Peer Community Journal

Section: Ecology

Research article

Published
2024-09-26

Cite as
Gaétan Morand, Alexis Joly, Tristan Rouyer, Titouan Lorieul and Julien Barde (2024) *Predicting species distributions in the open ocean with convolutional neural networks*, Peer Community Journal, 4: e93.

Correspondence
gaetan.morand@umontpellier.fr

Peer-review
Peer reviewed and recommended by PCI Ecology,
https://doi.org/10.24072/pci.ecology.100584

Predicting species distributions in the open ocean with convolutional neural networks

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

Volume **4** (2024), article e93

https://doi.org/10.24072/pcjournal.471

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 the scarcity of 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 ocean, discuss limitations and provide feedback for future improvements or adjustments. In this case study, we considered 38 taxa comprising pelagic fishes, elasmobranchs, marine mammals, marine turtles and birds. We trained a model to predict probabilities 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 classifier 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. Additionally, this purely correlative 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 and salinity as the most influential variables, in that order. These insights can prove valuable for future species-specific ecology studies.

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

Peer Community Journal is a member of the Centre Mersenne for Open Scientific Publishing
http://www.centre-mersenne.org/

e-ISSN 2804-3871

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 (Raffaelli et al., 2005; Robinson et al., 2011). It plays a critical role in regulating the Earth's climate and biogeochemical cycles (including nutrient cycles and carbon sequestration), making it a vital component of all life on Earth (Barrón and Duarte, 2015; Ganzeveld et al., 2009).

However, the ocean is facing a range of human-induced threats, including over-fishing, pollution and climate change (Jackson et al., 2001; Macías-Zamora, 2011; Sen Gupta et al., 2020). 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 (Selig et al., 2019).

To solve these most pressing challenges, a necessary first step is to understand how marine life is distributed within the open ocean. Species distribution models can provide valuable insights into where different species are likely to be found and how environmental factors drive their distribution (Miller, 2010). By developing accurate and reliable models, we can identify areas that are most threatened by foreseen local disturbances 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 (IPCC, 2019). 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 the open ocean 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

A wide variety of Species Distribution Models (SDMs) have been discussed in literature (Guisan and Thuiller, 2005). This is generally done through modelling a species-specific *environmental niche* where environmental conditions are favourable to the species in the long term, shaped by natural selection (Guisan and Zimmermann, 2000). Predictors are chosen empirically to try and predict the species' ranges from observed species' occurrences.

Usually, SDMs use climatological summaries of environmental data, at the location where the observation is recorded. These spatially isolated data are unable to convey the full nature of the environmental seascape around animals, as single values cannot represent more complex bathymetry features such as trenches for example. The same applies to other variables, which spatial structure may be more important than isolated values: algal blooms, temperature fronts, eddies, etc. Yet these spatial structures represent processes which are essential to ascertain species distributions (Baudena et al., 2021; Ramos et al., 1996).

A solution to this shortcoming is to include the environmental data in the neighbourhood of species occurrences, but the number of predictors 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 into fewer significant variables. This work may be carried out manually, which enables the model to take advantage of scientists' expert knowledge. This is how some spatial features are added into SDMs (Brodie et al., 2015), but it limits the performance of the model to the scope of existing knowledge and prevents the discovery of previously unknown influential factors.

Furthermore, the use of climatological summaries prevents taking advantage of these spatial features, as they are lost when averaging the values over time. While some climatology products try to mitigate this shortcoming (*e.g.* frequency of presence of fronts (Miller and Christodoulou, 2014)), countless such products would be necessary to capture all types of features. Therefore, the only remaining solution is to use instantaneous values at the time of species occurrence (Mannocci et al., 2017). This slightly changes the objective of the model: it does not try to model the ecological niche anymore, but dynamic distributions of the species and becomes a dynamic SDM (Milanesi et al., 2020). This development is necessary if we want to include the aforementioned spatial structures into predictors. Incidentally, this makes the predictions highly dependent on time, which has two unintended benefits: **1.** making it possible to include the variability of environmental data into SDM predictors, which has been identified as a way to improve their performance (Bateman et al., 2012) and **2.** allowing modelling the variations of distributions over time, which is especially interesting for highly mobile species or those that have rapid population dynamics (Fernandez et al., 2017; Melo-Merino et al., 2020).

This calls for new methods to extend the scope of SDMs to fully take into account the complex spatial structures of environmental seascapes and, as a direct benefit, their variability over time.

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

Convolutional neural networks (CNNs) were designed for image processing, so they have embedded feature extractors that are designed to detect multiple levels of details using convolution layers (He et al., 2016). As the training advances deeper within the layers, small details are increasingly pooled together to be able to detect much more complex shapes. 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 a given point
(2) Homogeneity of the variable in the neighbourhood of occurrence: 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 (Botella et al., 2018). The CNN architecture proved especially useful to capture spatial features, as well as to transfer knowledge from better-known species to lesser-known species (Deneu et al., 2021). CNN-based SDMs, as described here, usually predict species distributions by providing a classification rather than regressions.

While these studies were mostly based on satellite imagery (Sentinel-2), optical data is not enough to represent the state of the oceanic environment, although it yields some interesting products such as chlorophyll concentration. These products and many other significant oceanic variables are available as processed data sets, which should be used for a more comprehensive view of oceanic conditions. Another difference with this previous work is the high temporal variability of the oceanic seascape. Here we present an adaptation of the work of Botella et al. (2018) and Deneu et al. (2021) that includes these adaptations to the specificity of the open ocean.

1.4. Objectives of the study

Through the present study, we explore the possibilities that deep-learning-based SDMs offer in the open ocean. 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 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 relate species presence to environmental data. To achieve this, we used occurrence data from the Global Biodiversity Information Facility (GBIF) (GBIF, 2023) and downloaded environmental data in a buffer around each of their locations, at the date of their occurrence. All the data sets are freely available.

It is important to note that as training data is presence-only, we cannot predict abundance or any absolute measure of presence. That's why we modelled a multivariate output, where predictions are observation probabilities, relative to the 38 taxa that are the subject of this study. 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). The full process is summarized in Figure 1 and is explained in detail in this section.

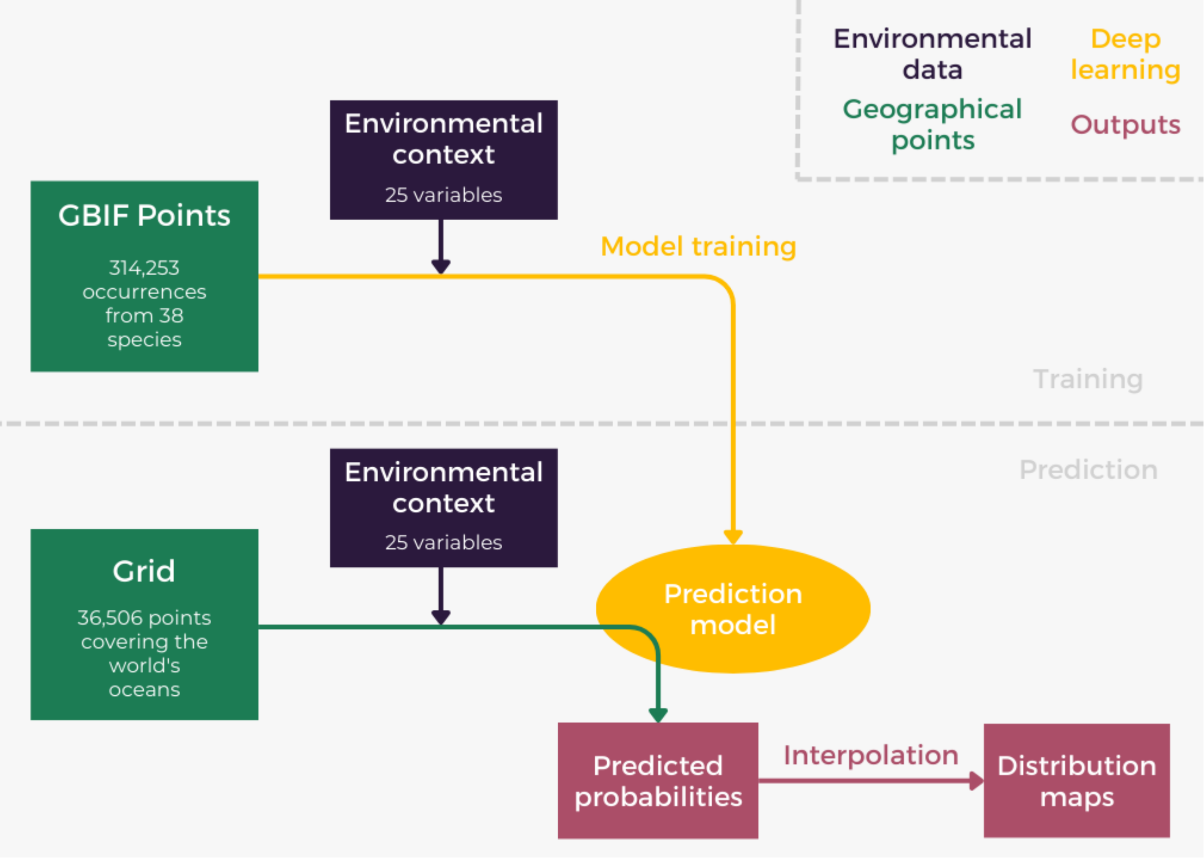


Figure 1 – Summary view of the analysis process: model training in the top half and predictions in the bottom half.

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 pelagic fishes, elasmobranchs, turtles, sea mammals and two species of marine birds (see Table 1). They were chosen based on the availability of occurrence data, and special attention was given to their diversity in order to stress-test our model. The sample contains highly mobile and sessile species, widespread and local ones, some which live in biodiversity hotspots and others in less frequented waters. While sessile species (*Acropora*) cannot move in response to changing environmental conditions, the model may learn suitable conditions from geographical or long-term patterns, which could be useful to study the potential impact of temporary episodes (*e.g.* El Niño Southern Oscillation) or slower trends (*e.g.* ocean warming).

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	N samples	DOI
Yellowfin tuna	*Thunnus albacares*	9,998	10.15468/dl.gr2wbb
Longfin tuna	*Thunnus alalunga*	9,991	10.15468/dl.aqjv3y
Atlantic bluefin tuna	*Thunnus thynnus*	8,908	10.15468/dl.nnyeyb
Southern bluefin tuna	*Thunnus maccoyii*	2,022	10.15468/dl.tw97qj
Bigeye tuna	*Thunnus obesus*	9,999	10.15468/dl.c96qpp
Skipjack tuna	*Katsuwonus pelamis*	9,986	10.15468/dl.6y2zzm
Frigate Tuna	*Auxis thazard*	4,855	10.15468/dl.kfm6kq
Sailfish	*Istiophorus*	9,996	10.15468/dl.f48dug
Black marlin	*Istiompax indica*	705	10.15468/dl.b5acky
Blue marlin	*Makaira*	2,767	10.15468/dl.sygtaw
Swordfish	*Xiphias gladius*	9,996	10.15468/dl.hazqd2
Dolphinfish	*Coryphaena*	9,992	10.15468/dl.q67bqt
Humphead wrasse	*Cheilinus undulatus*	2,446	10.15468/dl.9g76hq
Oceanic Whitetip	*Carcharhinus longimanus*	2,160	10.15468/dl.b5ws4q
Whitetip	*Carcharhinus albimarginatus*	9,991	10.15468/dl.vpc772
Silk shark	*Carcharhinus falciformis*	9,998	10.15468/dl.vg4rwh
Sandbar shark	*Carcharhinus plumbeus*	9,993	10.15468/dl.7fczpa
Grey reef shark	*Carcharhinus amblyrhynchos*	10,000	10.15468/dl.ccqyws
Mako shark	*Isurus oxyrinchus*	6,240	10.15468/dl.h5akxk
Blue shark	*Prionace glauca*	9,973	10.15468/dl.zqkssk
Devil ray	*Mobula mobular*	1,064	10.15468/dl.p4e2sx
Reef manta	*Mobula alfredi*	7,928	10.15468/dl.bkjkgu
Eagle ray	*Myliobatis*	9,974	10.15468/dl.3u3v7k
Humpback whale	*Megaptera novaeangliae*	9,980	10.15468/dl.yzg4n3
Fin whale	*Balaenoptera physalus*	9,996	10.15468/dl.r9kaq8
Blue whale	*Balaenoptera musculus*	9,973	10.15468/dl.28f7xd
Bottlenose	*Tursiops*	9,952	10.15468/dl.bec9p4
Spinner dolphin	*Stenella longirostris*	7,394	10.15468/dl.xz5eds
Common dolphin	*Delphinus delphis*	9,974	10.15468/dl.u5be7v
Sperm whale	*Physeter macrocephalus*	9,984	10.15468/dl.7pf4ue
Harbour porpoise	*Phocoena phocoena*	9,937	10.15468/dl.afr2fn
Southern right whale	*Eubalaena australis*	9,963	10.15468/dl.e3hdkj
Green turtle	*Chelonia mydas*	9,835	10.15468/dl.6gs9rp
Loggerhead	*Caretta caretta*	9,941	10.15468/dl.dmb6ds
Hawksbill turtle	*Eretmochelys imbricata*	9,721	10.15468/dl.e6w44w
Emperor penguin	*Aptenodytes forsteri*	9,981	10.15468/dl.s5unhs
Wedge-tailed shearwater	*Puffinus pacificus*	9,964	10.15468/dl.vyztue
Acropora coral	*Acropora*	8,676	10.15468/dl.vg752f

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 for 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 (see Table 2). Furthermore, convolutional neural networks are known to be robust against occasional labelling mistakes (Chen et al., 2020).

Digital object identifiers (DOIs) for the download of each species are available in Table 1. When more than 10,000 occurrences of a taxon were available, 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). We removed points located on the continents and no other filtering was conducted on geographical precision.

We made the debatable choice not to intervene in the input data, as we cannot assume any consistent rules over all data sets. For example duplicates might be multiple sightings or simply mistakes. This choice is arbitrary and more thorough data cleaning could be beneficial to future work.

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

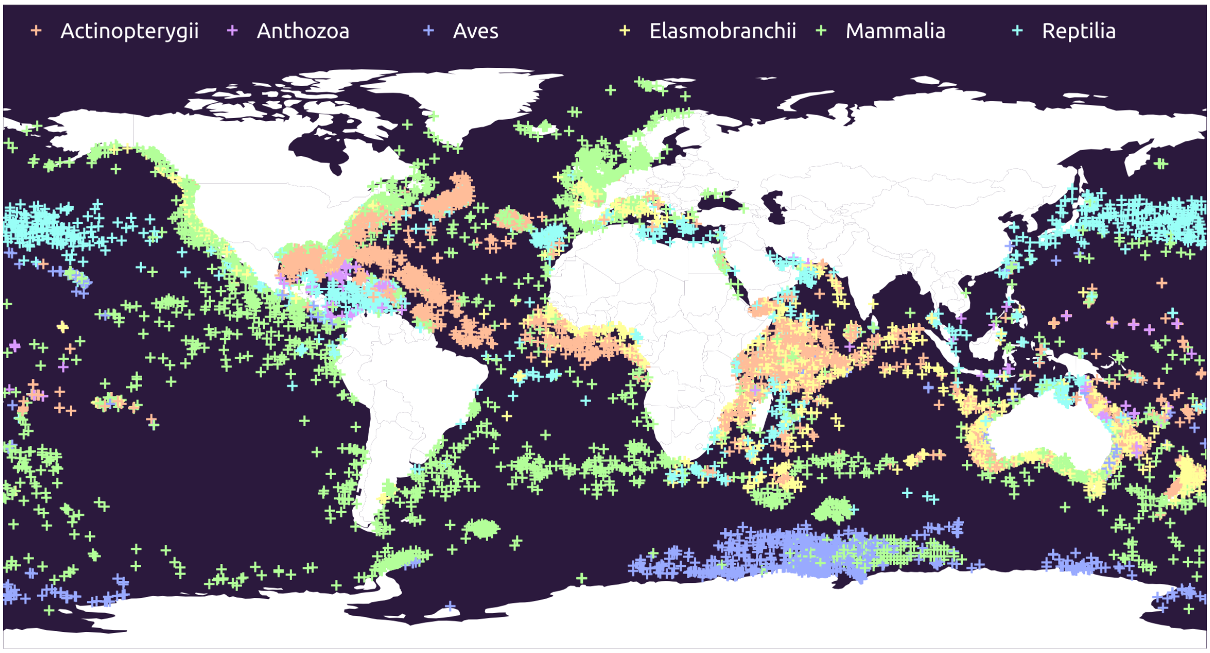


Figure 2 – Random sample (10%) of the training data set, coloured by taxonomic class.

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 values: both strength and orientation components (*i.e.*, polar coordinates) for finite-size Lyapunov exponents (FSLEs), and both zonal and meridional components for surface wind and geostrophic current. 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 (Moraes et al., 2012). Finally, four geographical variables were added (see Section 2.3.2). Even though these four variables were constant over each patch, they were encoded as layers with equal dimensions to the other variables. This was done for simplicity of implementation, as well as to take advantage of GPUs' efficiency at parallelized computation. Overall, this amounts to a total of 29 layers of data, shown in Table 2.

Each of these layers is two-dimensional: for 3D data only the surface layer was downloaded. However all the tools that we use in this study are compatible with 3D environmental data. We chose to focus on other aspects in this study, but the vertical dimension may be included into future work.

Table 2 – 29 layers used as input data (re-ordered for clarity).
NB: Eddy kinetic energy is calculated using geostrophic current data. Res. = Resolution.
P.S.U. = Practical salinity Unit. CMS = Copernicus Marine Service.

Variable	Source	Source type	Res.	Time Res.	Unit
Bathymetry	GEBCO, 2022	Observations	0.0042°		m
Salinity	European Union-CMS, 2020	Observations	0.25°	weekly	$P.S.U.$
Wave Height	European Union-CMS, 2021b	Observations	2°	daily	m
Surface wind (u)	CCMP (Mears et al., 2022)	Observations	0.25°	6 hours	$m.s^{-1}$
Surface wind (v)	CCMP (Mears et al., 2022)	Observations	0.25°	6 hours	$m.s^{-1}$
Oxygen	European Union-CMS, 2018, 2019	Models	0.25°	daily	$mmol.m^{-3}$
pH	European Union-CMS, 2018, 2019	Models	0.25°	monthly	
FSLEs (strength)	LOCEAN/CLS/CTOH/CNES, 2021	Observations	0.04°	daily	$days^{-1}$
FSLEs (orientation)	LOCEAN/CLS/CTOH/CNES, 2021	Observations	0.04°	daily	$degrees$
Geostrophic Current (u)	European Union-CMS, 2017, 2021a	Observations	0.25°	daily	$m.s^{-1}$
Geostrophic Current (v)	European Union-CMS, 2017, 2021a	Observations	0.25°	daily	$m.s^{-1}$
Eddy kinetic energy	Calculated		0.25°	daily	$m^2.s^{-2}$
Chlorophyll	OCCI (Sathyendranath et al., 2021)	Observations	0.042°	daily	$mg.m^{-3}$
Chlorophyll (D-5)	OCCI (Sathyendranath et al., 2021)	Observations	0.042°	daily	$mg.m^{-3}$
Chlorophyll (D-15)	OCCI (Sathyendranath et al., 2021)	Observations	0.042°	daily	$mg.m^{-3}$
SST	MUR (NASA/JPL, 2019)	Observations	0.25°	daily	$kelvin$
SST (D-5)	MUR (NASA/JPL, 2019)	Observations	0.25°	daily	$kelvin$
SST (D-15)	MUR (NASA/JPL, 2019)	Observations	0.25°	daily	$kelvin$
Mixed layer thickness	European Union-CMS, 2020	Observations	0.25°	weekly	m
Diatoms	European Union-CMS, 2022	Observations	4km	monthly	$mg.m^{-3}$
Dinophytes	European Union-CMS, 2022	Observations	4km	monthly	$mg.m^{-3}$
Haptophytes	European Union-CMS, 2022	Observations	4km	monthly	$mg.m^{-3}$
Green algae	European Union-CMS, 2022	Observations	4km	monthly	$mg.m^{-3}$
Prochlorophytes	European Union-CMS, 2022	Observations	4km	monthly	$mg.m^{-3}$
Prokaryotes	European Union-CMS, 2022	Observations	4km	monthly	$mg.m^{-3}$
Atlantic Ocean	Calculated				
Indian Ocean	Calculated				
Pacific Ocean	Calculated				
North hemisphere	Calculated				

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 is made available to other researchers for a wide range of uses in the GitHub IRDG2OI/geoenrich repository (Morand and Poulain, 2023). The package implements caching, so that it does not make requests to the server when data has been downloaded already.

Data was downloaded for the closest available date to the occurrence. A spatial buffer of 115 km 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 (Fromentin and Lopuszanski, 2014; Fujioka et al., 2018). All the available data within this buffer were downloaded into arrays.

These data arrays with various resolutions (minimum 1×1 for wave height, maximum 493×493 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. This grid has a resolution of approximately 7 km.

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: natural barriers prevent animals from navigating anywhere in the long term, namely continents and for some species, the warm waters around the equator.

Because we propose a type of dynamic SDM, we cannot capture these long-term barriers, so we have to include them artificially. 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.

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 replaced with the corresponding extrema and the scaling factors were saved to be reapplied to any subsequent input data.

Some data are missing 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. This does not allow the model to differentiate unobserved data (*e.g.* because of clouds) from no-data areas (*e.g.* coast), but this is not an issue as land pixels are already explicitly provided in the bathymetry layer.

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.

2.4. Training the model

The modelling technique that we describe in this study was developed for plant species distributions (Deneu et al., 2021). We used the *Malpolon* framework (Lorieul et al., 2023) after some adaptations to our use case. It was built on top of PyTorch (Paszke et al., 2019) and PyTorch Lightning (Falcon et al., 2020) frameworks.

The *Malpolon* framework implements a convolutional network with the *resnet50* feature extractor (He et al., 2016). We adapted it to use 29 inputs channels and 38 numerical outputs converted to relative probabilities by a Softmax function. Only the first (convolutional) layer and the last (linear + Softmax) layer were altered to adapt the number of inputs and outputs. It was trained from scratch in two sessions: one with a .1 learning rate and another one with a .01 learning rate to fine-tune the weights. We used a Binary Cross Entropy loss (averaged over the taxa and the training batch, weighted by taxa sample sizes).

Treating the problem as a classification task allows estimating the conditional probability of y (the observed species) given that an observation has been made in the environment x. It has the advantage to be (asymptotically) invariant to the spatial sampling effort but it is sensitive to the taxonomic reporting bias (the fact that some species are more observed than others) (Estopinan et al., 2024). In the absence of taxonomic reporting bias, the estimated probabilities would converge to the relative probability of each species given the environment. This is why sample size weights were used in the loss to compensate for this particular bias. Mapping those probabilities is thus equivalent to mapping the species suitability relatively to the other species suitability. It can be related to the "target-group" approach for generating pseudo-absences.

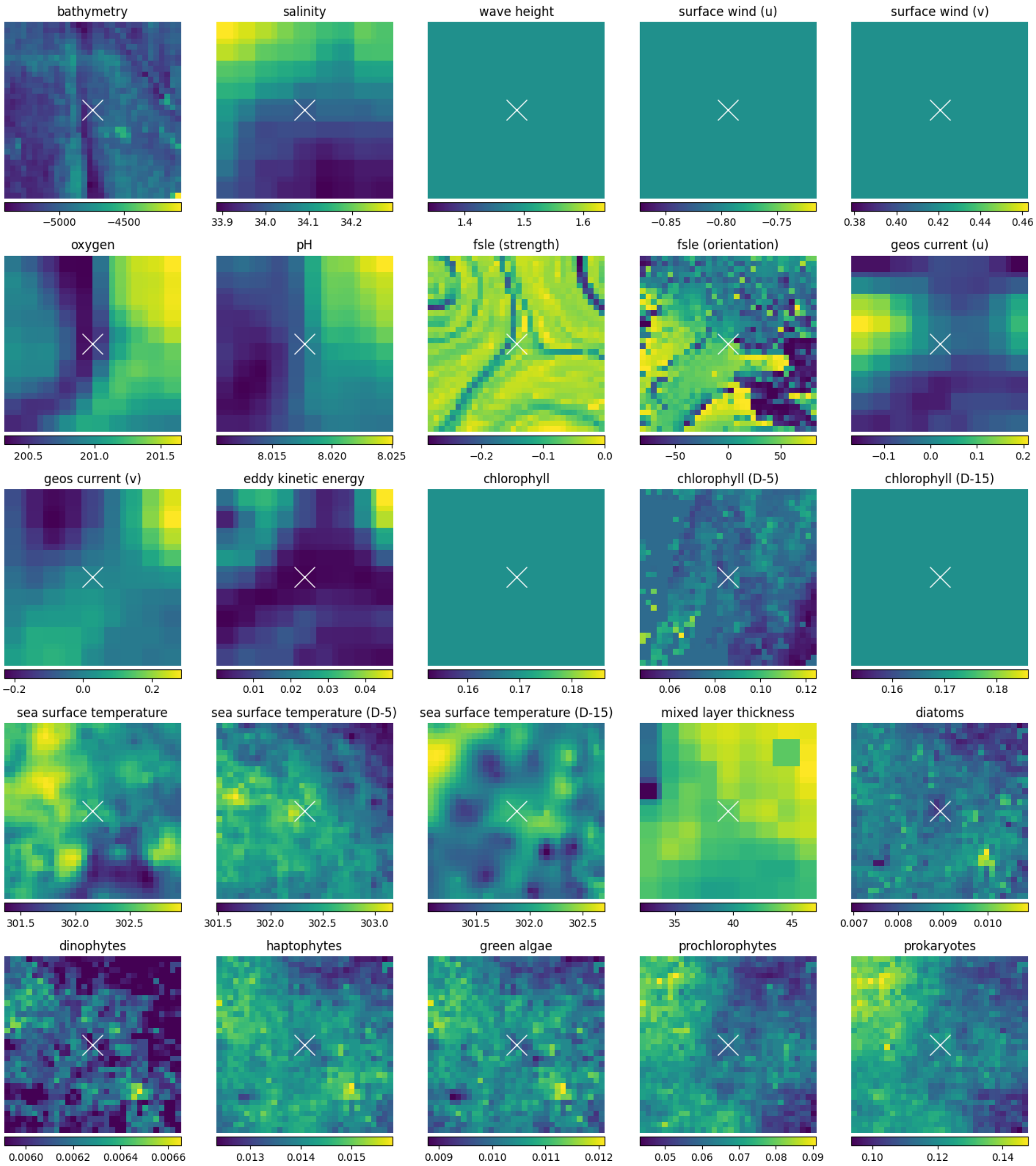


Figure 3 – Environmental variables around the point of coordinates -14.389°S, 78.918°E on March 20th, 2021.

Therefore, the target probabilities for training were set using one-hot encoding, *i.e.* a one for the observed species and zeroes for all others. This follows the principle of assumed negatives (Cole et al., 2021): 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, but it does not prevent co-occurrences. Indeed, if two species are present in the same environmental conditions, training will push the model towards a 50%-50% prediction.

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.

In order to assess the advantage of using a convolutional model, we removed geographical input data and retrained the same model from scratch, using only the center value of each patch, for each variable (Punctual-DNN (Deneu et al., 2021)). Since the tiles are 32x32 pixels, we had to use the average of the four center values. We then computed the same accuracy metrics as with our main model; they are displayed side by side in the Results Section.

2.6. From predicted probabilities 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, with an approximate 100 km stride. 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 predicted probabilities 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. We used cubic interpolation to generate outputs with a 3600x1800 pixels resolution for the World maps and 800x800 pixels for the Western Indian Ocean maps.

It is worth noting that since we are working with probabilities that are relative to our choice of studied taxa (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, which is a way to estimate the gradients of the scores with regard to the inputs, therefore gauging the importance of each input data point (Sundararajan et al., 2017). They were calculated for all points on the world grid at the four dates of 2021, using the Captum python package (Kokhlikyan et al., 2020). They were then aggregated over the whole study area (sum of absolute values) to represent variable importance over all the world oceans. Finally, to analyze this in a taxon-specific way, this process was repeated for each taxon on a random sample (N=1000) of the points where the taxon was the top prediction.

3. Results

3.1. Performance of the model

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. The corresponding score for the Punctual-DNN is significantly lower: 63%. See Table 3 for more complete accuracy results. These metrics prove the benefit of using spatial data, as hypothesized in the Introduction. Although the difference in scores is quite small, percentage points closer to 100% are much harder to gain than those close to 0%. Indeed, they represent the most difficult predictions.

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

Top N	Probability (Spatial input)	Probability (Punctual input)
1	69.15%	63.34%
2	83.19%	78.77%
3	89.13%	85.85%
4	92.83%	90.26%
5	95.13%	93.18%
6	96.63%	95.29%
7	97.48%	96.45%
8	97.99%	97.28%
9	98.43%	97.87%
10	98.75%	98.29%
38	100.00%	100.00%

All these statistics were computed on the test set, *i.e.* on data that was never used before.

A confusion matrix was computed on the test data set and is shown in Figure 4. It shows that some taxa were well predicted 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*. These two are among the taxa with the fewest occurrences, which could explain this result.

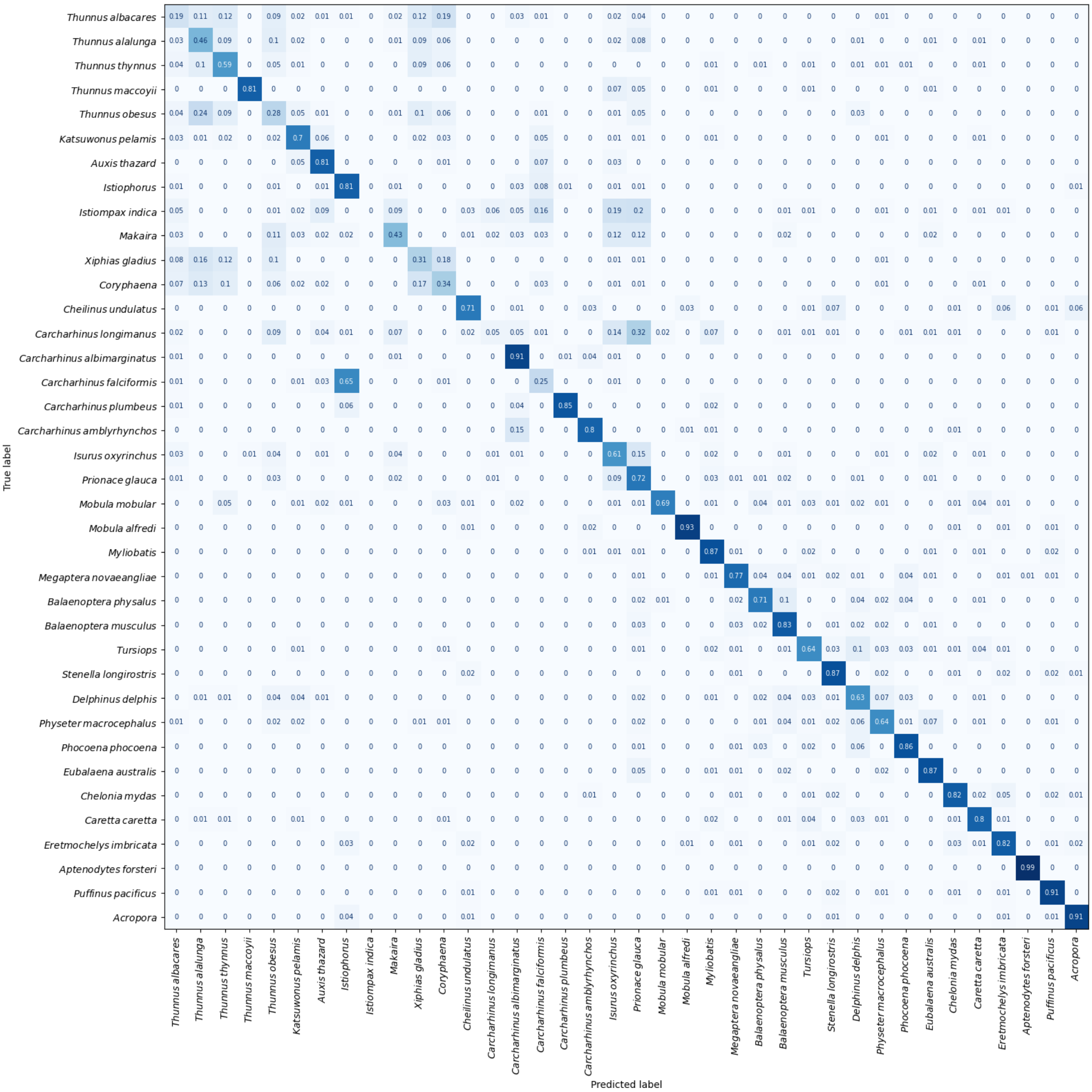
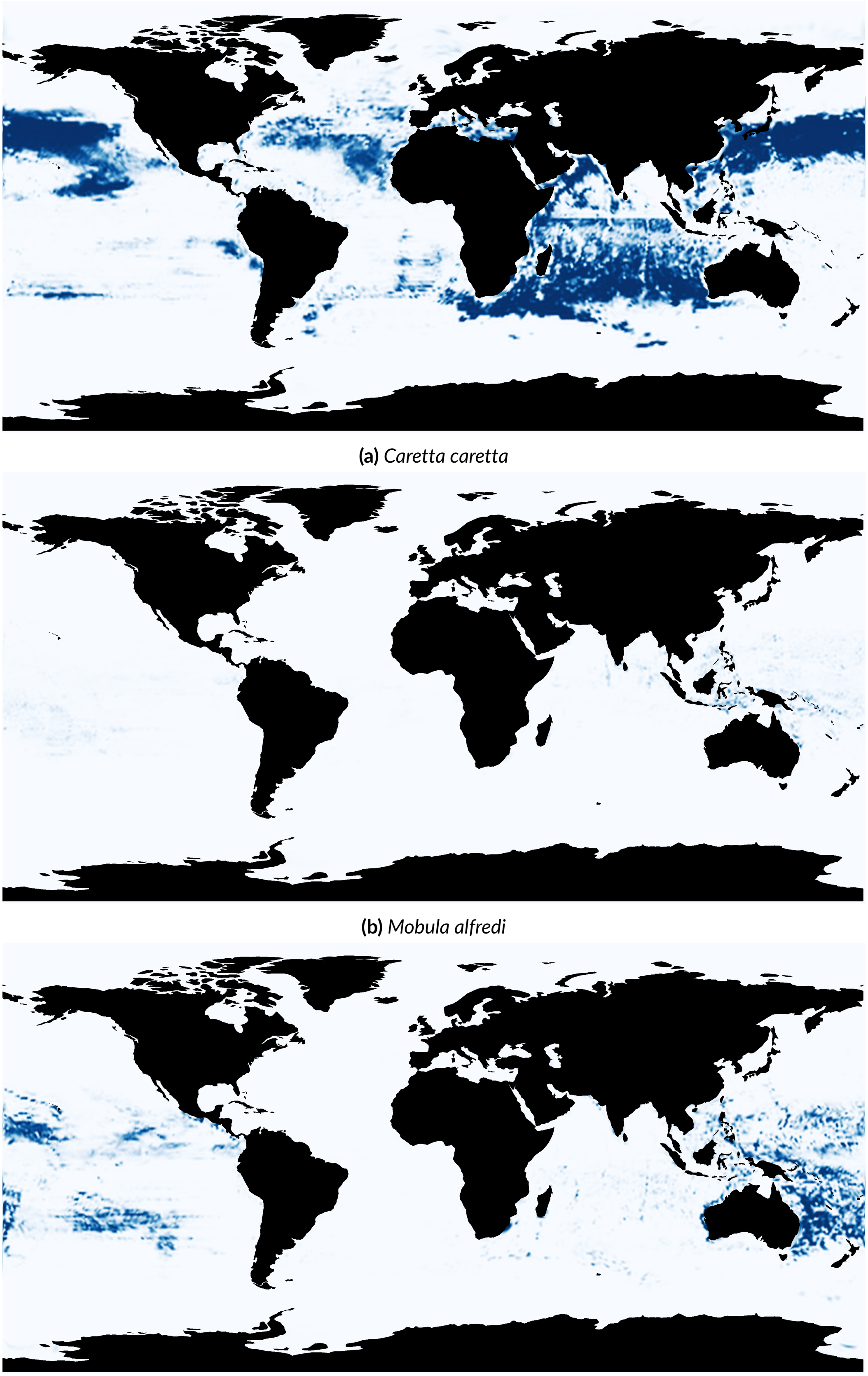


Figure 4 – The confusion matrix shows the predictions of the model on the test data set, for each actually observed taxon. Cell darkness is proportional to cross-probability.

3.2. Presentation of the species distributions maps

3.2.1. Global oceans. Distribution maps were calculated on four dates, all in 2021, corresponding to both solstices and both equinoxes, for the thirty-eight taxa. These maps represent the probability of presence *among the 38 studied genera*. Figure 5 shows these maps for three species on the spring equinox. All 152 distribution maps are available online (Morand, 2023d).



(a) *Caretta caretta*

(b) *Mobula alfredi*

(c) *Puffinus pacificus*

Figure 5 – Examples of distribution maps on March 20th, 2021, chosen to further discuss some interesting and contrasting patterns. All maps are available publicly (Morand, 2023d).

3.2.2. 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 (Morand, 2023c). Again, these maps represent the probability of presence *among the 38 studied genera*. An example for *Prionace glauca* is shown in Figure 6.

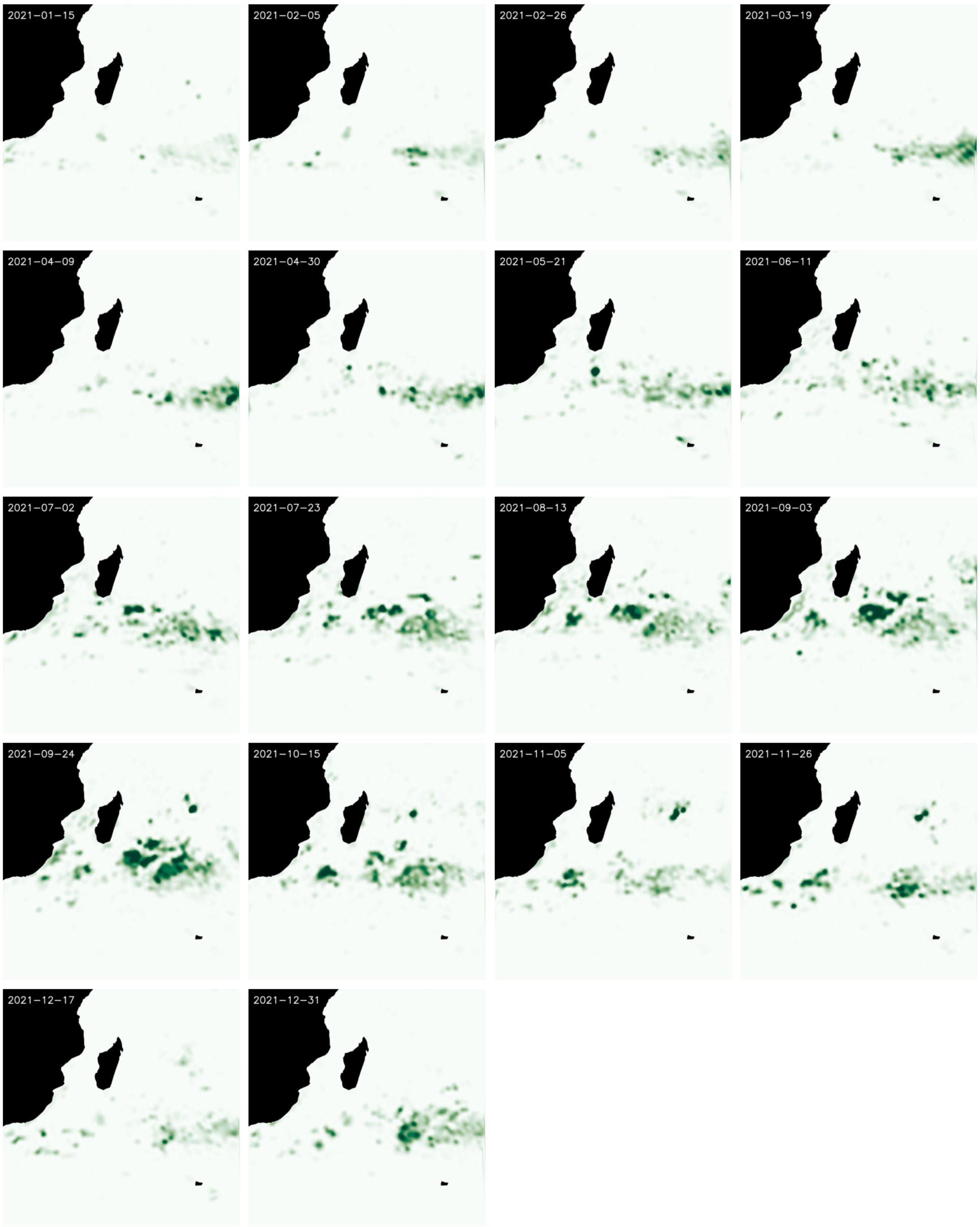
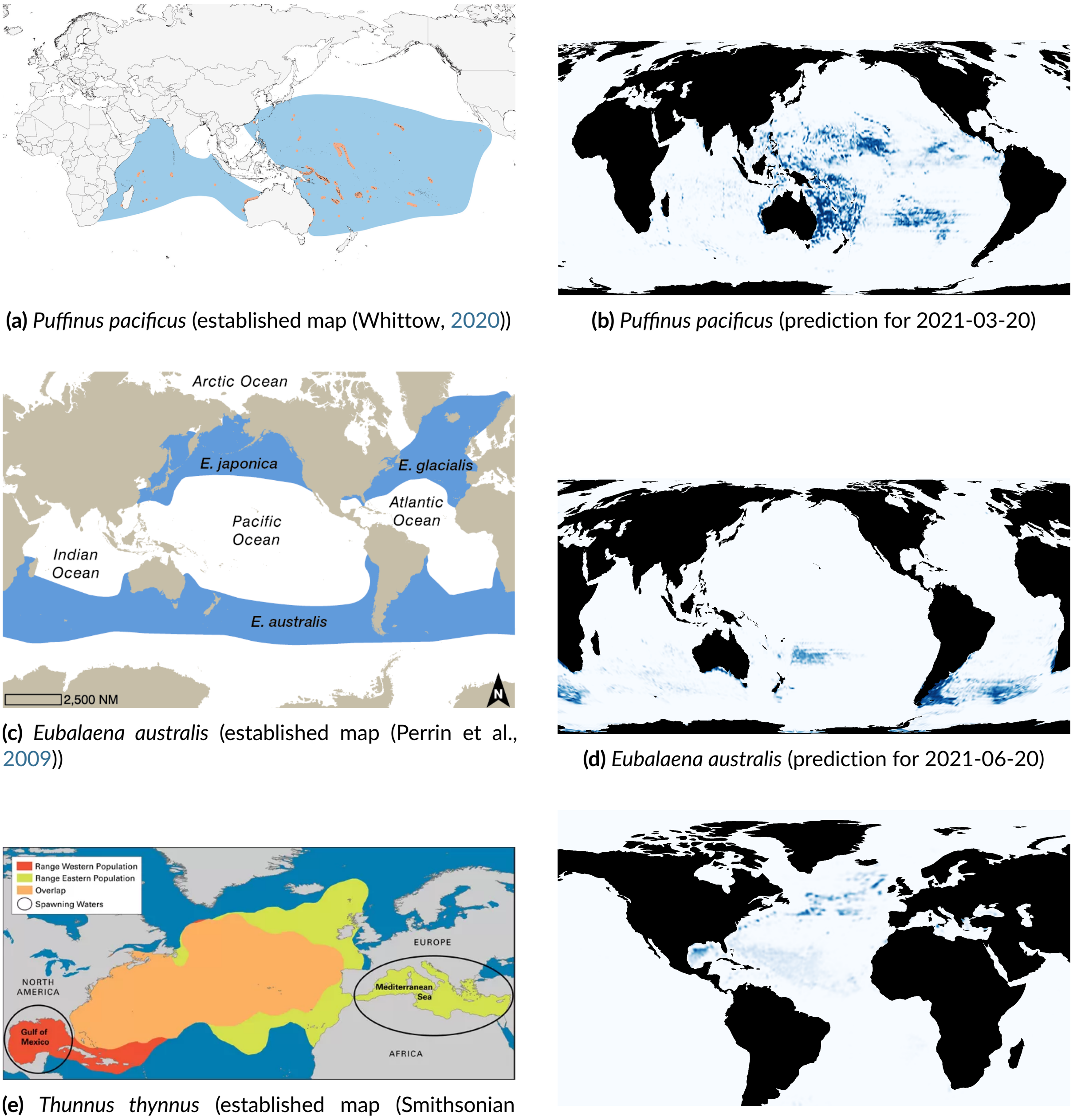


Figure 6 – Distribution maps for *Prionace glauca* every three weeks of 2021. The scale is different from other figures to improve visibility, hence the change in colours.

3.3. Comparison of predicted distribution maps to established maps

Validation of the distribution maps is challenging because existing distribution maps are usually the results of static studies (except sometimes broad seasonal variations). Yet the maps that we produce are dynamic, *i.e.* highly dependent on time, see Figure 6 for instance. Moreover, our maps show presence probabilities relatively to the set of 38 species, which yields results that are different in nature from classic distribution maps. Finally, observation data are spatially biased so they cannot be used for validation either.

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 (Whittow, 2020))

(b) *Puffinus pacificus* (prediction for 2021-03-20)

(c) *Eubalaena australis* (established map (Perrin et al., 2009))

(d) *Eubalaena australis* (prediction for 2021-06-20)

(e) *Thunnus thynnus* (established map (Smithsonian Ocean Team, 2009))

(f) *Thunnus thynnus* (prediction for 2023-03-20)

Figure 7 – Comparison between established distribution maps (left) (Perrin et al., 2009; Smithsonian Ocean Team, 2009; Whittow, 2020) and deep-learning generated maps (right). Prediction maps at other dates (Morand, 2023d) do not change the interpretation.

3.3.1. Puffinus pacificus. The prediction map 7a is consistent with the established one 7b, although it shows 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 the Indian Ocean stock may be an under-represented in our data set. In theory, this should not have impacted our results as the point of the method is to be geography-agnostic.

But in reality different stocks may have different reactions to environmental conditions, therefore introducing a correlation with geography (location of the stocks). Perhaps some stocks were under-represented in our training data set.

3.3.2. Eubalaena australis. The predicted distribution 7d fits within the known geographical range of *Eubalaena australis* 7c and the map shows 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. As our results are relative probabilities (*i.e.* proportions among all 38 species), variations in one distribution map may also ensue from variations in other species habitat preferences.

3.3.3. 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 a major population lives in these areas (Fromentin et al., 2014). 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.

3.4. Analysis of determining variables

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

Figure 8 shows the median integrated gradient for each taxon. While FSLEs and SST are the most important variables overall, this chart reveals the diversity of correlations between taxa and predictors. For example, bathymetry was the most important predictor for 6 taxa. Another interesting observation is that phytoplankton data was used significantly for one taxon only: *Thunnus thynnus*.

Table 4 – Statistics of the influence of variables over the 2021 predictions for the whole world ($\times 1,000$, sorted by median). Colour scale 0 to 1. MAD = Median Absolute Deviation.

Variable	min	25%	50%	75%	max	MAD
FSLEs (strength)	0.00	0.42	0.73	1.21	6.88	0.36
SST	0.00	0.17	0.44	0.95	5.13	0.33
pH	0.00	0.19	0.38	0.72	5.04	0.23
Salinity	0.00	0.19	0.32	0.63	3.56	0.17
FSLEs (orientation)	0.00	0.16	0.30	0.56	3.85	0.17
Bathymetry	0.00	0.11	0.28	0.66	8.39	0.22
Geos. Current (u)	0.00	0.13	0.24	0.41	3.25	0.12
Geos. Current (v)	0.00	0.14	0.24	0.40	2.93	0.12
Surface wind (v)	0.00	0.13	0.23	0.38	1.48	0.11
Surface wind (u)	0.00	0.12	0.20	0.34	1.43	0.10
Oxygen	0.00	0.05	0.20	0.61	4.26	0.18
Wave height	0.00	0.07	0.11	0.19	0.83	0.06
Mixed layer thickness	0.00	0.03	0.09	0.28	3.83	0.07
Pacific Ocean	0.00	0.00	0.08	0.57	3.29	0.08
Green algae	0.00	0.01	0.03	0.06	2.88	0.02
Prochlorophytes	0.00	0.01	0.03	0.06	3.62	0.02
Haptophytes	0.00	0.01	0.02	0.06	2.93	0.02
Prokaryotes	0.00	0.01	0.01	0.03	2.24	0.01
Chlorophyll	0.00	0.00	0.01	0.02	1.23	0.01
Dinophytes	0.00	0.00	0.01	0.02	3.14	0.01
Diatoms	0.00	0.00	0.00	0.02	1.96	0.00
North hemisphere	0.00	0.00	0.00	0.48	3.84	0.00
Atlantic Ocean	0.00	0.00	0.00	0.12	3.27	0.00
Indian Ocean	0.00	0.00	0.00	0.00	3.35	0.00

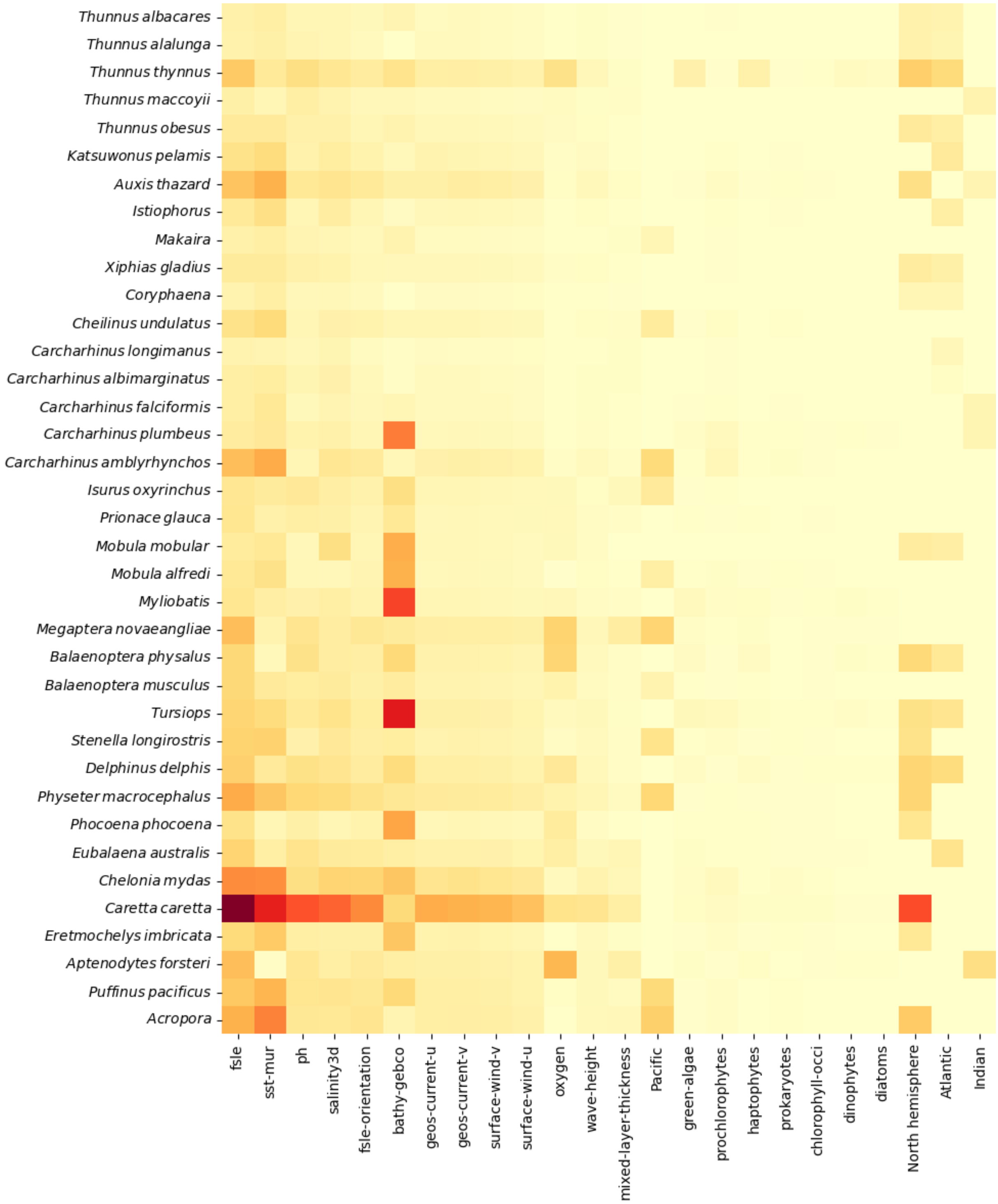


Figure 8 – Variables that had the most influence on the determination of each taxon presence (darker = stronger influence). NB: *Istiompax indica* is not present in this chart as it was never predicted to be the most likely taxon.

4. Discussion

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

The variables that were identified as important are coherent with past research. Specifically, FSLEs were identified as a particularly important predictor of movement for top marine predators (Tew Kai et al., 2009). Sea surface temperature was also expected to be an important predictor, since it has important physiological consequences and is therefore the most frequently used descriptor in marine SDMs (Melo-Merino et al., 2020) and was identified as the most relevant factor in an SDM review (Bosch et al., 2018).

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 and is coherent with previous findings (Bateman et al., 2012; Melo-Merino et al., 2020).

We noticed 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. This could be a legitimate and expected behavior, or overfitting due to auto-correlated training data, which is discussed in the next section.

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. Indeed we showed that the convolutional part of the model was taking advantage of spatial data, which lead to significantly higher accuracy than using data only from the point of occurrence. 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 metric. The present model is a classifier and as such, it behaves differently from usual SDMs. In particular, outputs are predicted probabilities, relatively to the set of 38 taxa. Therefore, prediction maps cannot be interpreted as the usual results of SDMs. For example, if a species is obviously present in some environmental conditions, probabilities for other species will be lower. A solution to this shortcoming would be to include a large number of species into the study (*e.g.* 4,520 in (Deneu et al., 2021)), which is a priority for any reproduction of this work.

The accuracy of the model could still be improved, depending on the ecological feasibility. Indeed, as individuals, members of a species may explore, behave erratically, or in any other way exercise their free will or at least their individual preference (Cerqueira et al., 2016). Their response to environmental predictors may even depend on the environment itself (Muñoz et al., 2015). As such, dynamic SDMs will never provide a perfect prediction of their distribution.

Furthermore, if some species are frequently seen together, the model cannot 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 habitat preferences, but this would remove the possibility of studying differences between such 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.

Consequently, accuracy is not an ideal metric for this use case, as we use a classifier to bypass the scarcity of training data, in particular the fact that almost all available data are presence-only. This should not be a deterrent as previous research showed that presence-only occurrence data could yield satisfactory results (Elith et al., 2006). This is why we provide 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 (Clegg et al., 2022).

The fact that the model has limited access to geographical information (only hemisphere and ocean basin) partly compensates for sampling effort heterogeneity. Indeed, only environmental conditions guide the predictions. But this is not flawless when various stocks of the same species have different behavior relatively to environmental conditions. This is the case of *Thunnus thynnus* which has two separate stocks (West and East Atlantic) (Viñas et al., 2011). When sampling is biased between stocks, the model might not fully learn the various responses to environmental responses. 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 may 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 ocean. 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, as previous research even showed that more data was more important than spatial bias in our context (Gaul et al., 2020).

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 (Rizzo and Schulte, 2009). The model also did not catch the *Thunnus thynnus* seasonal spawning in summer in the Mediterranean Sea.

The causes of these shortcomings are unclear, so we offer a range of possible explanations and ways to improve the present method, in the hope that these will help future research obtain higher quality results.

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, observer bias could be reduced by selecting data sources more appropriately. In particular, redundant data sets should be avoided and more importance should be given to the diversity of sampling methods.

As previously discussed, this method would probably benefit from including a large number of taxa. In particular, planktonic species may prove valuable as they are less prone to sampling biases (Gregg et al., 2017) and data sets are widely available (Righetti et al., 2020).

Further, It could be interesting to run the model on two separate occurrences data sets: one with fast moving species only and the other with sedentary or sessile species only. This would allow testing the efficiency of dynamic SDMs in two different contexts: real-time environmental conditions preferences and long-term distribution shifts, respectively.

4.3.2. Environmental data. For some variables, it could be beneficial to use other sources. For example the chlorophyll data that we used was quite incomplete and using the Copernicus product instead (European Union-CMS, 2022) could yield better results. It has also been suggested to include 3D environmental data, as most variables vary with depth and occurrence data are not limited to the surface (Duffy and Chown, 2017). Such data could easily be included in the input data with no change to our method. Finally, additional data may be beneficial, in particular the distance to the nearest coastline or level of anthropisation.

The encoding of variables could also be experimented with. In particular, vectors (FSLEs, wind and current) could be encoded with three variables instead of two: strength, cos(angle) and sin(angle). This would make the North-South and East-West components, as well as the total strength explicit.

4.3.3. Model training. Several choices could be made differently during the training phase. For example, *Acropora* presence is predicted in the open waters of the Pacific (too much to correspond to Pacific atolls), even though it is only present at very low depths in both the training and testing data sets. This may be the result of overfitting, due to autocorrelation between the training and validation data sets (Nurunnabi and Teferle, 2022). This is not visible in the covariance matrix because in that case the testing data set is also correlated with the two others. To remedy this, the split between the training, validation and testing data sets could be more sophisticated, by using block cross-validation or withholding a region/period only for testing, or even more complex methods such as adding a time lag to some observations (Zeraati et al., 2022).

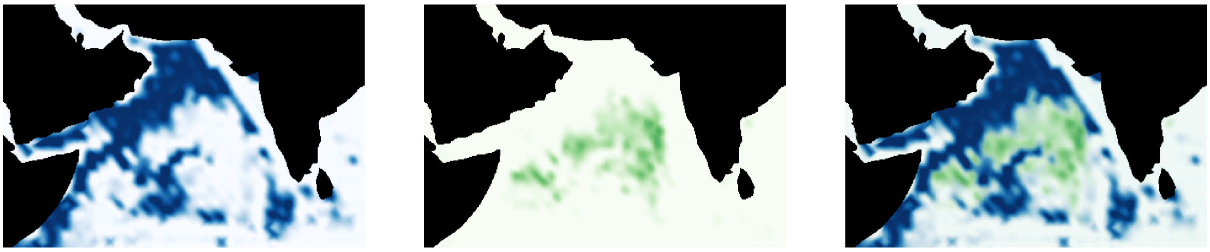
Second, although we experimented with loss functions, this can be continued to try and find alternatives more adapted to this context.

4.3.4. Removing artefacts. Three types of graphical artefacts are present in our results:

- The sharp divide between geographical areas caused by our binary geographical variables (see the Indian Ocean in Figure 5a for instance). Ideally, for a given species, only barriers outside of its range should be significant and therefore barriers should not affect the maps. Indeed, we included these artificial barriers as proxies of the historical zoogeographical barriers to colonization (Briggs, 1974). But the imperfections of the model and the low number of species (see third point) contributed to this flaw in the maps.

  This type of artefact could be mitigated by using a gradient between corresponding binary variables in the zones where theses areas meet. A more drastic solution would be to fully remove the binary geographical variables. This would imply either **1.** using the model only on smaller regions with full connectivity or **2.** accepting that the model predicts theoretical habitat suitability, independently from actual species presence.
- The spotted aspect of the map, which shows that the model uses small scale features for its prediction. Further investigations need to be conducted to determine whether they are justified by the environmental preferences of taxa or if they are the result of overfitting.
- The third one is less visible, but it is a consequence of our predicted probabilities being relative to the 38 species. When an area is favorable to species A, but species B dominates in part of this area, the distribution map for species A shows variations over the area (regardless of actual variations in suitability to species A). Figure 9 shows an example of this phenomenon where the presence of *Katsuwonus pelamis* causes a hole in the distribution of *Caretta caretta*.

  This makes it harder to interpret the maps, and a solution would be to increase the number of species, as previously suggested in other sections. This way the variation of suitability to one species would only marginally influence probabilities for other species. The present study shows that $N = 38$ is not enough to avoid this type of artefact.



(a) *Caretta caretta* **(b)** *Katsuwonus pelamis* **(c)** Both species

Figure 9 – Example of how one species distribution can influence results for an other.

4.3.5. 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

Dynamic SDMs provide a way to estimate species presence at all dates and all areas, provided environmental data is available. In addition, the present method leverages environmental data around occurrences, including complex patterns. While the dynamic nature makes it difficult to judge accuracy (available reference data are static), 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 (Deneu et al., 2021).

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. Our method would also greatly benefit from their input, as we do not have the species-specific expertise that is necessary to fine-tune training and predictions.

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.

Finally, we expect that experts of different taxa will rightfully criticize the maps we provide. We wish they did not have to, but developing such a model inevitably includes a trial and error phase, so we welcome their remarks which will lead to investigating issues and proposing improvements for subsequent studies.

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.

Acknowledgements

Preprint version 3 of this article has been peer-reviewed and recommended by Peer Community In Ecology (https://doi.org/10.24072/pci.ecology.100584; Munoz (2024)).

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.

This study was significantly improved thanks to numerous and thorough reviews from Jean-Olivier Irisson, Sakina-Dorothee Ayata and a third reviewer who chose to remain anonymous. We are also grateful for the platform provided by PCI Ecology for ethical reviewing and publishing, and the comments by recommender François Munoz.

Funding

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

Conflict of interest disclosure

The authors declare they comply with the PCI rule of having no financial conflicts of interest.

Data, script, code, and supplementary information availability

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 (https://doi.org/10.5281/zenodo.10809445) (Morand, 2024).

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 (https://doi.org/10.5281/zenodo.8188512) (Morand, 2023a) 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 (https://doi.org/10.5281/zenodo.8202914) (Morand, 2023b), 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 (https://doi.org/10.5281/zenodo.10809445) (Morand, 2024). It consists of Python files based on a custom version of Malpolon.

Results

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

- Global predictions as PNGs and GeoTIFFs (https://doi.org/10.5281/zenodo.8202261) (Morand, 2023d)
- Western Indian Ocean predictions as GIFs and GeoTIFFs (https://doi.org/10.5281/zenodo.8202056) (Morand, 2023c)

References

Barrón C, Duarte CM (2015). *Dissolved organic carbon pools and export from the coastal ocean*. *Global Biogeochemical Cycles* **29**, 1725–1738. https://doi.org/10.1002/2014GB005056.

Bateman BL, VanDerWal J, Johnson CN (2012). *Nice weather for bettongs: using weather events, not climate means, in species distribution models*. *Ecography* **35**, 306–314. https://doi.org/10.1111/j.1600-0587.2011.06871.x.

Baudena A, Ser-Giacomi E, D'Onofrio D, Capet X, Cotté C, Cherel Y, D'Ovidio F (2021). *Fine-scale structures as spots of increased fish concentration in the open ocean*. *Scientific Reports* **11**, 15805. https://doi.org/10.1038/s41598-021-94368-1.

Bosch S, Tyberghein L, Deneudt K, Hernandez F, De Clerck O (2018). *In search of relevant predictors for marine species distribution modelling Using the MarineSPEED benchmark dataset*. *Diversity and Distributions* **24**. Ed. by Alexandra Syphard, 144–157. https://doi.org/10.1111/ddi.12668.

Botella C, Joly A, Bonnet P, Monestiez P, Munoz F (2018). *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, pp. 169–199. https://doi.org/10.1007/978-3-319-76445-0_10.

Briggs JC (1974). *Operation of zoogeographic barriers*. *Systematic Biology* **23**, 248–256. https://doi.org/10.1093/sysbio/23.2.248.

Brodie S, Hobday AJ, Smith JA, Everett JD, Taylor MD, Gray CA, Suthers IM (2015). *Modelling the oceanic habitats of two pelagic species using recreational fisheries data*. *Fisheries Oceanography* **24**, 463–477. https://doi.org/10.1111/fog.12122.

Cerqueira M, Rey S, Silva T, Featherstone Z, Crumlish M, MacKenzie S (2016). *Thermal preference predicts animal personality in Nile Tilapia* Reochromis niloticus. *Journal of Animal Ecology* **85**. Ed. by Sissel Jentoft, 1389–1400. https://doi.org/10.1111/1365-2656.12555.

Chen R, Wang M, Lai Y (2020). *Analysis of the role and robustness of artificial intelligence in commodity image recognition under deep learning neural network*. *PLOS ONE* **15**, e0235783. https://doi.org/10.1371/journal.pone.0235783.

Clegg TL, Fuglebakk E, Ono K, Vølstad JH, Nedreaas K (2022). *A simulation approach to assessing bias in a fisheries self-sampling programme*. *ICES Journal of Marine Science* **79**. Ed. by Ernesto Jardim, 76–87. https://doi.org/10.1093/icesjms/fsab242.

Cole E, Mac Aodha O, Lorieul T, Perona P, Morris D, Jojic N (2021). *Multi-label learning from single positive labels*. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 933–942. https://doi.org/10.1109/cvpr46437.2021.00099.

Deneu B, Servajean M, Bonnet P, Botella C, Munoz F, Joly A (2021). *Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment*. *PLOS Computational Biology* **17**, e1008856. https://doi.org/10.1371/journal.pcbi.1008856.

Duffy GA, Chown SL (2017). *Explicitly integrating a third dimension in marine species distribution modelling*. *Marine Ecology Progress Series* **564**, 1–8. https://doi.org/10.3354/meps12011.

Elith J, H. Graham C, P. Anderson R, Dudík M, Ferrier S, Guisan A, J. Hijmans R, Huettmann F, R. Leathwick J, Lehmann A, Li J, G. Lohmann L, A. Loiselle B, Manion G, Moritz C, Nakamura M, Nakazawa Y, McC. M. Overton J, Townsend Peterson A, J. Phillips S, et al. (2006). *Novel methods improve prediction of species' distributions from occurrence data*. *Ecography* **29**, 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x.

Estopinan J, Servajean M, Bonnet P, Joly A, Munoz F (2024). *AI-based mapping of the conservation status of orchid assemblages at global scale*. Version 1. https://doi.org/10.48550/ARXIV.2401.04691. preprint.

European Union-CMS (2017). *Global ocean gridded L4 sea surface heights and derived variables NRT*. Mercator Ocean International. https://doi.org/10.48670/MOI-00149.

European Union-CMS (2018). *Global ocean biogeochemistry hindcast*. Mercator Ocean International. https://doi.org/10.48670/MOI-00019.

European Union-CMS (2019). *Global ocean biogeochemistry analysis and forecast*. Mercator Ocean International. https://doi.org/10.48670/MOI-00015.

European Union-CMS (2020). *Multi observation global ocean 3D temperature salinity height geostrophic current and MLD*. Mercator Ocean International. https://doi.org/10.48670/MOI-00052.

European Union-CMS (2021a). *Global ocean gridded L4 sea surface heights and derived variables reprocessed (1993-ongoing)*. Mercator Ocean International. https://doi.org/10.48670/MOI-00148.

European Union-CMS (2021b). *Global ocean L4 significant wave height from reprocessed satellite measurements*. Mercator Ocean International. https://doi.org/10.48670/MOI-00177.

European Union-CMS (2022). *Global ocean colour (Copernicus-GlobColour), bio-geo-chemical, L4 (monthly and interpolated) from satellite observations (1997-ongoing)*. Mercator Ocean International. https://doi.org/10.48670/MOI-00281.

Falcon W, Borovec J, Wälchli A, Eggert N, Schock J, Jordan J, Skafte N, Ir1dXD, Bereznyuk V, Harris E, Tullie Murrell, Yu P, Præsius S, Addair T, Zhong J, Lipin D, Uchida S, Shreyas Bapat, Schröter H, Dayma B, et al. (2020). *PyTorchLightning/Pytorch-Lightning: 0.7.6 release*. Version 0.7.6. Zenodo. https://doi.org/10.5281/ZENODO.3828935.

Fernandez M, Yesson C, Gannier A, Miller PI, Azevedo JM (2017). *The importance of temporal resolution for niche modelling in dynamic marine environments*. *Journal of Biogeography* **44**, 2816–2827. https://doi.org/10.1111/jbi.13080.

Fromentin JM, Lopuszanski D (2014). *Migration, residency, and homing of Bluefin Tuna in the western Mediterranean Sea*. *ICES Journal of Marine Science* **71**, 510–518. https://doi.org/10.1093/icesjms/fst157.

Fromentin JM, Reygondeau G, Bonhommeau S, Beaugrand G (2014). *Oceanographic changes and exploitation drive the spatio-temporal dynamics of Atlantic Bluefin Tuna* (Thunnus thynnus. *Fisheries Oceanography* **23**, 147–156. https://doi.org/10.1111/fog.12050.

Fujioka K, Fukuda H, Tei Y, Okamoto S, Kiyofuji H, Furukawa S, Takagi J, Estess E, Farwell CJ, Fuller DW, Suzuki N, Ohshimo S, Kitagawa T (2018). *Spatial and temporal variability in the trans-Pacific migration of Pacific Bluefin Tuna (Thunnus orientalis) revealed by archival tags*. *Progress in Oceanography* **162**, 52–65. https://doi.org/10.1016/j.pocean.2018.02.010.

Ganzeveld L, Helmig D, Fairall CW, Hare J, Pozzer A (2009). *Atmosphere-ocean ozone exchange: a global modeling study of biogeochemical, atmospheric, and waterside turbulence dependencies*. *Global Biogeochemical Cycles* **23**, 2008GB003301. https://doi.org/10.1029/2008GB003301.

Gaul W, Sadykova D, White HJ, Leon-Sanchez L, Caplat P, Emmerson MC, Yearsley JM (2020). *Data quantity is more important than its spatial bias for predictive species distribution modelling*. *PeerJ* **8**, e10411. https://doi.org/10.7717/peerj.10411.

GBIF (2023), https://www.gbif.org/.

GEBCO (2022). *The GEBCO_2022 grid - a continuous terrain model of the global oceans and land.* Version 1. NERC EDS British Oceanographic Data Centre NOC. https://doi.org/10.5285/E0F0BB80-AB44-2739-E053-6C86ABC0289C.

Gregg WW, Rousseaux CS, Franz BA (2017). *Global trends in ocean phytoplankton: a new assessment using revised ocean colour data*. *Remote Sensing Letters* **8**, 1102–1111. https://doi.org/10.1080/2150704X.2017.1354263.

Guisan A, Thuiller W (2005). *Predicting species distribution: offering more than simple habitat models*. *Ecology Letters* **8**, 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x.

Guisan A, Zimmermann NE (2000). *Predictive habitat distribution models in ecology*. *Ecological Modelling* **135**, 147–186. https://doi.org/10.1016/S0304-3800(00)00354-9.

He K, Zhang X, Ren S, Sun J (2016). *Deep residual learning for image recognition*. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. https://doi.org/10.1109/cvpr.2016.90.

IPCC (2019). *Summary for policymakers*. In: *Special Report on the Ocean and Cryosphere in a Changing Climate*. Cambridge University Press. https://doi.org/10.1017/9781009157964.001.

Jackson JBC, Kirby MX, Berger WH, Bjorndal KA, Botsford LW, Bourque BJ, Bradbury RH, Cooke R, Erlandson J, Estes JA, Hughes TP, Kidwell S, Lange CB, Lenihan HS, Pandolfi JM, Peterson CH, Steneck RS, Tegner MJ, Warner RR (2001). *Historical overfishing and the recent collapse of coastal ecosystems*. *Science (New York, N.Y.)* **293**, 629–637. https://doi.org/10.1126/science.1059199.

Kokhlikyan N, Miglani V, Martin M, Wang E, Alsallakh B, Reynolds J, Melnikov A, Kliushkina N, Araya C, Yan S, Reblitz-Richardson O (2020). *Captum: a unified and generic model interpretability library for PyTorch*. https://doi.org/10.48550/arXiv.2009.07896.

LOCEAN/CLS/CTOH/CNES (2021). *FSLE - Finite-Size Lyapunov Exponents and orientations of the associated eigenvectors (Version DT2021)*. [Dataset]. CNES. https://doi.org/10.24400/527896/A01-2022.002.

Lorieul T, Larcher T, Joly A (2023), Plantnet/Malpolon: Deep–SDM framework. https://github.com/plantnet/malpolon.

Macías-Zamora JV (2011). *Chapter 19 - Ocean Pollution*. In: *Waste*. Ed. by Trevor M. Letcher and Daniel A. Vallero. Boston: Academic Press, pp. 265–279. https://doi.org/10.1016/B978-0-12-381475-3.10019-1.

Mannocci L, Boustany AM, Roberts JJ, Palacios DM, Dunn DC, Halpin PN, Viehman S, Moxley J, Cleary J, Bailey H, Bograd SJ, Becker EA, Gardner B, Hartog JR, Hazen EL, Ferguson MC, Forney KA, Kinlan BP, Oliver MJ, Perretti CT, et al. (2017). *Temporal resolutions in species distribution models of highly mobile marine animals: recommendations for ecologists and managers*. *Diversity and Distributions* **23**. Ed. by Maria Beger, 1098–1109. https://doi.org/10.1111/ddi.12609.

Mears C, Lee T, Ricciardulli L, Wang X, Wentz F (2022). *RSS cross-calibrated multi-platform (CCMP) 6-hourly ocean vector wind analysis on 0.25 Deg grid, version 3.0*. Remote Sensing Systems. https://doi.org/10.56236/RSS-uv6h30.

Melo-Merino SM, Reyes-Bonilla H, Lira-Noriega A (2020). *Ecological niche models and species distribution models in marine environments: a literature review and spatial analysis of evidence*. *Ecological Modelling* **415**, 108837. https://doi.org/10.1016/j.ecolmodel.2019.108837.

Milanesi P, Della Rocca F, Robinson RA (2020). *Integrating dynamic environmental predictors and species occurrences: toward true dynamic species distribution models*. *Ecology and Evolution* **10**, 1087–1092. https://doi.org/10.1002/ece3.5938.

Miller J (2010). *Species distribution modeling*. *Geography Compass* **4**, 490–509. https://doi.org/10.1111/j.1749-8198.2010.00351.x.

Miller PI, Christodoulou S (2014). *Frequent locations of oceanic fronts as an indicator of pelagic diversity: application to marine protected areas and renewables*. *Marine Policy* **45**, 318–329. https://doi.org/10.1016/j.marpol.2013.09.009.

Moraes LE, Paes E, Garcia A, Jr OM, Vieira J (2012). *Delayed response of fish abundance to environmental changes: a novel multivariate time-lag approach*. *Marine Ecology Progress Series* **456**, 159–168. https://doi.org/10.3354/meps09731.

Morand G (2023a). *Deep-SDMs in the open oceans - INPUT DATA*. Zenodo. https://doi.org/10.5281/zenodo.8188512.

Morand G (2023b). *Deep-SDMs in the open oceans - MODEL CHECKPOINT*. https://doi.org/10.5281/zenodo.8202914.

Morand G (2023c). *Deep-SDMs in the open oceans - OUTPUTS - western Indian Ocean*. https://doi.org/10.5281/zenodo.8202056.

Morand G (2023d). *Deep-SDMs in the open oceans - OUTPUTS - World*. https://doi.org/10.5281/zenodo.8202261.

Morand G (2024). *Deep-SDMs in the open oceans - CODE*. Version 2. Zenodo. https://doi.org/10.5281/zenodo.10809445.

Morand G, Poulain S (2023). *GeoEnrich v0.5.8: a new tool for scientists to painlessly enrich species occurrence data with environmental variables*. Version v0.5.8. https://doi.org/10.5281/ZENODO.6458090.

Munoz F (2024). *The potential of convolutional neural networks for modeling species distributions*. *Peer Community in Ecology*, 100584. https://doi.org/10.24072/pci.ecology.100584.

Muñoz AR, Márquez AL, Real R (2015). *An approach to consider behavioral plasticity as a source of uncertainty when forecasting species' response to climate change*. *Ecology and Evolution* **5**, 2359–2373. https://doi.org/10.1002/ece3.1519.

NASA/JPL (2019). *GHRSST Level 4 MUR 0.25deg global foundation sea surface temperature analysis (v4.2)*. NASA Physical Oceanography DAAC. https://doi.org/10.5067/GHM25-4FJ42.

Nurunnabi A, Teferle FN (2022). *Resampling methods for a reliable validation set in deep learning based point cloud classification*. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **XLIII-B2-2022**, 617–624. https://doi.org/10.5194/isprs-archives-XLIII-B2-2022-617-2022.

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, et al. (2019). *PyTorch: an imperative style, high-performance deep learning library*. https://doi.org/10.48550/arXiv.1912.01703.

Perrin WF, Würsig B, Thewissen JGM (2009). *Right Whales*. In: *Encyclopedia of Marine Mammals*. ISBN: 9780080919935. Academic Press.

Raffaelli D, Solan M, Webb TJ (2005). *Do marine and terrestrial ecologists do it differently? Marine Ecology Progress Series* **304**, 283–289. JSTOR: 24869863.

Ramos AG, Santiago J, Sangra P, Canton M (1996). *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*. *International Journal of Remote Sensing* **17**, 749–759. https://doi.org/10.1080/01431169608949042.

Righetti D, Vogt M, Zimmermann NE, Guiry MD, Gruber N (2020). *PhytoBase: a global synthesis of open-ocean phytoplankton occurrences*. *Earth System Science Data* **12**, 907–933. https://doi.org/10.5194/essd-12-907-2020.

Rizzo L, Schulte D (2009). *A review of Humpback Whales' migration patterns worldwide and their consequences to gene flow*. *Journal of the Marine Biological Association of the United Kingdom* **89**, 995–1002. https://doi.org/10.1017/S0025315409000332.

Robinson LM, Elith J, Hobday AJ, Pearson RG, Kendall BE, Possingham HP, Richardson AJ (2011). *Pushing the limits in marine species distribution modelling: lessons from the land present challenges and opportunities: marine species distribution models*. *Global Ecology and Biogeography* **20**, 789–802. https://doi.org/10.1111/j.1466-8238.2010.00636.x.

Sathyendranath S, Jackson T, Brockmann C, Brotas V, Calton B, Chuprin A, Clements O, Cipollini P, Danne O, Dingle J, Donlon C, Grant M, Groom S, Krasemann H, Lavender S, Mazeran C, Mélin F, Müller D, Steinmetz F, Valente A, et al. (2021). *ESA Ocean Colour Climate Change Initiative: Version 5.0 Data*. NERC EDS Centre for Environmental Data Analysis. https://doi.org/10.5285/1DBE7A109C0244AAAD713E078FD3059A.

Selig ER, Hole DG, Allison EH, Arkema KK, McKinnon MC, Chu J, Sherbinin A, Fisher B, Glew L, Holland MB, Ingram JC, Rao NS, Russell RB, Srebotnjak T, Teh LC, Troëng S, Turner WR, Zvoleff A (2019). *Mapping global human dependence on marine ecosystems*. *Conservation Letters* **12**, e12617. https://doi.org/10.1111/conl.12617.

Sen Gupta A, Thomsen M, Benthuysen JA, Hobday AJ, Oliver E, Alexander LV, Burrows MT, Donat MG, Feng M, Holbrook NJ, Perkins-Kirkpatrick S, Moore PJ, Rodrigues RR, Scannell HA, Taschetto AS, Ummenhofer CC, Wernberg T, Smale DA (2020). *Drivers and impacts of the most extreme marine heatwave events*. *Scientific Reports* **10** (1, 1), 19359. https://doi.org/10.1038/s41598-020-75445-3.

Smithsonian Ocean Team (2009). *Atlantic Bluefin Tuna (Thunnus Thynnus)*. Smithsonian Institute. https://ocean.si.edu/ocean-life/fish/atlantic-bluefin-tuna-thunnus-thynnus.

Sundararajan M, Taly A, Yan Q (2017). *Axiomatic attribution for deep networks*. https://doi.org/10.48550/arXiv.1703.01365.

Tew Kai E, Rossi V, Sudre J, Weimerskirch H, Lopez C, Hernandez-Garcia E, Marsac F, Garçon V (2009). *Top marine predators track Lagrangian coherent structures*. *Proceedings of the National Academy of Sciences* **106**, 8245–8250. https://doi.org/10.1073/pnas.0811034106.

Viñas J, Gordoa A, Fernández-Cebrián R, Pla C, Vahdet Ü, Araguas RM (2011). *Facts and uncertainties about the genetic population structure of Atlantic Bluefin Tuna (Thunnus thynnus) in the Mediterranean. Implications for fishery management*. *Reviews in Fish Biology and Fisheries* **21**, 527–541. https://doi.org/10.1007/s11160-010-9174-6.

Whittow GC (2020). *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. https://doi.org/10.2173/bow.wetshe.01.

Zeraati R, Engel TA, Levina A (2022). *A flexible Bayesian framework for unbiased estimation of timescales*. *Nature Computational Science* **2**, 193–204. https://doi.org/10.1038/s43588-022-00214-3.