and Daniel A. Vallero. Boston: Academic Press, 1st Jan. 2011, pp. 265–279. DOI: 10.1016/B978-0-12-381475-3.10019-1.

[5] Alex Sen Gupta et al. 'Drivers and Impacts of the Most Extreme Marine Heatwave Events'. In: *Scientific Reports* 10 (1, 1 9th Nov. 2020), p. 19359. DOI: 10.1038/s41598-020-75445-3.

[6] Elizabeth R. Selig et al. 'Mapping Global Human Dependence on Marine Ecosystems'. In: *Conservation Letters* 12.2 (2019), e12617. DOI: 10.1111/conl.12617.

[7] Jennifer Miller. 'Species Distribution Modeling'. In: *Geography Compass* 4.6 (2010), pp. 490–509. DOI: 10.1111/j.1749-8198.2010.00351.x.

[8] IPCC. 'Summary for Policymakers'. In: *Special Report on the Ocean and Cryosphere in a Changing Climate*. In press, 2019.

[9] Antoine Guisan and Wilfried Thuiller. 'Predicting Species Distribution: Offering More than Simple Habitat Models'. In: *Ecology Letters* 8.9 (Sept. 2005), pp. 993–1009. DOI: 10.1111/j.1461-0248.2005.00792.x.

[10] Antoine Guisan and Niklaus E. Zimmermann. 'Predictive Habitat Distribution Models in Ecology'. In: *Ecological Modelling* 135.2-3 (Dec. 2000), pp. 147–186. DOI: 10.1016/S0304-3800(00)00354-9.

[11] Antonio G. Ramos, J. Santiago, Pablo Sangra and M. Canton. 'An Application of Satellite-Derived Sea Surface Temperature Data to the Skipjack (*Katsuwonus Pelamis* Linnaeus, 1758) and Albacore Tuna (*Thunnus Alalunga* Bonaterre, 1788) Fisheries in the North-East Atlantic'. In: *International Journal of Remote Sensing* 17.4 (Mar. 1996), pp. 749–759. DOI: 10.1080/01431169608949042.

[12] Jean-Noël Druon, Simone Panigada, Léa David, Alexandre Gannier, Pascal Mayol, Antonella Arcangeli, Ana Cañadas, Sophie Laran, Nathalie Di Méglio and Pauline Gauffier. 'Potential Feeding Habitat of Fin Whales in the Western Mediterranean Sea: An Environmental Niche Model'. In: *Marine Ecology Progress Series* 464 (19th Sept. 2012), pp. 289–306. DOI: 10.3354/meps09810.

[13] Stephanie Brodie, Alistair J. Hobday, James A. Smith, Jason D. Everett, Matt D. Taylor, Charles A. Gray and Iain M. Suthers. 'Modelling the Oceanic Habitats of Two Pelagic Species Using Recreational Fisheries Data'. In: *Fisheries Oceanography* 24.5 (2015), pp. 463–477. DOI: 10.1111/fog.12122.

[14] Sara M. Melo-Merino, Héctor Reyes-Bonilla and Andrés Lira-Noriega. 'Ecological Niche Models and Species Distribution Models in Marine Environments: A Literature Review and Spatial Analysis of Evidence'. In: *Ecological Modelling* 415 (1st Jan. 2020), p. 108837. DOI: 10.1016/j.ecolmodel.2019.108837.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. 'Deep Residual Learning for Image Recognition'. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016, pp. 770–778.

[16] Christophe Botella, Alexis Joly, Pierre Bonnet, Pascal Monestiez and François Munoz. 'A Deep Learning Approach to Species Distribution Modelling'. In: *Multimedia Tools and Applications for Environmental & Biodiversity Informatics*. Ed. by Alexis Joly, Stefanos Vrochidis, Kostas Karatzas, Ari Karppinen and Pierre Bonnet. Cham: Springer International Publishing, 2018, pp. 169–199. DOI: 10.1007/978-3-319-76445-0_10.

[17] Benjamin Deneu, Maximilien Servajean, Pierre Bonnet, Christophe Botella, François Munoz and Alexis Joly. 'Convolutional Neural Networks Improve Species Distribution Modelling by Capturing the Spatial Structure of the Environment'. In: *PLOS Computational Biology* 17.4 (19th Apr. 2021), e1008856. DOI: 10.1371/journal.pcbi.1008856.

[18] *GBIF*. URL: https://www.gbif.org/ (visited on 17/07/2023).

[19] Rui Chen, Meiling Wang and Yi Lai. 'Analysis of the Role and Robustness of Artificial Intelligence in Commodity Image Recognition under Deep Learning Neural Network'. In: *PLOS ONE* 15.7 (7th July 2020), e0235783. DOI: 10.1371/journal.pone.0235783.

[20] Leonardo E. Moraes, Eduardo Paes, Alexandre Garcia, Osmar Möller Jr and João Vieira. 'Delayed Response of Fish Abundance to Environmental Changes: A Novel Multivariate Time-Lag Approach'. In: *Marine Ecology Progress Series* 456 (7th June 2012), pp. 159–168. DOI: 10.3354/meps09731.

[21] Gaétan Morand and Sylvain Poulain. *GeoEnrich v0.5.8: A New Tool for Scientists to Painlessly Enrich Species Occurrence Data with Environmental Variables*. Version v0.5.8. 30th May 2023. DOI: 10.5281/ZENODO.6458090.

[22] Ko Fujioka et al. 'Spatial and Temporal Variability in the Trans-Pacific Migration of Pacific Bluefin Tuna (Thunnus Orientalis) Revealed by Archival Tags'. In: *Progress in Oceanography* 162 (Mar. 2018), pp. 52–65. DOI: 10.1016/j.pocean.2018.02.010.

[23] Jean-Marc Fromentin and Daniel Lopuszanski. 'Migration, Residency, and Homing of Bluefin Tuna in the Western Mediterranean Sea'. In: *ICES Journal of Marine Science* 71.3 (1st Apr. 2014), pp. 510–518. DOI: 10.1093/icesjms/fst157.

[24] GEBCO Bathymetric Compilation Group 2022. *The GEBCO_2022 Grid - a Continuous Terrain Model of the Global Oceans and Land*. Documents,Network common data form. Version 1. NERC EDS British Oceanographic Data Centre NOC, 2022. DOI: 10.5285/E0F0BB80-AB44-2739-E053-6C86ABC0289C.

[25] European Union-Copernicus Marine Service. *Multi Observation Global Ocean 3D Temperature Salinity Height Geostrophic Current and MLD*. Mercator Ocean International, 2020. DOI: 10.48670/MOI-00052.

[26] European Union-Copernicus Marine Service. *GLOBAL OCEAN L4 SIGNIFICANT WAVE HEIGHT FROM REPROCESSED SATELLITE MEASUREMENTS*. Mercator Ocean International, 2021. DOI: 10.48670/MOI-00177.

[27] Carl Mears, Tong Lee, Lucrezia Ricciardulli, Xiaochun Wang and Frank Wentz. *RSS Cross-Calibrated Multi-Platform (CCMP) 6-Hourly Ocean Vector Wind Analysis on 0.25 Deg Grid, Version 3.0*. Remote Sensing Systems, 10th Aug. 2022. DOI: 10.56236/RSS-uv6h30.

[28] European Union-Copernicus Marine Service. *Global Ocean Biogeochemistry Hindcast*. Mercator Ocean International, 2018. DOI: 10.48670/MOI-00019.

[29] European Union-Copernicus Marine Service. *Global Ocean Biogeochemistry Analysis and Forecast*. Mercator Ocean International, 2019. DOI: 10.48670/MOI-00015.

[30] LOCEAN/CLS/CTOH/CNES. *FSLE - Finite-Size Lyapunov Exponents and Orientations of the Associated Eigenvectors: Version 2021*. Ed. by Guillaume Taburet. NetCDF 4. Version DT2021. CNES, 2021. DOI: 10.24400/527896/A01-2022.002.

[31] European Union-Copernicus Marine Service. *GLOBAL OCEAN GRIDDED L4 SEA SURFACE HEIGHTS AND DERIVED VARIABLES REPROCESSED (1993-ONGOING)*. Mercator Ocean International, 2021. DOI: 10.48670/MOI-00148.

[32] European Union-Copernicus Marine Service. *GLOBAL OCEAN GRIDDED L4 SEA SURFACE HEIGHTS AND DERIVED VARIABLES NRT*. Mercator Ocean International, 2017. DOI: 10.48670/MOI-00149.

[33] Shubha Sathyendranath et al. *ESA Ocean Colour Climate Change Initiative (Ocean_Colour_cci): Version 5.0 Data*. NERC EDS Centre for Environmental Data Analysis, 2021. DOI: 10.5285/1DBE7A109C0244AAAD713E078FD3059A.

[34] NASA/JPL. *GHRSST Level 4 MUR 0.25deg Global Foundation Sea Surface Temperature Analysis (v4.2)*. NASA Physical Oceanography DAAC, 2019. DOI: 10.5067/GHM25-4FJ42.

[35] European Union-Copernicus Marine Service. *Global Ocean Colour (Copernicus-GlobColour), Bio-Geo-Chemical, L4 (Monthly and Interpolated) from Satellite Observations (1997-Ongoing)*. Mercator Ocean International, 2022. DOI: 10.48670/MOI-00281.

[36] Titouan Lorieul, Théo Larcher and Alexis Joly. *Plantnet/Malpolon: Deep-SDM Framework*. URL: https://github.com/plantnet/malpolon (visited on 15/07/2023).

[37] Adam Paszke et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. 3rd Dec. 2019. URL: http://arxiv.org/abs/1912.01703 (visited on 15/06/2023). preprint.

[38] William Falcon et al. *PyTorchLightning/Pytorch-Lightning: 0.7.6 Release*. Version 0.7.6. Zenodo, 15th May 2020. DOI: 10.5281/ZENODO.3828935.

[39] Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris and Nebojsa Jojic. 'Multi-Label Learning From Single Positive Labels'. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021, pp. 933–942.

[40] Mukund Sundararajan, Ankur Taly and Qiqi Yan. *Axiomatic Attribution for Deep Networks*. 12th June 2017. URL: http://arxiv.org/abs/1703.01365 (visited on 26/07/2023). preprint.

[41] Narine Kokhlikyan et al. *Captum: A Unified and Generic Model Interpretability Library for PyTorch*. 2020.

[42] Gaétan Morand. *Deep-SDMs in the Open Oceans - OUTPUTS - World*. 1st Aug. 2023. DOI: 10.5281/zenodo.8202261.

[43] Gaétan Morand. *Deep-SDMs in the Open Oceans - OUTPUTS - Western Indian Ocean*. 1st Aug. 2023. DOI: 10.5281/zenodo.8202056.

[44] G. Causey Whittow. 'Wedge-Tailed Shearwater (Ardenna Pacifica)'. In: *Birds of the World*. Ed. by Shawn M. Billerman, Brooke K. Keeney, Paul G. Rodewald and Thomas S. Schulenberg. Cornell Lab of Ornithology, 4th Mar. 2020. DOI: 10.2173/bow.wetshe.01.

[45] William F. Perrin, Bernd Würsig and J. G. M. Thewissen. 'Right Whales'. In: *Encyclopedia of Marine Mammals*. Academic Press, 26th Feb. 2009.

[46] Stanford University and TAG-A-Giant Foundation. *Atlantic Bluefin Tuna (Thunnus Thynnus) | Smithsonian Ocean*. URL: https://ocean.si.edu/ocean-life/fish/atlantic-bluefin-tuna-thunnus-thynnus (visited on 17/07/2023).

[47] Jean-Marc Fromentin, Gabriel Reygondeau, Sylvain Bonhommeau and Gregory Beaugrand. 'Oceanographic Changes and Exploitation Drive the Spatio-Temporal Dynamics of Atlantic Bluefin Tuna *(Thunnus Thynnus)*'. In: *Fisheries Oceanography* 23.2 (Mar. 2014), pp. 147–156. DOI: 10.1111/fog.12050.

[48] Sylvain Lenoir, Gregory Beaugrand and Éric Lecuyer. 'Modelled Spatial Distribution of Marine Fish and Projected Modifications in the North Atlantic Ocean: NICHE MODELLING AND FISH BIOGEOGRAPHY'. In: *Global Change Biology* 17.1 (Jan. 2011), pp. 115–129. DOI: 10.1111/j.1365-2486.2010.02229.x.

[49] Richard W. Brill. 'A Review of Temperature and Oxygen Tolerance Studies of Tunas Pertinent to Fisheries Oceanography, Movement Models and Stock Assessments'. In: *Fisheries Oceanography* 3.3 (1994), pp. 204–216. DOI: 10.1111/j.1365-2419.1994.tb00098.x.

[50] Gaétan Morand. *Deep-SDMs in the Open Oceans - OUTPUTS - World +2°C*. 1st Aug. 2023. DOI: 10.5281/zenodo.8202709.

[51] Emilie Tew Kai, Vincent Rossi, Joel Sudre, Henri Weimerskirch, Cristobal Lopez, Emilio Hernandez-Garcia, Francis Marsac and Veronique Garçon. 'Top Marine Predators Track Lagrangian Coherent Structures'. In: *Proceedings of the National Academy of Sciences* 106.20 (19th May 2009), pp. 8245–8250. DOI: 10.1073/pnas.0811034106.

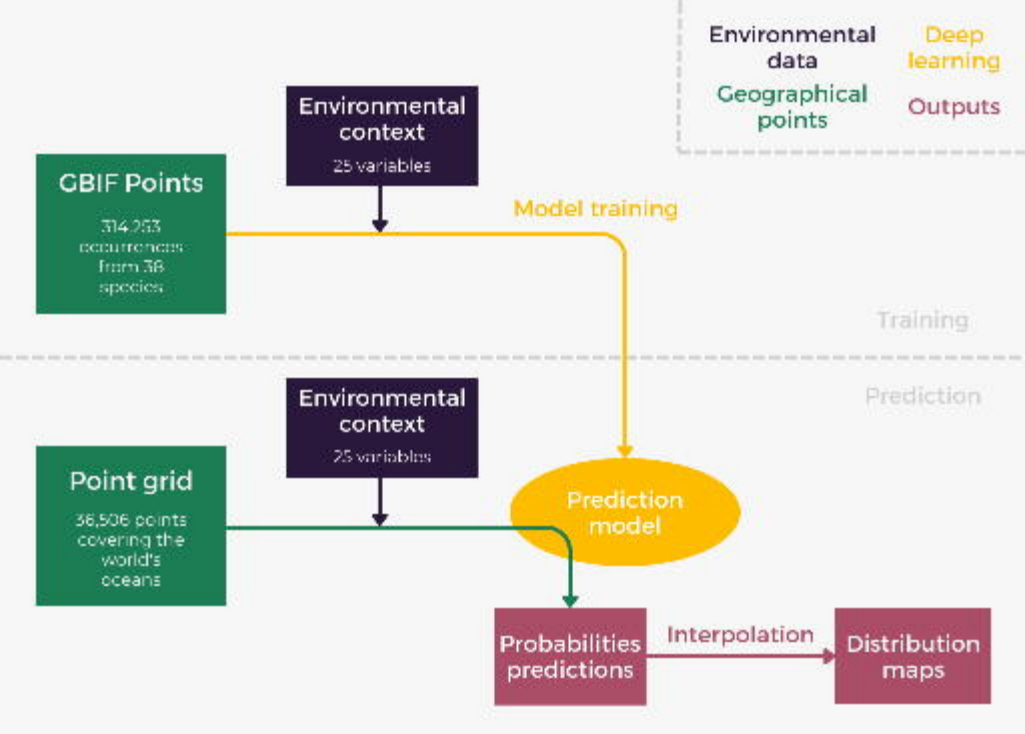
[52] Samuel Bosch, Lennert Tyberghein, Klaas Deneudt, Francisco Hernandez and Olivier De Clerck. 'In Search of Relevant Predictors for Marine Species Distribution Modelling Using the MarineSPEED Benchmark Dataset'. In: *Diversity and Distributions* 24.2 (Feb. 2018). Ed. by Alexandra Syphard, pp. 144–157. DOI: 10.1111/ddi.12668.

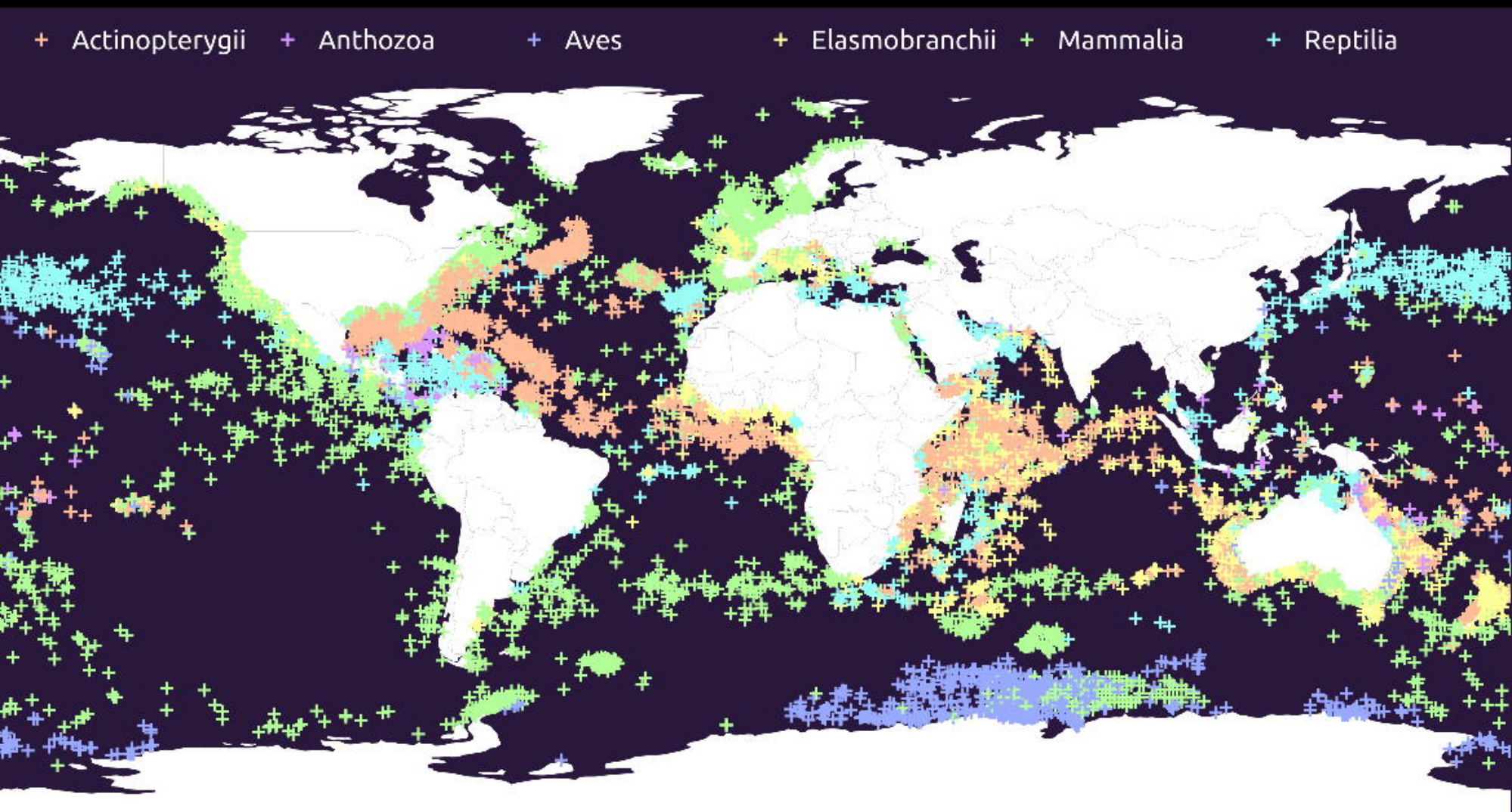
[53] Grant A. Duffy and Steven L. Chown. 'Explicitly Integrating a Third Dimension in Marine Species Distribution Modelling'. In: *Marine Ecology Progress Series* 564 (3rd Feb. 2017), pp. 1–8. DOI: 10.3354/meps12011.

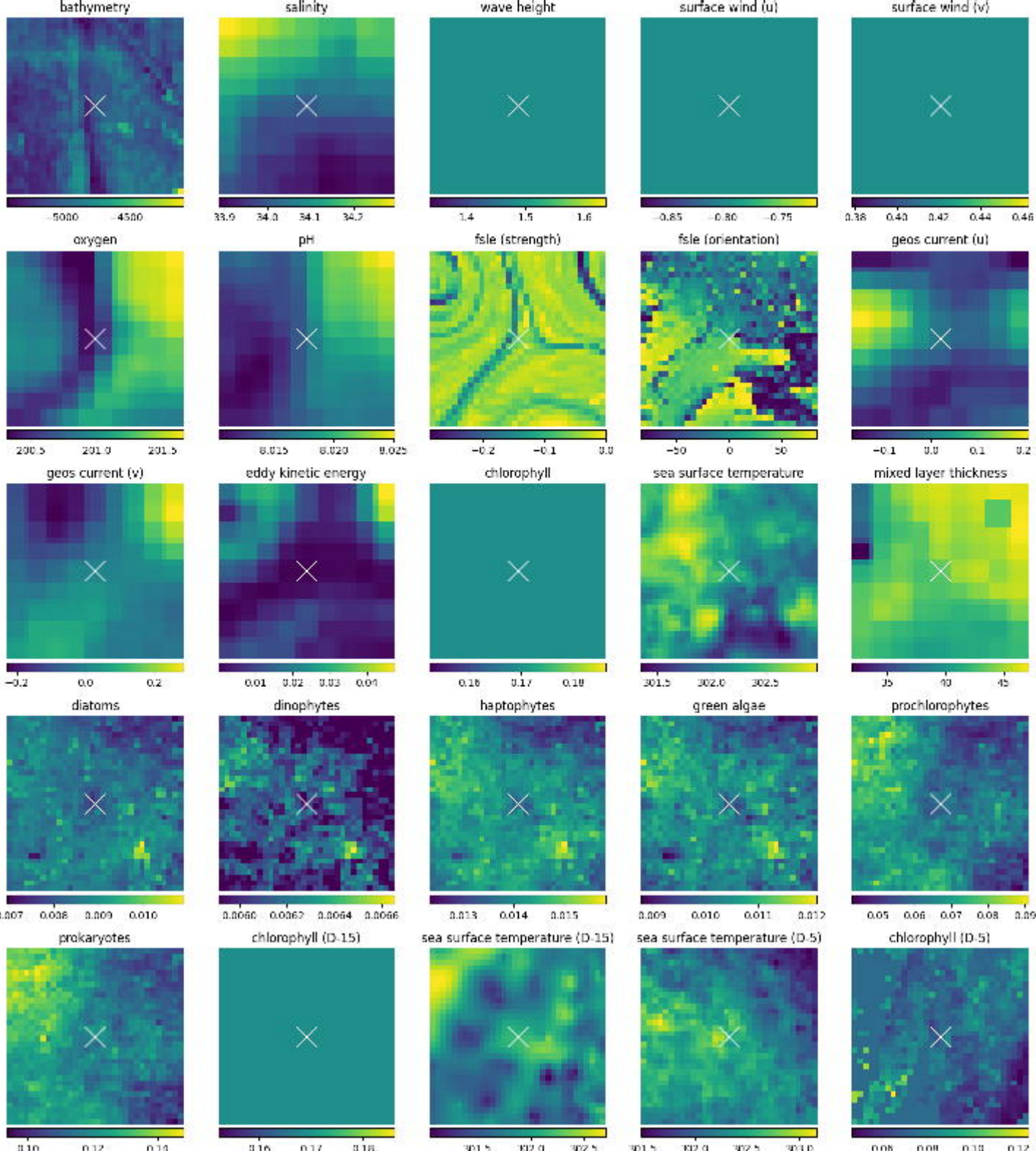
[54] Gaétan Morand. *Deep-SDMs in the Open Oceans - CODE*. Zenodo, 8th Aug. 2023. DOI: 10.5281/zenodo.8222155.

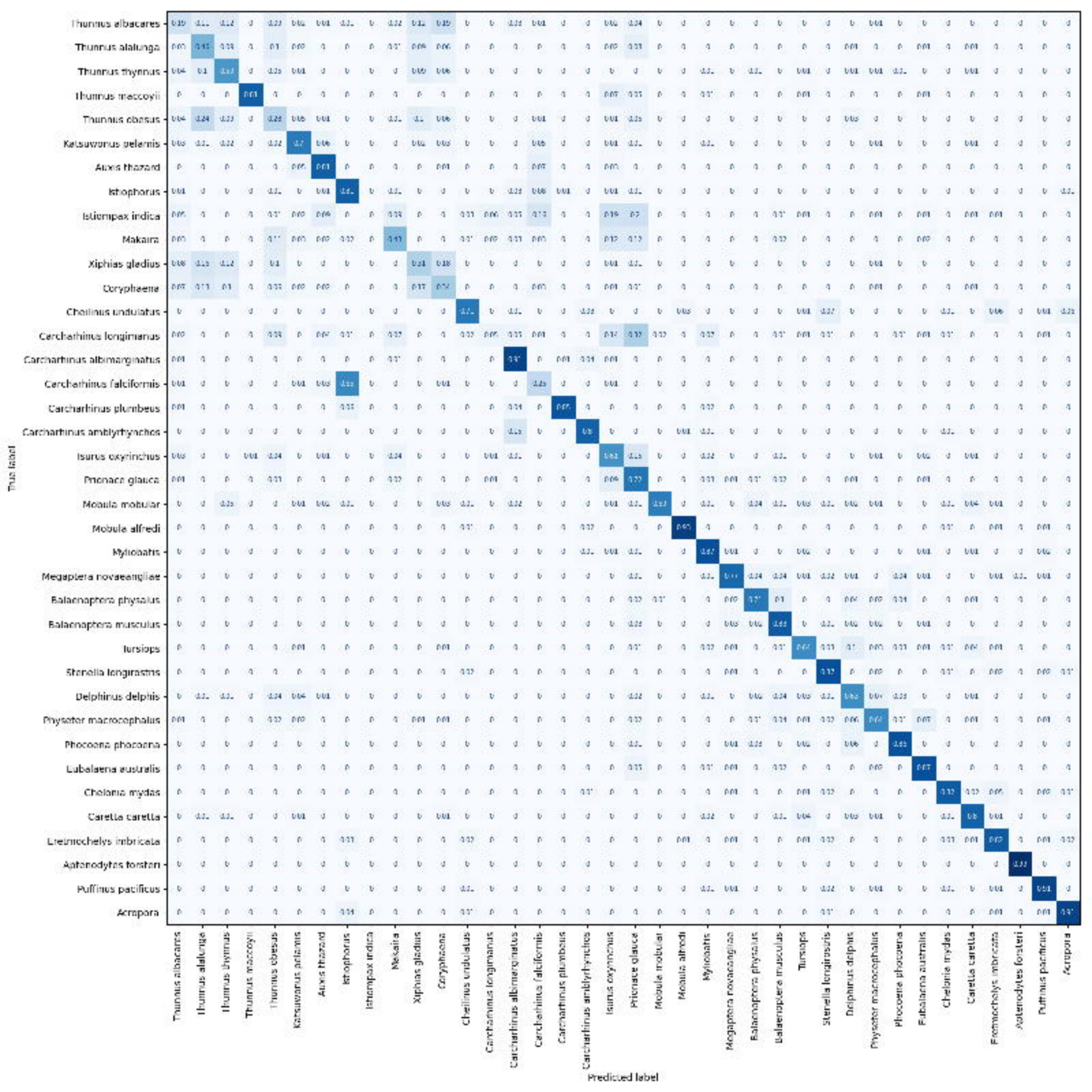
[55] Gaétan Morand. *Deep-SDMs in the Open Oceans - INPUT DATA*. Zenodo, 27th July 2023. DOI: 10.5281/zenodo.8188512.

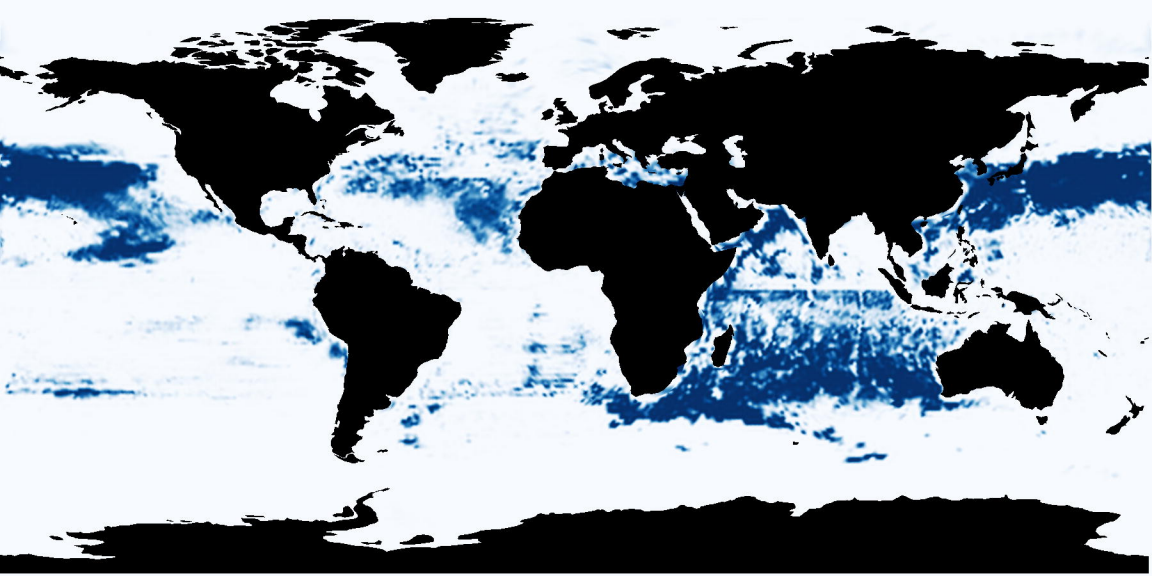
[56] Gaétan Morand. *Deep-SDMs in the Open Oceans - MODEL CHECKPOINT*. 1st Aug. 2023. DOI: 10.5281/zenodo.8202914.

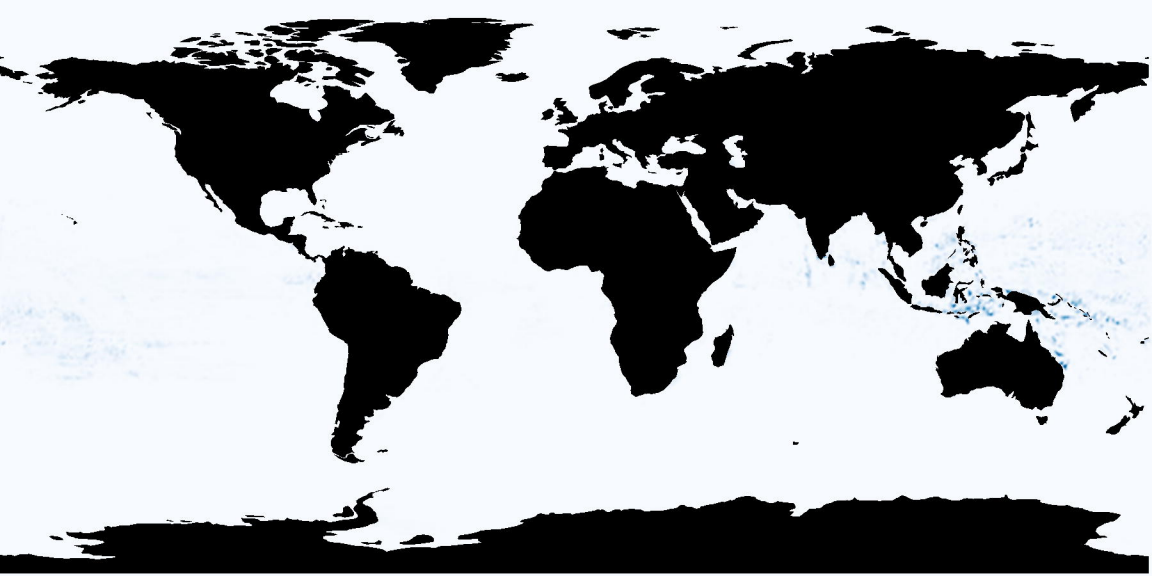

Environmental data
Deep learning
Geographical points
Outputs
Environmental context
25 variables
GBIF Points
314 263 occurrences from 38 species
Model training
Training
Prediction
Environmental context
25 variables
Point grid
36,506 points covering the world's oceans
Prediction model
Probabilities predictions
Interpolation
Distribution maps

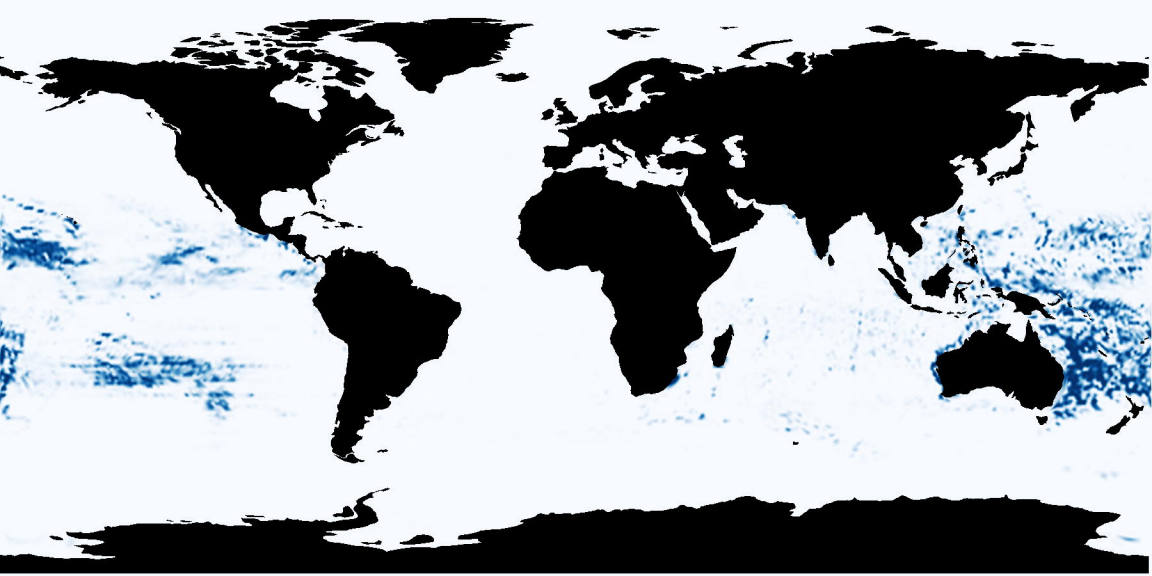

Actinopterygii
Anthozoa
Aves
Elasmobranchii
Mammalia
Reptilia

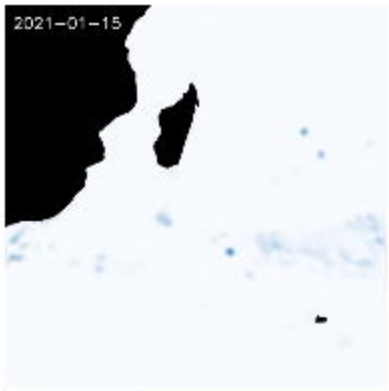

bathymetry
−5000
−4500
salinity
33.9
34.0
34.1
34.2
wave height
1.4
1.5
1.6
surface wind (u)
−0.85
−0.80
−0.75
surface wind (v)
0.38
0.40
0.42
0.44
0.46
oxygen
200.5
201.0
201.5
pH
8.015
8.020
8.025
fsle (strength)
−0.2
−0.1
0.0
fsle (orientation)
−50
0
50
geos current (u)
−0.1
0.0
0.1
0.2
geos current (v)
−0.2
0.0
0.2
eddy kinetic energy
0.01
0.02
0.03
0.04
chlorophyll
0.16
0.17
0.18
sea surface temperature
301.5
302.0
302.5
mixed layer thickness
35
40
45
diatoms
0.007
0.008
0.009
0.010
dinophytes
0.0060
0.0062
0.0064
0.0066
haptophytes
0.013
0.014
0.015
green algae
0.009
0.010
0.011
0.012
prochlorophytes
0.05
0.06
0.07
0.08
0.09
prokaryotes
0.10
0.12
0.14
chlorophyll (D-15)
0.16
0.17
0.18
sea surface temperature (D-15)
301.5
302.0
302.5
sea surface temperature (D-5)
301.5
302.0
302.5
303.0
chlorophyll (D-5)
0.08
0.09
0.10
0.12

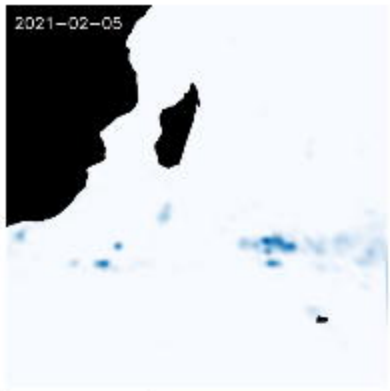

True label
Predicted label
Thunnus albacares
Thunnus alalunga
Thunnus thynnus
Thunnus maccoyii
Thunnus obesus
Katsuwonus pelamis
Auxis thazard
Istiophorus
Istiompax indica
Makaira
Xiphias gladius
Coryphaena
Cheilinus undulatus
Carcharhinus longimanus
Carcharhinus albimarginatus
Carcharhinus falciformis
Carcharhinus plumbeus
Carcharhinus amblyrhynchos
Isurus oxyrinchus
Prionace glauca
Mobula mobular
Mobula alfredi
Mylioatis
Megaptera novaeangliae
Balaenoptera physalus
Balaenoptera musculus
Tursiops
Stenella longirostris
Delphinus delphis
Physeter macrocephalus
Phocoena phocoena
Eubalaena australis
Chelonia mydas
Caretta caretta
Eretmochelys imbricata
Aptenodytes forsteri
Puffinus pacificus
Acropora

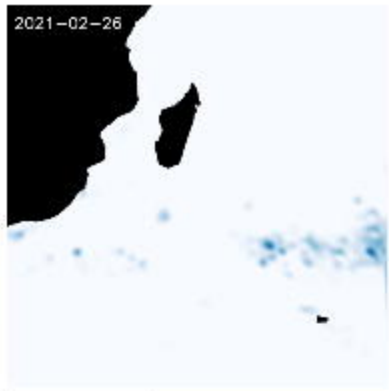


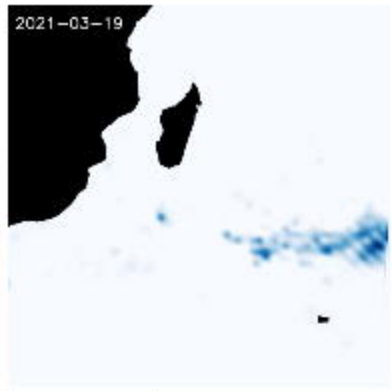


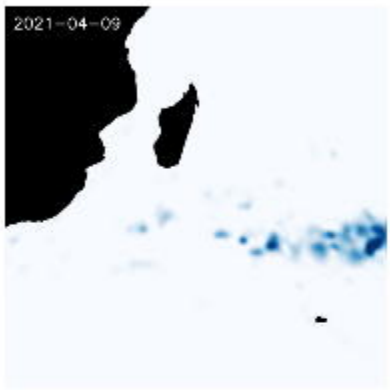


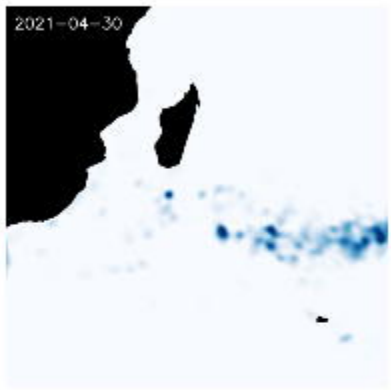

2021-01-15

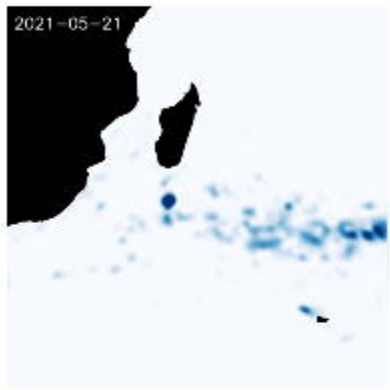

2021-02-05

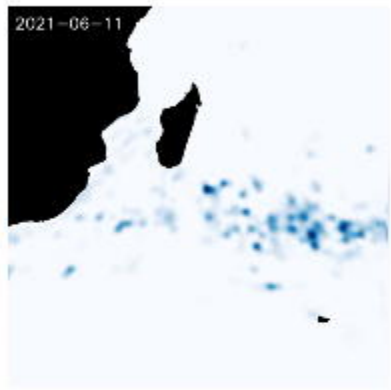

2021-02-26

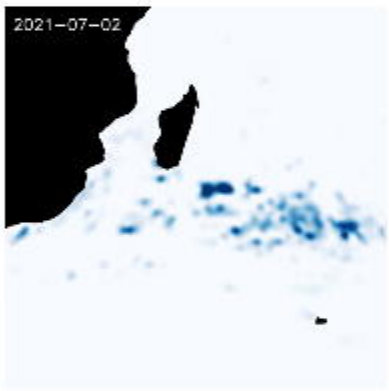

2021-03-19

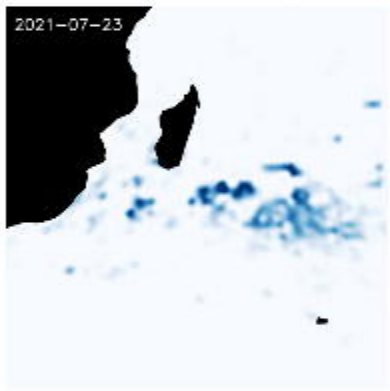

2021-04-09

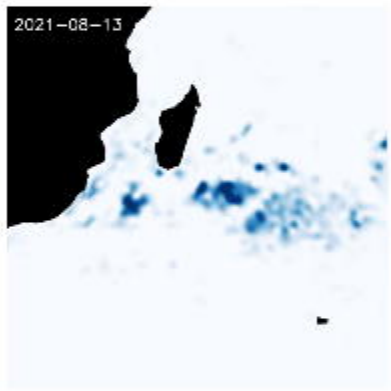

2021-04-30

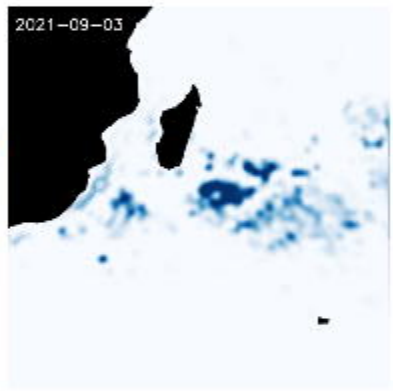

2021-05-21

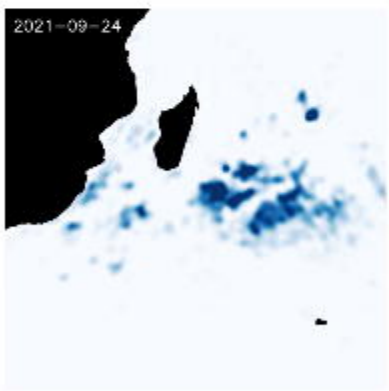

2021-06-11

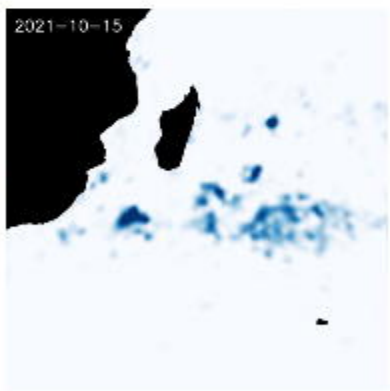

2021-07-02

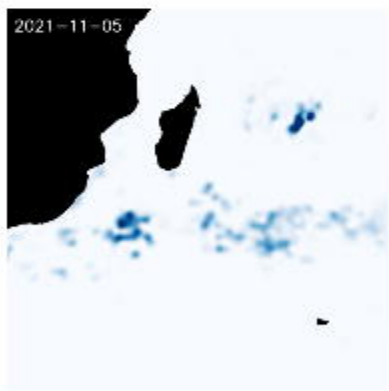

2021-07-23

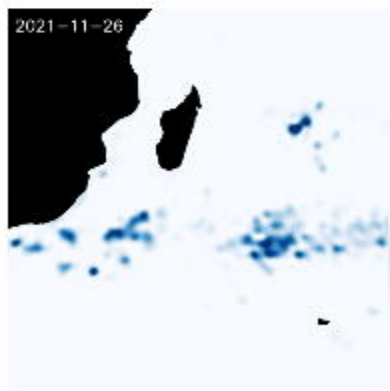

2021-08-13

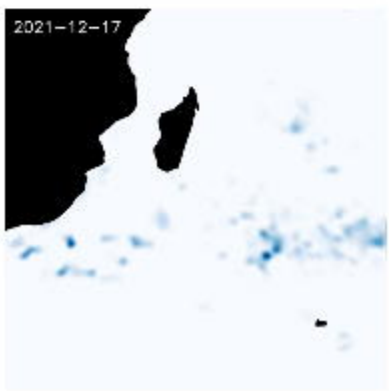

2021-09-03

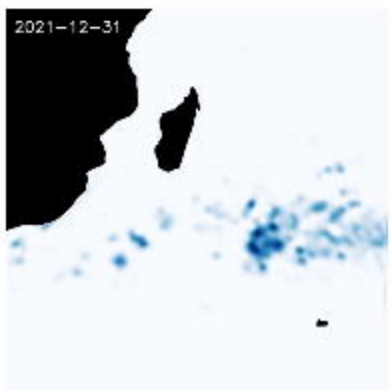

2021-09-24

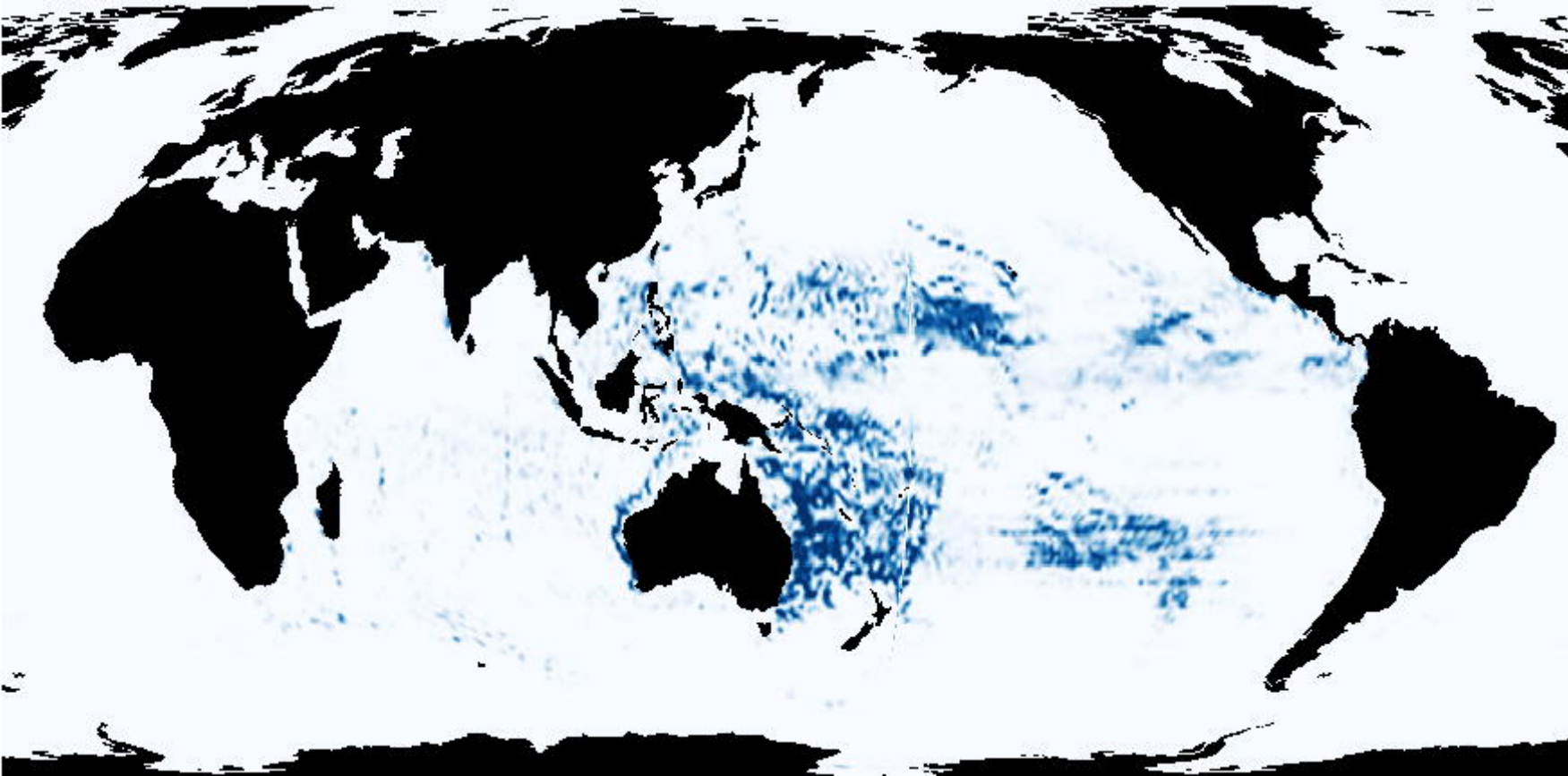

2021-10-15


2021-11-05

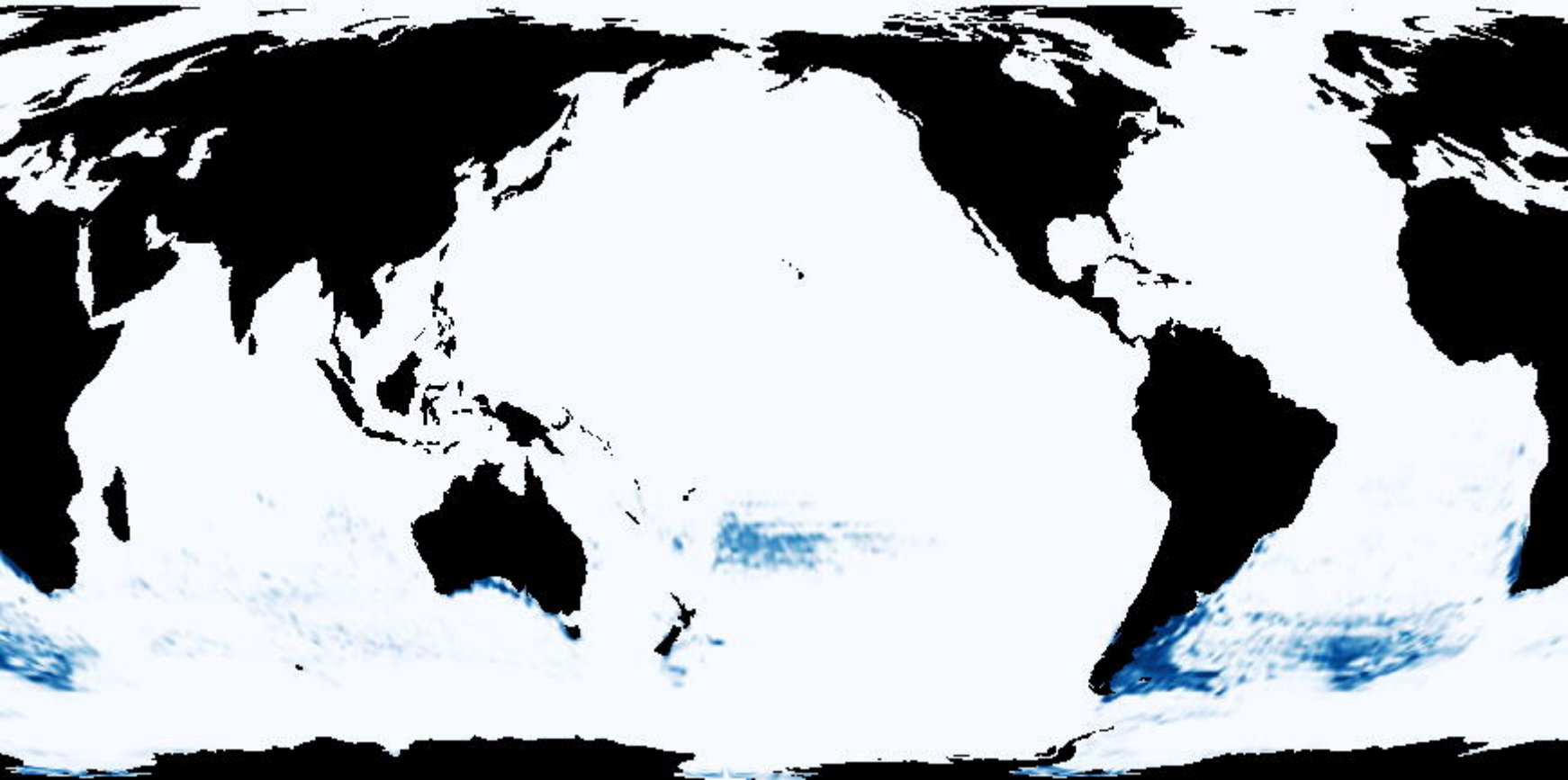

2021-11-26

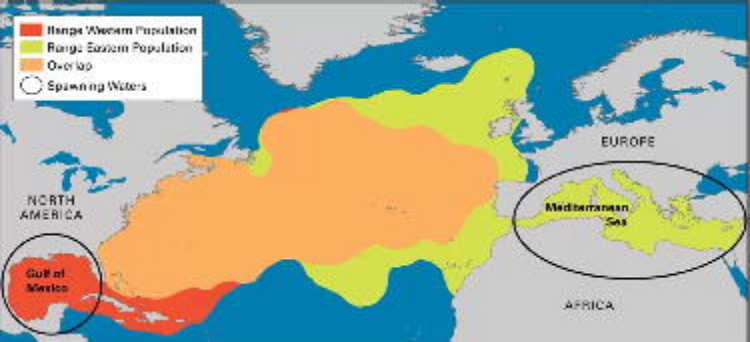

2021-12-17


2021-12-31




Arctic Ocean
E. japonica
E. glacialis
Atlantic Ocean
Pacific Ocean
Indian Ocean
E. australis
2,500 NM
N




Range Western Population
Range Eastern Population
Overlap
Spawning Waters
EUROPE
NORTH AMERICA
Mediterranean Sea
Gulf of Mexico
AFRICA

Thunnus albacares
Thunnus alalunga
Thunnus thynnus
Thunnus maccoyii
Thunnus obesus
Katsuwonus pelamis
Auxis thazard
Istiophorus
Makaira
Xiphias gladius
Coryphaena
Cheilinus undulatus
Carcharhinus longimanus
Carcharhinus albimarginatus
Carcharhinus falciformis
Carcharhinus plumbeus
Carcharhinus amblyrhynchos
Isurus oxyrinchus
Prionace glauca
Mobula mobular
Mobula alfredi
Myliobatis
Megaptera novaeangliae
Balaenoptera physalus
Balaenoptera musculus
Tursiops
Stenella longirostris
Delphinus delphis
Physeter macrocephalus
Phocoena phocoena
Lubalaena australis
Chelonia mydas
Caretta caretta
Eretmochelys imbricata
Aptenodytes forsteri
Puffinus pacificus
Acropora
bathymetry
salinity
wave height
surface wind (u)
surface wind (v)
oxygen
pH
fsle (strength)
fsle (orientation)
geos current (u)
geos current (v)
chlorophyll
sea surface temperature
mixed layer thickness
diatoms
dinophytes
haptophytes
green algae
prochlorophytes
prokaryotes
Atlantic Ocean
Indian Ocean
Pacific Ocean
North hemisphere