OCEAN
Operator-Centred Enhancement of Awareness in Navigation

D4.2 - Environmental niche models for selected whale species

30 September 2024

Co-funded by
the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101076983. UK participants in Project OCEAN are supported by UKRI grant numbers 10038659 (Lloyds Register) and 10052942 (The Nautical Institute).

Project information

Project acronym:	OCEAN
Full title:	Operator-Centred Enhancement of Awareness in Navigation
Grant agreement:	101076983
Programme and call:	Horizon Europe, CL5-2022-D6-01
Project coordinator:	HVL
Contact:	ocean@hvl.no
Project duration:	36
Project website:	www.ocean-navigation-awareness.eu

Deliverable information

Deliverable number	D4.2
Deliverable title:	Environmental niche models for whale species
Dissemination level:	Public
Deliverable type:	Other
License:	N/A
Status:	Submitted
Due date:	30/09/2024
Submission date:	30/09/2024
Work Package:	04
Lead Beneficiary:	UAC

Disclaimer

Copyright and license information

Revision History

Date	Version	Author	Comment
13/06/2023	V0.01	Rui Prieto, UAC	Creation of document outline
23/06/2023	V0.02	Simon Berrow, IWDG	First draft content of Section 1
31/07/2024	V0.1.1	Rui Prieto, UAC	First draft of report
14/08/2024	V0.1.1	Simon Berrow, IWDG	Peer review
19/08/2024	V0.1.2	Rui Prieto, UAC	Changes according to peer-review. Updated author and acronym list. Updated glossary.
11/09/2024	V0.1.3	Rui Prieto, UAC	Replied to comments by Reviewer 2 and made small changes along the text, and added new Section (7) accordingly to comments/corrections. Added definitions to Glossary.
24/09/2024	V0.1.4	Risk and Quality Manager,, HVL	Final check and approval for submission
26/09/2024	V1.0	Coordinator, HVL	Final check and submission to the granting authority

Author List

Institution	First name and Name	Contact information
UAC	Rui Prieto	rcabprieto@gmail.com
UAC	Sergi Pérez-Jorge	
UAC	Marta Tobeña	
UAC	Mónica Almeida e Silva	

Abbreviations and acronyms

Abbreviation or acronym used in this document	Explanation
4D-SAD	OCEAN 4D Situation Awareness Display
ANN	Artificial Neural Networks (model family)
app	Software application (usually for handheld equipment)
AVHRR	Advanced Very High-Resolution Radiometer
BIOCLIM	Bioclimatic Envelope Model (model family)
BRT	Boosted Regression Trees (model family)

CMEMS	Copernicus Marine Environment Monitoring Service
DOI	Digital Object Identifier
ENHI	European Navigational Hazard Infrastructure
ENM	Environmental Niche Model(s)
EPSG	European Petroleum Survey Group (geodetic parameter dataset)
ERA-INTERIM	European Centre for Medium-range Weather Forecasts global reanalysis
ETOPO-1	NOAA Global Relief Model version 1
EU	European Union
GAM	Generalized Additive Model (model family)
GARP	Genetic algorithm for rule-set prediction (model family)
GIS	Geographic Information System
GLM	Generalized Linear Model (model family)
GLMM	Generalized Linear Mixed Model (model family)
IUCN	International Union for the Conservation of Nature
IWC	International Whaling Commission
km	Kilometre
LTL	Low-trophic level
MARS	Multivariate Adaptive Regression Splines (model family)
MaxEnt	Maximum Entropy (model family)
MTL	Mid-trophic level
NASA	National Aeronautics and Space Administration
NCEI	National Centers for Environmental Information
NGO	Non-Governmental Organization
NOAA	National Oceanic and Atmospheric Administration
ODPS	Ocean Data Processing System
PANGEA	Data Publisher for Earth & Environmental Sciences
PAR	Photosynthetically active radiation
POPA	Programa de Observação das Pescas dos Açores [Azores Fisheries Observer Program]
RF	Random Forest (model family)

SD	Standard Deviation (statistics)
SDM	Species Distribution Model(s)
SEAPODYM	Spatial Ecosystem and POpulation DYnamics Model
SST	Sea Surface Temperature
TLR	Technology Readiness Level
US	United States of America
WP	Work Package

Glossary

Term	Definition used or meaning in the Acronym project	Reference or source for the definition if applicable
Bathypelagic	Marine layer between ~1000 and ~4000 metres depth.	https://handwiki.org/wiki/Earth:Mesopelagic%20zone
BAYES	Any method/algorithm using the Bayes estimator.	https://handwiki.org/wiki/Bayes_estimator
Centroid	The geometric centre of a plane polygon consisting of the arithmetic mean position of all points in the surface.	https://handwiki.org/wiki/Centroid
Ecological niche	Relation between a taxon to specific environmental conditions and other members of its community.	https://handwiki.org/wiki/Biology:Ecological%20niche
Epipelagic	Marine layer between the surface and ~200 metres depth.	https://handwiki.org/wiki/Earth:Mesopelagic%20zone
Heuristic [model]	Algorithms using approximation techniques to find the most approximate solution when the optimal solution is either impractical or impossible to find.	https://handwiki.org/wiki/Algorithm
Macrozooplankton	Free-swimming marine fauna, 2 millimetres to 2 centimetres in body size.	https://handwiki.org/wiki/Biology:Micronekton
Mesopelagic	Marine layer between ~200 and ~1000 metres depth.	https://handwiki.org/wiki/Earth:Mesopelagic%20zone
Micronekton	Free-swimming marine fauna, 2-20 centimetres in body size and distributed from the surface to ~1000 metres depth.	https://handwiki.org/wiki/Biology:Micronekton
Taxon	Any unit used in biological classification (taxonomy). Plural: 'taxa'.	https://www.britannica.com/science/taxon
Vagile	A taxon that is able to move freely.	https://www.merriam-webster.com/dictionary/vagile

Executive Summary

This document reports the work developed under Task 4.2, between months 1 to 24 of the OCEAN project, and relates to Deliverable D4.2 (environmental niche model) that is made public through the GitHub repository in https://github.com/AzWhaleLab/OCEAN.

The description of the work planned for Task 4.2 in the Grant Agreement is reproduced below:

> "Environmental Niche Models (ENMs) will be developed for large whale species using Bayesian and frequentist statistical models. Models will be fit to existing sighting data provided by the UAC and IWDG. A set of explanatory co-variates will be tested, including static and remote-sensed oceanographic co-variates, as well as mid-trophic level prey (micronekton) from the SEAPODYM [Spatial Ecosystem And Population Dynamics Model] [...]. Predictive performance of ENMs will be tested using standard statistical methods. This step will provide the ability of forecasting areas with enhanced conditions for the occurrence of whale aggregations."

Thus, Task 4.2 output should be one or more models enabling the identification of areas with high probability of whale occurrence. Furthermore, output of Task 4.2 should be representative of other similar models, to aid in the design and testing of the European Navigational Hazard Infrastructure developed in Work Package 6.

A set of sub-tasks necessary for the accomplishment of the goals set forth in the Grant Agreement was identified early on in the work-planning phase, consisting of nine steps:

- Identification of case study area(s);
- Identification of case study species;
- Informed decision on case study(ies) to develop;
- Identification of available ecologically relevant predictive co-variates;
- Selection of modelling approach;
- Data compilation, preparation;
- Model fitting;
- Model performance testing;
- Informed selection of modelling approach to be used in Task 4.3.

The identification of a case study area and species benefited from inputs of experts who participated in the *Marine Mammal Ship Strike Mitigation* held under Task 4.1, and followed a systematic approach based on expert opinion. This process resulted in the selection of the sperm whale in the Azores (Portugal) as the case study to be developed in Task 4.2, following a quantitative methodology fully described in this document (Section 2).

Four competing models were considered. The first was a model already published, based on a machine learning approach using a relatively small whale sightings dataset. The other three models were specifically created in the scope of Task 4.2, using a larger and longer time series dataset fitted to Generalized Additive Models. Two of the later models included predictive co-variates from the SEAPODYM mid-trophic level model, as defined in the Grant Agreement.

The two models including SEAPODYM co-variates were rejected due to low performance and were considered inadequate to be used within the OCEAN framework. Discussing the reasons for the low performance is not within the scope of this document but suffice to say that the SEAPODYM co-variates probably do not explain the sperm whale prey distribution in the study region.

The two remaining models were compared after taking into consideration model performance, spatial and temporal resolution, and adequacy for implementation within the OCEAN framework. From this exercise, the remaining model developed within the scope of Task 4.2 (termed ENM 2a) was selected as the best solution to support the work required in Task 4.3.

The model had a moderate predictive performance in line with other similar models for the same species, and has several characteristics that contributed to the choice, including the use of co-variates available from the EU-run Earth Observation Programme, which includes the possibility of supporting daily predictions of whale presence.

However, as with any similar model, the ENM produced in Task 4.2 also has limitations that must be acknowledged, and that can affect its utility in certain situations. In a conservative scenario, where most of the weight is put on ship strike avoidance, the areas identified by the model may become too large to be practical for navigation (for example if they imply route changes in the order of hundreds of kilometres). On the other hand, the predictive performance of models can be improved by increasing the dataset used to fit the models or by advances in modelling techniques. For those reasons, the infrastructure created in the OCEAN project can accommodate the replacement of obsolete models by better performing ones.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Task goals and approach

The overarching goal of Work Package 4 – *Detection and tracking of Marine Mammals in high-density areas* (WP4), is to harness information from a variety of sources to inform mariners about the risk of ship strikes to marine mammals along their routes.

Specifically, environmental niche models (ENM) are included as a powerful way of informing navigation about the danger of collisions with marine mammals, where direct detection is unavailable. Task 4.2 main goal is to demonstrate how ENMs can be leveraged to predict areas of high probability of marine mammal occurrence to inform management and the operation of shipping activity.

Apart from in exceptional cases such as rare species with very restricted distributions, or sessile species, knowing the exact distribution of entire biological populations is seldom possible, especially in the case of highly mobile marine organisms. Instead, researchers and managers often use models to try to infer species distributions and how they may change over time. However, to do so, a profound understating of the drivers of the spatio-temporal distribution of the species of interest is necessary, but which is a challenging endeavour. This is the challenged tackled in Task 4.2.

The declared outcome of Task 4.2, as spelled out in the Grant Agreement, is to create ENMs to enable "[...] forecasting areas with enhanced conditions for the occurrence of whale aggregations.". These forecasts will, in turn, serve to inform mariners, so that they can take proper action, through the European Navigational Hazard Infrastructure (ENHI), developed in WP6, either through the reporting app (OCEAN Task 4.4) or the 4D-SAD (OCEAN WP7). The ENM created in Task 4.2 should be representative of other models that may eventually be integrated within the ENHI developed in WP6, not only for demonstration purposes, but to guarantee proper design and testing of the infrastructure.

However, creating ENMs for all whale species in EU waters would not be feasible. Instead, a strategy to select the area(s), and species, to be used as case studies had to be established, followed by the actual modelling work.

Thus, to realize Task 4.2, several sub-tasks had to be accomplished:

- Identification of case study area(s);
- Identification of case study species;
- Informed decision on case study(ies) to develop;
- Identification of available ecologically relevant predictive co-variates;
- Selection of modelling approach;
- Data compilation, preparation;
- Model fitting;
- Model performance testing;
- Informed selection of modelling approach to be used in Task 4.3.

Based on expert opinion through a systematic approach, fully described in Section 2.2, a case study was chosen for the purposes established in the Grant Agreement for Task 4.2: namely the sperm whale (*Physeter macrocephalus*) in the Azores region (Portugal).

Subsequently, four competing models were considered. The first (ENM 1) was an existing model published in 2016 and based on a machine-learning approach driven by six environmental predictive covariates fitted to a sighting's dataset comprising 381 sperm whale sightings recorded by a fisheries observer program (POPA[1]) between 2004 and 2009. The other three models (ENM 2a, ENM 2b, ENM 2c) were created in the scope of Task 4.2 and consisted of variations in statistical Generalized Additive Models (GAM) driven by different classes of predictive co-variates fitted to 1580 sightings collected by POPA between 2004 and 2016 and transformed to sightings per unit effort.

After the evaluation of model performance, ENM 2c and ENM 2d were discarded due to poor performance and deemed inadequate for the goals of Task 4.2.

Finally, models ENM 1 and ENM 2a were compared taking into consideration model performance, spatial and temporal resolution, and adequacy for implementation in the OCEAN framework. From this exercise, ENM 2a was selected as the best solution to support the ensuing work in Task 4.3.

Finally, it must be stressed that the ENM developed in Task 4.2 has limitations, as is always the case with this approach. No model should be ever considered definitive, and their limitations must be acknowledged so that they can be used according to the goals set for management, conservation, or operation.

For the goals of the OCEAN project, and within its timeframe, ENM 2a fulfils the objectives set in the Grant Agreement and will be carried on supporting other tasks within the project. However, the framework that has been established in OCEAN is purposely flexible enough to allow the replacement of ENMs deemed obsolete at any time. This attribute of the OCEAN framework makes it robust and allows for improved performance overtime, as more data and new modelling techniques become available.

1.2 Intended readership

This report details the objectives, methodology and results of Task 4.2. The methodologies developed and applied during the development of Task 4.2 may be of interest to different stakeholders:

- European Commission;
- OCEAN partners;
- Science practitioners;
- Decision/policymakers;
- Shipping industry;
- Environmental NGOs.

1.3 Structure of the document

This document is structured in several sections to facilitate consultation and readability. The strategy and methods used for selection of a case study and for development of ENMs are detailed in Sections 2 and 3. Section 2 also includes a very succinct introduction to the theory and some methods used to fit ENMs. In no way it intends to be an in-depth review on ecological modelling but act only to give some context to readers not familiar with these methods.

In Section 5 the four competing models considered in Task 4.2 are compared on different aspects (predictive power, spatio-temporal resolution, adequacy for implementation in the OCEAN framework). Subsequently, the reasoning for the selection of the model to be used in Task 4.3 is explained.

[1] https://www.popaobserver.org

Section 6 summarizes the work done in Task 4.2 and highlights the limitations of the resulting ENM, and of any ENM in general, and good practices involved in utilizing this strategy to inform decision making.

1.4 Relationship with other deliverables

This document will support deliverables 4.3, 4.9, 6.4.

2 Environmental Niche Models

2.1 Succinct introduction to Environmental Niche Models

Natural ecosystems possess intrinsic economic and social values but are increasingly influenced by anthropogenic activities, making reliable information about the spatio-temporal distribution of organisms essential for creating robust management frameworks (Costanza et al., 1997; Grumbine, 1994; Guisan et al., 2013; World Bank, 2006). Notwithstanding, for most animal species, especially aquatic ones, information on their distribution is scarce, fragmented, and typically extremely costly to obtain.

To overcome data scarcity for vast, often unsurveyed, areas researchers and managers often resource to ecological models relating biological records to relevant independent environmental variables (Austin, 2007; Elith & Leathwick, 2009; Guisan & Zimmermann, 2000). Given the challenges of surveying the entire oceans, this approach has been increasingly used for inferring the potential distribution of marine organisms (Melo-Merino, Reyes-Bonilla, & Lira-Noriega, 2020).

The terms *Species Distribution Models* (SDMs) and *Ecological Niche Models* (ENMs) are often used interchangeably in the literature although they translate slightly different concepts (Elith & Leathwick, 2009; Soberón, Osorio-Olvera, & Peterson, 2017). SDMs are tailored at estimating the real distribution and densities of a taxonomic group (usually a species) while ENMs are used mostly to infer the potential distribution of a taxon by characterization of the organism's ecological niche (Peterson et al., 2011; Soberón et al., 2017). Thus, SDMs are constrained by the survey area and period, but translate real distribution, while ENMs are less constrained, especially by survey area, but only translate a potential distribution, which may or not be realized.

Even with optimal conditions, organisms can be absent from an area for different reasons, such as not having physical access to that area, or having a population too small to be able to occupy all the available habitat. Thus, the potential distribution of a species is translated by areas with good conditions for its occurrence, regardless of the actual presence of the species there. Only by surveying these areas we can be sure if the species is present in those areas, but until having these data, for conservation purposes, the species should be considered as potentially being present.

For clarity, here we refer to ENMs since the goal is to predict where animals may be present at any given moment, regardless of the modelling approach used, within and beyond the area surveyed.

Choosing the appropriate modelling approach to infer a species distribution is often a non-trivial matter, as there are several methods with differing levels of complexity, requirements and assumptions (Austin, 2007; Elith et al., 2006; Elith & Leathwick, 2009).

The most widely used approaches for that purpose fall into five broad categories (Table I):

1. Heuristic models (also known as envelope models), that characterise sites that are located within the climactic niche occupied by a species regardless of other biophysical and biotic variables, of which BIOCLIM is the most eminent;

2. Statistical, including generalised linear models (GLM), generalised linear mixed models (GLMM) and generalised additive models (GAM), multivariate adaptive regression splines (MARS), among others, that are extensively used for the strong statistical foundation and proven ability to realistically model ecological relationships;
3. Bayesian approach, based on inductive learning process carried out using Bayes theorem, such as the BAYES algorithm;
4. Machine-learning techniques stemming from informatics science and including artificial neural networks (ANN), boosted regression trees (BRT), random forests (RF), and maximum entropy (MaxEnt);
5. Evolutionary computation, such as the genetic algorithm for rule-set prediction (GARP) modelling environment that uses machine-learning approaches but is particular for using a genetic algorithm.

Table I: Main modelling approaches categories and examples. See text for explanation of acronyms.

Model family	Examples	References
Heuristic	BIOCLIM*	(Beaumont, Hughes, & Poulsen, 2005; Nix, 1986)
Statistical	GAM; GLM; GLMM; MARS	(Guisan, Edwards Jr, & Hastie, 2002; Guisan & Zimmermann, 2000; Hastie & Tibshirani, 1986; Nelder & Wedderburn, 1972)
Bayesian	BAYES	(Aspinall, 1992)
Machine-learning	RF; BRT; ANN; MAXENT*	(Breiman, 2001a; Elith et al., 2006; Elith, Leathwick, & Hastie, 2008; Elith et al., 2011; Lawler, White, Neilson, & Blaustein, 2006; Olden, 2008; Phillips, Anderson, & Schapire, 2006)
Evolutionary computation	GARP*	(D. Stockwell & Peters, 1999; D. R. B. Stockwell & Noble, 1992)

*Methods specifically designed to be used with presence-only data.

The multiplicity of approaches to try to identify species distributions (either realized or potential) gave rise to much discussion and confusion about each method's strengths and weaknesses, and their utility (e.g. (Elith et al., 2006; Fitzpatrick, Gotelli, & Ellison, 2013; Monk, Ierodiaconou, Harvey, Rattray, & Versace, 2012; Pueyo, 2012; Sinclair, White, & Newell, 2010; Uriarte & Yackulic, 2009; Yackulic et al., 2012)). However, a robust body of review and synthesis of past work offers guidance and best practice in choosing the best approach for different questions and data types (Elith & Leathwick, 2009; Guisan & Thuiller, 2005; Guisan & Zimmermann, 2000; Richards, Carstens, & Lacey Knowles, 2007; Schröder, 2008; Soberón, 2007; Soberón et al., 2017).

Fundamental differences among the aforementioned methods relate to the goals and data available for the studies (Breiman, 2001b; Elith & Leathwick, 2009; Mac Nally, 2000). Often the main goal of these models is to gain insight into the ecological drivers of species distributions. On the other hand, when the focus is on conservation and space management, understanding the ecological processes involved may not be strictly necessary as long as predictive performance is good.

Methods in the "Statistical" family in Table I have been long favoured for the former case, but they tend to be difficult to implement due to very strict assumptions. "Machine-learning" and "Bayesian" methods, on the other hand, can produce accurate predictions of distributions even with scarce or non-systematic data, (Breiman, 2001b; Elith et al., 2006; Elith & Leathwick, 2009). Methods in these two families (Machine-learning and Bayesian) are sometimes perceived as 'black boxes' (*sensu* Fraser (1968)), in part because their application is more recent as a result of being computationally intensive. Nevertheless, due to the good results they bring and their flexibility, these methods are increasingly utilized to predict the potential distribution of species (Breiman, 2001b; Clark & Gelfand, 2006; Elith & Leathwick, 2009; Olden, 2008; Uriarte & Yackulic, 2009).

The nature of the dependent variable (the presence data) often dictates the type of approach that can be used. Traditionally, the investigation of species distribution was only performed using presence-absence data, consisting of known occurrences (presence) and known absences (absence), collected under specifically designed surveys. The use of presence-absence data enables investigating survey bias and computing the detection probability of vagile species, also termed the species prevalence (Phillips et al., 2009).

Unfortunately, for many species and areas, such as vast expanses of the oceans, presence-absence data is either too expensive or even impossible to obtain with current methods. In those cases, the only available records may consist of presence-only data (museum/historical records, non-systematic surveys, platforms of opportunity, telemetry data, citizen science, etc.). Machine-learning methods are often utilized with presence-only data, or data collected non-systematically, but call for a good understanding of their limitations.

Regardless of the modelling approach, when used to inform decision-making it is essential that the methods, data sources, and model limitations are fully disclosed and plainly presented to the users and be subject to subsequent field validation and improvement in order to increase efficiency of management decisions (Sofaer et al., 2019).

2.2 Development of Species Distribution Models for the OCEAN project

2.2.1 Species and area selection

Work in Task 4.2 started by devising a strategy to select the focal marine mammal species and area. The decision on the species and area to be used as a case study were intentionally not taken *a priori* in the Grant Agreement, in order obtain feedback from the specialists invited for the *Marine Mammal Ship Strike Mitigation* held under Task 4.1 (Prieto, 2023).

The most suitable demonstration location in the scope of the project fell naturally on the Azores (Portugal), both due to the diversity of cetacean species present, as well as the data available and expertise for ENM creation by OCEAN partner UAC.

In order to decide which should be the focal marine mammal species in Task 4.2, we had informal discussions with specialists present at the workshop mentioned above, and within the OCEAN consortium. From those discussions it was decided to focus only on large whales (>8 metres mean adult size), as these are perceived as the most problematic group regarding ship strikes, not only for the harm caused to the animals but also regarding the damage that can occur to vessels involved in the incidents (F. Ritter, 2012; Schoeman, Patterson-Abrolat, & Plön, 2020; Winkler, Panigada, Murphy, & Ritter, 2020).

Subsequently, the list of large whales present in the North Atlantic was compiled and a set of selection criteria was established:

1. IWC database incident reporting frequency (%): relative number ship strikes by species (including definitive, probable and possible cases) as % of the total number of records in the IWC ship-strikes database (period 1820-2019; total number of records: 933), obtained from (Winkler et al., 2020) [only informative; not used in final score calculation].
2. Species IUCN Red List global conservation status (IUCN, 2023). Regional level assessment was not considered because it was not available for all species. Category definitions can be obtained from IUCN[2] [only informative; not used in final score calculation].

[2] https://www.iucnredlist.org

3. Species of concern: score from 1 (low) to 3 (high) based on specialist opinion of the threat ship strikes pose to animals but also ships, taking into account frequency of recorded cases, species mean adult size, species conservation status.
4. Data availability: score from 1 (no data available) to 3 (enough data available) about data availability to fit environmental niche models.
5. Compatibility with other systems being tested in WP4: sores from 1 (not compatible) to 3 (compatible if animals present) with other systems being developed in WP4, namely passive acoustic systems, and ship-borne visual detections.
6. Adequacy regarding the project's timeframe: scores from 1 (not possible within project's timeframe) to 3 (possible within project's timeframe), taking into account data availability and amount of effort for data preparation and model fitting (data compilation, cleaning, processing, etc.).

Lastly, a decision matrix with ten large whale species present in the North Atlantic was created and a body of specialists from UAC gave scores to criteria 3-6 (Table I).

Based on those criteria, the sperm whale (*Physeter macrocephalus*) scored high in all criteria and was chosen as the ideal demonstration species. The fin whale (*Balaenoptera physalus*) was the second highest-scoring species (total score of 10) but was ultimately discarded because the data preparation would be even more time-consuming than for the sperm whale. Attempting to create models for these two species simultaneously could jeopardize the project's schedule and affect other work packages, specifically WP6 that depends on results from WP4.

Table II: Decision matrix for candidate large whale species for creation of ENM under Task 4.2. Ship strike reporting frequency from (Winkler et al., 2020); IUCN Red List global conservation status category from (IUCN, 2023).

	IWC database ship strike reporting frequency (% of records)	IUCN global category	Species of concern	Data availability	Compatible with other OCEAN systems	Realistic in project timeframe	Score	Comments
Bowhead whale (*Balaena mysticetus*)	<1	Least concern	2	1	2	1	**6**	Not present in the Azores; access to data from other regions problematic
Right whale (*Eubalaena glacialis*)	4,7	Critically endangered	3	1	2	1	**7**	Not present in the Azores; access to data from other regions problematic
Blue whale (*Balaenoptera musculus*)	1,7	Endangered	3	2	2	1	**8**	Limited data; only acoustically detectable during short periods of the year; highly seasonal
Fin whale (*Balaenoptera physalus*)	20,3	Vulnerable	3	3	2	2	**10**	Enough data; only acoustically detectable during short periods of the year; highly seasonal
Sei whale (*Balaenoptera borealis*)	1,5	Endangered	2	2	2	2	**8**	Enough data; only acoustically detectable during short periods of the year; highly seasonal
Bryde's whale (*Balenoptera edeni*)	3,5	Least concern	2	1	2	1	**6**	Not enough data; only acoustically detectable during short periods of the year; highly seasonal and rare visitor
Minke whale (*Balaenoptera acutorostrata*)	3,0	Least concern	2	1	2	1	**6**	Very limited data; only acoustically detectable during short periods of the year; highly seasonal
Humpback whale (*Megaptera novaeangliae*)	17,4	Least concern	3	1	2	1	**7**	Very limited data; only acoustically detectable during short periods of the year; highly seasonal
Sperm whale (*Physeter macrocephalus*)	10,9	Vulnerable	3	3	3	3	**12**	Enough data; acoustically detectable when present; present throughout the year
Bottlenose whale (*Hyperoodon amplullatus*)	0	Near threatened	1	1	3	1	**6**	Not enough data; acoustically detectable when present; highly seasonal

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101076983. UK participants in Project OCEAN are supported by UKRI grant numbers 10038659 (Lloyds Register) and 10052942 (The Nautical Institute).

2.2.2 Modelling approach

The sperm whale in the Azores (Portugal) was selected as the case study for the development of an ENM, as detailed in Section 2.2.1. The sperm whale is associated with deep waters in all oceans and the Mediterranean Sea. The species is widely distributed from the tropics to the ice edge in both hemispheres, with males and females prevalent in distinct parts of their range (Whitehead, 2003). The sperm whale is still recovering from population reduction caused by whaling and is subject to several threats, including ship strikes, and is classified as a vulnerable species by the International Union for Conservation of Nature (Grossi et al., 2021; Taylor et al., 2019; Winkler et al., 2020).

The work developed in Task 4.2 builds on the works by Mónica A. Silva et al. (2014) and Tobeña, Prieto, Machete, and Silva (2016b) that represent a baseline for the spatial and temporal distribution of the sperm whale in the Azores and for the identification of the main predictive co-variates. However, these works were considered preliminary (Tobeña et al., 2016b), calling for a larger dataset both regarding whales and explanatory co-variates, as well as alternative modelling approaches.

Tobeña et al. (2016b), co-lead by two of the authors of this report, fitted models for 16 cetacean taxa (15 species and 1 genus), using a subset of a sightings dataset obtained from a fisheries observer program (POPA[3]). Up to the work developed in Task 4.2, the results published by Tobeña et al. (2016b) for the sperm whale established the state of the art regarding ENMs for this species in the Azores region.

The approach taken by Tobeña et al. (2016b) relied on a modelling approach in the machine learning family (Maxent: Phillips et al. (2006); Table I). In that work, the models did not take into account sighting effort, nor did they have access to true species absences, using instead a statistical methodology to create what is known in the literature as pseudo-absences (Phillips et al., 2009).

In Task 4.2, we considered the sperm whale ENM published by Tobeña, Prieto, Machete, and Silva (2016a) as our baseline model (ENM 1), further detailed in Section 3, below. Furthermore, following the recommendations by the authors (Tobeña et al., 2016a), we then set to fit new ENMs with more sighting and absence data, alternative explanatory co-variates, and an alternative modelling method.

That work resulted in the development of ENM2, further detailed in Section 4, below, and required several steps:

- Increasing the sighting dataset using the same data provider (POPA);
- Calculating surveying effort;
- Obtaining true animal absences;
- Testing a new modelling approach: Generalized Additive Models (GAM);
- Testing new predictive co-variates from the Copernicus Marine Service[4], including modelled prey fields;
- Evaluating competing modelling approaches;
- Evaluating suitability of models approaches in the scope of issuing regular navigational warnings (NW).

---

[3] https://www.popaobserver.org

[4] https://marine.copernicus.eu

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101076983. UK participants in Project OCEAN are supported by UKRI grant numbers 10038659 (Lloyds Register) and 10052942 (The Nautical Institute).

3 ENM 1

As already mentioned in Section 2.2, above, ENM 1 was published in Tobeña et al. (2016b). The Open Access publication and supplementary materials detailing the methods and results can be accessed using the publication DOI 10.3389/fmars.2016.00202. Maps of the monthly habitat suitability predictions are also publicly available in PANGEA, through DOI 10.1594/PANGAEA.864511.

Here a succinct description of the model is given for context and to simplify comparison with ENM 2 (Section 4, below). Sections of text extracted literally from the publication are presented between quotation marks (“”) or, in the case of Tables and Figures, mentioned in the captions, followed by the reference. Within quotations, square brackets signal either comments [*comment*] or deletions [...].

3.1 ENM 1 Methods

3.1.1 Sighting data

“Cetacean occurrences were obtained from the Azores Fisheries Observer Programme (POPA), from May to November, between 2004 and 2009 [...]. POPA places trained observers aboard tuna-fishing vessels to monitor and collect information on the fishery and on the presence and behavior of cetaceans, seabirds and turtles. Cetacean surveying effort is conducted when the vessel is cruising or searching for fish schools. During on-effort periods, vessel position and environmental conditions are recorded every 30 minutes or whenever vessel course changes >20°. All sightings and vessel positions are georeferenced using global positioning system with datum São Braz (EPSG 2190). Sightings are coded according to reliability of species identification, from 0 (low confidence) to 3 (definitive). In this study we analyzed only sightings recorded during on-effort survey periods conducted in sea states on the Beaufort scale ≤3 and with an identification score of 3. Each sighting was considered as a single occurrence, irrespective of the number of individuals within the group.” (Tobeña et al., 2016b).

For the sperm whale, 381 sightings were available (Figure 1). These sightings were split in two subsets: 338 were used to fit the models; the remaining 43 sightings (11%) were selected randomly and set apart for model performance testing.

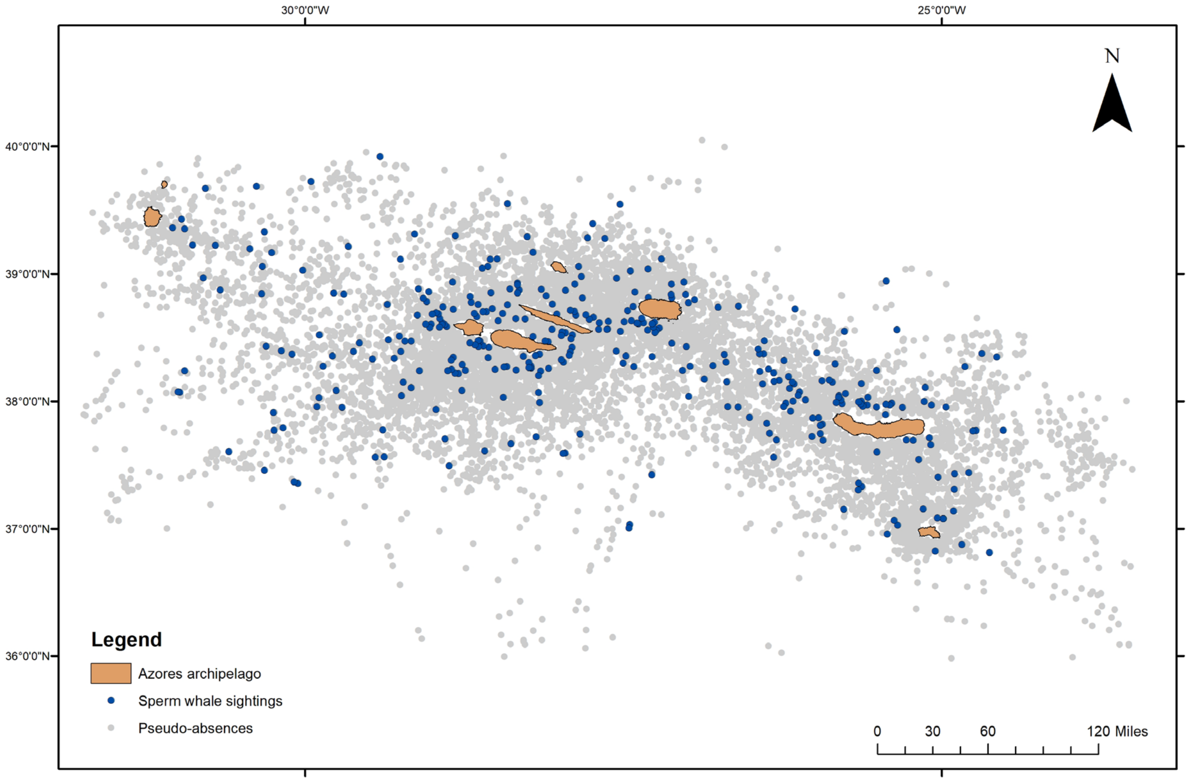

Figure 1: Sperm whale sightings recorded by POPA observers between 2004-2009 (blue) and pseudo-absences (grey) used to fit the Maxent Environmental Niche Model (ENM 1).

3.1.2 Co-variates

"A set of 18 candidate environmental variables […] were selected based on their perceived ecological relevance for cetaceans" (Tobeña et al., 2016b) (and references therein).

The co-variates used to fit the models by Tobeña et al. (2016b) are presented in Table III. Four main remote-sensed and topographic variables were considered, and the remaining were derived from those using statistical tools in the Geographic Information System (GIS) ArcGIS. The full reasoning for the choice and processing of co-variates variables is given in Tobeña et al. (2016b) main text and supplementary materials.

Table III: Candidate environmental variables used in the variable selection procedure prior to model fitting. Reproduced from (Tobeña et al., 2016b).

Environmental variable	**Acronym**	**Transformation**	**Resolution Spatial/temporal**	**Units**	**Source**
Depth	Depth	none	1 arc-minute/static	m	NationalGeophysicalDataCenter(NGDC), National OceanicandAtmospheric Administration (NOAA) http://www.ngdc.noaa.gov/mgg/global/global.html. (Amante & Eakins, 2008)
Night-time sea surface temperature	NSST	none	2.5 arc-minute/month	°C	National Aeronautics and Space Administration (NASA) Goddard Space Flight Center's Ocean Data Processing System (ODPS) http://oceancolor.gsfc.nasa.gov. (Campbell, Blaisdell, & Darzi, 1995).
Chlorophyll-a concentration	Chl-a	log10	2.5 arc-minute/month	mg/m^3	National Aeronautics and Space Administration (NASA) Goddard Space Flight Center's Ocean Data Processing System (ODPS) http://oceancolor.gsfc.nasa.gov. (Campbell et al., 1995).
Seamounts	*None*	none	10 meters	unitless	www.int-res.com/articles/suppl/m357p017_app.pdf. (Morato et al., 2008).
Derived environmental variables					*Original variable*
Slope within a 3x3 pixel kernel	Slope	log10	1 arc-minute/static	degrees from the horizontal	Depth
Euclidean distance to shoreline	Distance to shore	square root	1 arc-minute/static	m	Depth
Euclidean distance to 200 meters isobath	Dist(200)	square root	1 arc-minute/static	m	Depth
Euclidean distance to 500 meters isobath	Dist(500)	square root	1 arc-minute/static	m	Depth
Euclidean distance to 1000 meters isobath	Dist(1000)	square root	1 arc-minute/static	m	Depth
Euclidean distance to 2000 meters isobath	Dist(2000)	square root	1 arc-minute/static	m	Depth
Seamount density within 8x8 pixel kernel	d-Seamounts	none	1 arc-minute/static	seamounts /km^2	Seamounts
Environmental variable	**Acronym**	**Transformation**	**Resolution Spatial/temporal**	**Units**	*Original variable*

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101076983. UK participants in Project OCEAN are supported by UKRI grant numbers 10038659 (Lloyds Register) and 10052942 (The Nautical Institute).

Minimum depth seamounts	Seamount_dpt	none	1 arc-minute/static	m	Seamounts
Time-lagged Chlorophyll-a concentration (-1 month)	Chl-a (-1m)	log10	2.5 arc-minute/month	mg/m^3	Chl-a
Time-lagged Chlorophyll-a concentration (-2 months)	Chl-a (-2m)	log10	2.5 arc-minute/month	mg/m^3	Chl-a
Chlorophyll-a local variation (calculated as standard deviation within a 8x8 pixel kernel of log-transformed Chlorophyll-a)	V-Chl-a	none	2.5 arc-minute/month	SD log10(mg/m^3)	Chl-a
Time-lagged Chlorophyll-a local variation (-1 month)	V-Chl-a (-1m)	none	2.5 arc-minute/month	SD log10(mg/m^3)	Chl-a
Time-lagged Chlorophyll-a local variation (-2 months)	V-Chl-a (-2m)	none	2.5 arc-minute/month	SD log10(mg/m^3)	Chl-a
Nigh-time sea surface temperature local variation (calculated as standard deviation within a 3x3 pixel kernel of NSST)	V-NSST	none	2.5 arc-minute/month	SD °C	NSST

3.1.3 Fitting and model selection

The "[...] dataset comprised only presence records thus [...] the software MaxEnt 3.3.3k [...]" was used to fit the models (Dudík, Phillips, & Schapire, 2007; Phillips et al., 2006; Tobeña et al., 2016b). The software MaxEnt 3.3.3k mentioned above, uses the machine learning MAXENT method (Table I) and was developed to infer the potential distribution of organisms using presence-only data, as a function of a set of co-variates that are known or considered to be ecologically relevant to the organism in question (Dudík et al., 2007; Phillips et al., 2006).

Further details about model tuning, selection and performance evaluation are given in Tobeña, et al. [52] main text and supplementary materials. However, it is noteworthy that the authors were careful to counteract effects of sample selection bias, which is a common problem with presence only records.

"MaxEnt predictions are strongly affected by sample selection bias (Phillips et al., 2009); models suffering from that type of bias can be considerably improved by drawing the environmental samples from a distribution of locations with the same selection bias as the occurrence data to create an 'informed' model (Kramer-Schadt et al., 2013; Phillips et al., 2009). POPA survey effort is dependent on fish distribution and fishing strategies of the boat captains and is neither random, nor homogeneously distributed (M. A. Silva, Feio, & Prieto, 2002; Mónica A. Silva et al., 2011)." (Tobeña et al., 2016b). For that reason, the authors "[...] dealt with sample selection bias in the POPA dataset by drawing environmental samples from a set of 10,000 randomly chosen vessel data points, thus creating informed models [...]" (Tobeña et al., 2016b) (Figure 1).

Covariates were evaluated for co-linearity and ecological relevance to strip the models from uninformative or redundant variables. This pruning process is detailed in Tobeña, et al. [52] main text and supplementary materials.

Another aspect that was addressed when fitting the models was model overfitting. As stated by the authors, "[o]verfitting will occur when the model fits the training data too closely, hampering the generalization ability of the model […]; when relying on MaxEnt default settings, models can easily become overfit as more functions are added to the model […]". That problem was addressed by running several test trials and inspecting the resulting graphical representation of the response curves relating sightings to environmental co-variates, looking for signs of "[…] biologically nonsensical responses (ie, highly jagged or multimodal).". Following that procedure, "[m]odels were chosen based on the best combination of their ecological plausibility and predictive power based on test […] scores".

The authors also applied a distance buffer of 150 nautical miles from shore to the model predictions to avoid predicting outside the environmental conditions used when fitting the model (Tobeña et al., 2016b), which is patented in the prediction maps (Figure 2).

3.2 ENM 1 Results

After variable pruning and model tuning, the best model for the sperm whale created by Tobeña et al. (2016b) retained six explanatory co-variates: depth; sea surface temperature (night time); primary productivity in the previous month (translated by chlorophyll-a concentration); local variation of productivity in the previous second month; distance to shore; and seamount density (seamounts/km^2).

Overall, the discrimination power of the model was considered moderate based on the performance tests that were performed and are fully presented in Tobeña et al. (2016b).

Figure 2 gives examples of predictions based on monthly averages of the predictive co-variates used in ENM 1, just to serve as an example of the model output. Predictions can be created for shorter periods of time (daily, weekly) in the same manner. In the monthly averages, some patters do emerge, with better habitat quality especially near the islands in the eastern group of islands and to the banks southwest of the central group of islands, as well as near some

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101076983. UK participants in Project OCEAN are supported by UKRI grant numbers 10038659 (Lloyds Register) and 10052942 (The Nautical Institute).

seamounts. The seasonal changes in habitat quality are also perceived by a degradation of the conditions towards the end of the summer.

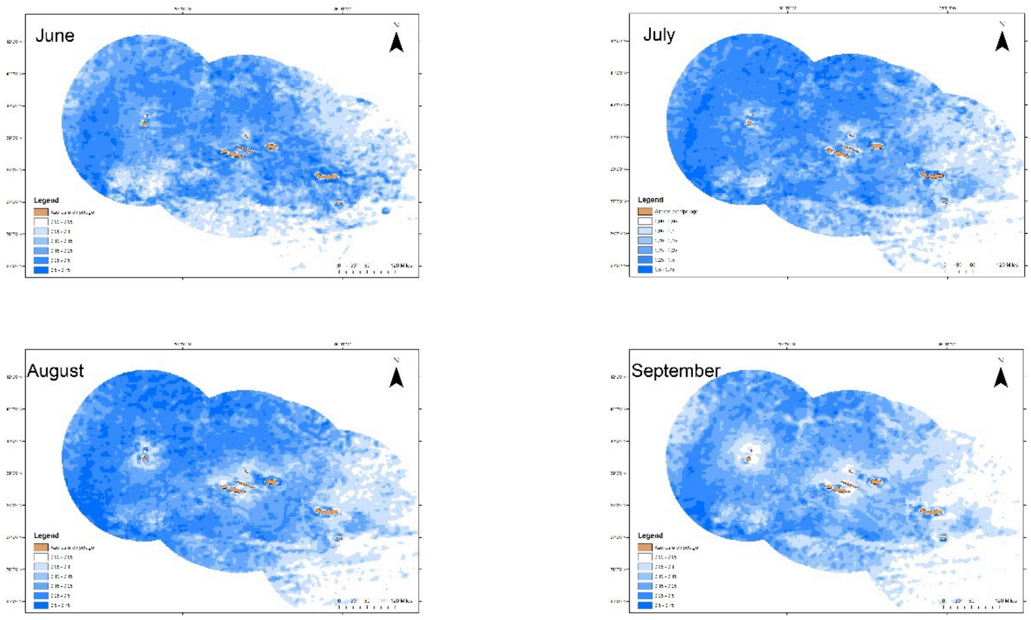

Figure 2: Predicted monthly sperm whale habitat quality from June to September based on ENM 1. Values represent a unitless habitat quality score with higher values representing better habitat quality for the species.

3.3 ENM 1 Conclusions

The model for the sperm whale created by Tobeña et al. (2016b) was considered at the time of publication to have moderate predictive power, which is usual in this type of model (Table VI). As it stands, the model is at TLR 7 and can be used for the goals established in the OCEAN Grant Agreement, namely supporting the subsequent work predicted in Task 4.3.

However, as the authors themselves cautioned, the small sample size used to fit the models, can be an issue that could lead to inconsistent results (Tobeña et al., 2016b). Repeatability and consistency in model results are key factors in informing decision making, and efforts should be made to ensure that (Sofaer et al., 2019).

It is still to be tested if the sperm whale model by Tobeña et al. (2016b) suffers from any inconsistencies from small sample size, but the effort required for that testing would be similar to the effort required to create a new model with a larger sample size and wider temporal span.

Given that since the publication of the sperm whale model by Tobeña et al. (2016b) new data is now available, the most prudent line of work was to use a more robust dataset to test competing models and come to a final decision on the best model to support Task 4.3. That work is detailed in Section 4.

4 ENM 2

Unlike ENM 1 (Section 3, above), which is already published (Tobeña et al., 2016b), the work for ENM 2 is novel and, as such, here we present a detailed description of the methods and results.

For this work, we decided to fit Generalized Additive Models (GAM) to a larger (and longer) sightings dataset than that used in Tobeña et al. (2016b). The choice of modelling method fell over GAM because for this work survey effort was calculated and true absences were obtained. In that situation, a statistical method was deemed preferable over a machine-learning method (Table I).

Three main models were tested, ENM 2a, ENM 2b, and ENM 2c, varying on the type of predictor co-variates utilized to fit the models.

4.1 ENM 2 Methods

4.1.1 Sighting data

Sperm whale sightings and effort data was collected by POPA between 2001 and 2016 (Mónica A. Silva, Feio, Prieto, Goncalves, & Santos, 2002; Mónica A. Silva et al., 2014) (Figure 3), using the same methodology described by Tobeña et al. (2016b) and references therein (see Section 3, above).

The study area was defined between 35°- 41° N and 22°- 33° W. Effort outside this box was not considered for the analysis (Figure 3). To minimize the impact of unbalanced survey effort, the track lines of the vessels were split into segments of 10 km. The final effort data included 13795 segments with a mean length of 9.95 km (SD= 1.29 km) and ranging from 8 to 15 km. Tracks shorter than 8 km were excluded from the analysis to reduce the effect of zero inflated count data (Welsh, Cunningham, Donnelly, & Lindenmayer, 1996).

The number of sperm whale sightings were computed for each on-effort segment, yielding a measure of sightings per unit effort (whales/km). To avoid duplication of the same sightings recorded by different tuna-fishing vessels within the same day, contiguous segments at <5 km were selected and the segments with the larger number of sperm whale sightings were chosen. The POPA program runs from April to October, but due to the low number of sperm whale sightings during April and October, these two months were excluded from the analysis. After this process, in total 1580 sightings were available for the model fitting process (Figure 3).

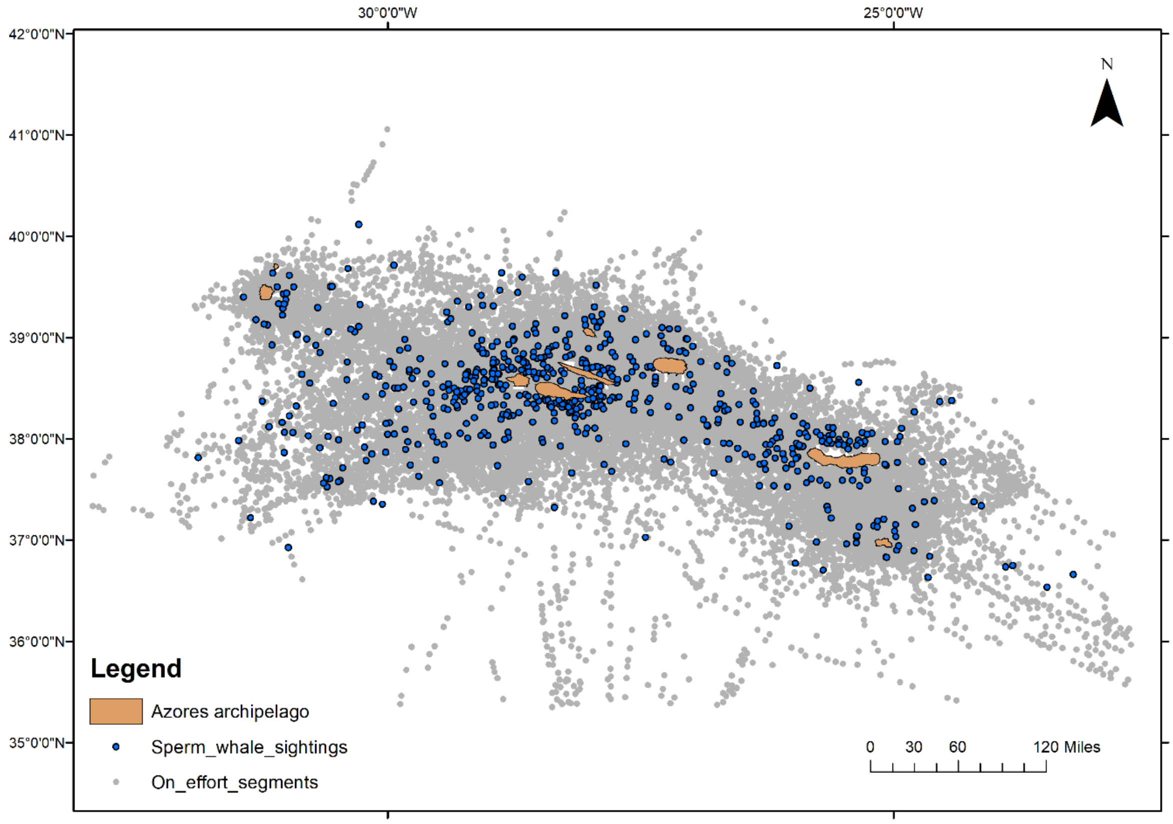

Figure 3: Sperm whale sightings recorded by POPA observers between 2001-2016 (blue) and on effort absences (grey) used to fit the GLM Environmental Niche Models (ENM 2a, b, c).

4.1.2 Co-variates

A total of 19 candidate prey and environmental variables were selected for the GAM models based on the species' ecological preferences and regional oceanography and physiography characteristics (Lehodey, Murtugudde, & Senina, 2010; Mannocci et al., 2014) (Table IV).

Potential prey biomass distributions were obtained from the mid-trophic module of the Spatial Ecosystem And POpulation DYnamics Model (SEAPODYM-MTL, (Lehodey et al., 2014; Lehodey et al., 2010); Table IV). This model simulates the spatio-temporal dynamics of microkenton biomass density biomass between the surface and 1,000 m depth. Micronekton is defined by a size ranging from 2 to 20 cm, including a large diversity of crustaceans, fish, squid and gelatinous species. Six functional groups of micronekton are simulated based on the dynamics of their vertical distribution: epipelagic, migrant upper mesopelagic, upper mesopelagic, highly migrant lower mesopelagic, migrant lower mesopelagic and lower mesopelagic (Table IV).

The SEAPODYM-MTL model is driven by ocean currents, temperature and net primary productivity, using a robust system of advection-diffusion-reaction equations (Lehodey et al., 2010). Ocean currents and temperature are obtained from the Copernicus Marine Service multiyear product. Net primary production is computed from satellite-derived chlorophyll a, sea surface temperature and photosynthetically active radiation (PAR) observations (obtained from CMEMS multiyear product for chlorophyll, NOAA NCEI AVHRR-only Reynolds for sea surface temperature, and ERA-INTERIM for PAR). The analysis also incorporated the outputs of the low-trophic module of the SEAPODYM (SEAPODYM-LTL). The SEAPODYM-LTL follows the modelling framework developed for the mid-trophic module, with only a single functional group for all mesozooplankton organisms. The biomass of mesozooplankton is integrated over the whole epipelagic layer defined by the euphotic depth (Conchon, 2016)

(Table IV). The low and mid-trophic modules of the SEAPODYM model were validated with biomass estimates from zooplankton and micronekton sampling cruises and acoustic backscatter data (Lehodey et al., 2014; Lehodey et al., 2010).

In addition to these prey-related variables, we included surface chlorophyll-a concentration (Chl-a) and sea surface temperature (SST). A time-lag of 2 months was also considered for Chlorophyll-a as it was found to be an important predictor of the spatio-temporal distribution of sperm whales around the study area (Tobeña et al., 2016b).

Depth was obtained from the ETOPO-1 digital elevation model (Amante & Eakins, 2008). Derived static variables were slope, calculated using the elevation values of the four immediately adjacent depths (P. Ritter, 1987), Euclidean distance to coast, seamounts (Morato et al., 2008) and to the isobaths 200, 500, 1000, 3000 and 4000 metres (Table IV).

For each on-effort segment, the environmental and prey-related variables were extracted in the segment centroid position, from the original co-variates netcdf files using the packages “raster” (Hijmans & van Ettern, 2012), “ncdf4” (Pierce, 2019) and “rgdal” (Bivand, Keitt, & Rowlingson, 2019) from the R software (version 2.15.3 (R Development Core Team, 2015)).

On-effort segments with incomplete data were excluded from the final dataset. Prior to running the GAM models, the Pearson's correlation coefficient was used to assess the collinearity between pairs of predictors. Less correlated variables (Pearson coefficient <0.7) and more ecologically relevant were kept on the dataset (Dormann et al., 2013). This led to the removal the variables: distance to coast, distance to 200 m and distance to 3000 m.

Table IV. Candidate variables used to fit the GAM models for sperm whale in the Azores.

Environmental variable	**Acronym**	**Transformation**	**Spatio-temporal resolution**	**Units**	**Source**
Depth	depth	none	1 arc-minute/static	m	ETOPO-1 digital elevation model (Amante & Eakins, 2008)
Seamounts	*None*	none	10 meters/static	unitless	Morato et al. (2008)
Euclidean distance to seamount	d-Seamounts	none	1 arc-minute/static	m	Seamounts
Sea surface temperature	SST		0.083 degrees/daily	Degrees Celsius	CMEMS Global Ocean Physics Reanalysis (1993-2021)
Mass concentration of chlorophyll a in sea water	chl		4 km /daily	mg/m^3	CMEMS Global Ocean Colour
Derived environmental variables					*Original variable*
Slope within a 3x3 pixel kernel	Slope	log10	1 arc-minute/static	degrees from the horizontal	depth
Euclidean distance to coast	Distance to coast	square root	1 arc-minute/static	m	depth
Euclidean distance to 200 meters isobath	Dist(200)	square root	1 arc-minute/static	m	depth
Euclidean distance to 500 meters isobath	Dist(500)	square root	1 arc-minute/static	m	depth
Euclidean distance to 1000 meters isobath	Dist(1000)	square root	1 arc-minute/static	m	depth
Euclidean distance to 3000 meters isobath	Dist(3000)	square root	1 arc-minute/static	m	depth

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101076983. UK participants in Project OCEAN are supported by UKRI grant numbers 10038659 (Lloyds Register) and 10052942 (The Nautical Institute).

Environmental variable	Acronym	Transformation	Spatio-temporal resolution	Units	Source
Euclidean distance to 4000 meters isobath	Dist(4000)	square root	1 arc-minute/static	m	depth
Euclidean distance to seamount	d-Seamounts	none	1 arc-minute/static	m	seamounts
Time-lagged Chlorophyll-a concentration (_2 month)	log_chl_2m	log10	0.083 degrees/daily	mg/m^3	chl
Outputs from SEAPODYM ocean model Lehodey et al. (2010)					
Epipelagic micronekton(pb)	epi_pb		0.083 degrees/daily	g/m^2	SEAPODYM ocean model (Lehodey et al., 2014; Lehodey et al., 2010)
Mesopelagic micronekton(pb)	meso_pb		0.083 degrees/daily	g/m^2	SEAPODYM ocean model (Lehodey et al., 2014; Lehodey et al., 2010)
Migrant mesopelagic micronekton(pb)	mmeso_pb		0.083 degrees/daily	g/m^2	SEAPODYM ocean model (Lehodey et al., 2014; Lehodey et al., 2010)
High migrant bathypelagic micronekton (pb)	hmbathy_pb		0.083 degrees/daily	g/m^2	SEAPODYM ocean model (Lehodey et al., 2014; Lehodey et al., 2010)
Lower trophic level plankton(pb)	pk_pb		0.083 degrees/daily	g/m^2	SEAPODYM ocean model (Lehodey et al., 2014; Lehodey et al., 2010)

4.1.3 Model fitting and selection

Generalized Additive Models (GAMs) were run using the presence or absence of sperm whales per segment as a response variable with a binomial distribution. Three different models were created: 1) one model including only the environmental variables as predictors (ENM 2a); 2) a second model including only the prey-related variables (ENM 2b); and 3) a model combining environmental and prey-related variables (ENM 2c; Table V). To avoid overfitting and preserve the ecological interpretability of functional relationships, the number of knots of the smoothers were limited whenever appropriate. The restricted maximum likelihood (REML) optimization method was used to fit the models (S. N. Wood, 2006, 2010).

A cross-validation method was used to assess the accuracy of the GAM models. Models were built with 75% of the original sighting's dataset (training dataset), randomly selected, and evaluated using the remaining 25% (test dataset). The predictive performance was measured using the concordance index (C-index) through the R package 'Hmisc' (Harrell Jr, 2019). The predictive discrimination of this index is related to a rank correlation between predicted and observed outcomes. The C-index ranges between 0 and 1 with values from 0.6-0.7 indicating "satisfactory" discrimination, 0.7-0.8 "good discrimination", 0.8-0.9 "very good" and 0.9-1 "excellent discrimination" (Swets, 1988). Models were built using the package "MASS" (Ripley & Venables, 2019) and "mgcv" (S. Wood, 2019) within the R environment. The modelling procedure is summarized in Figure 4.

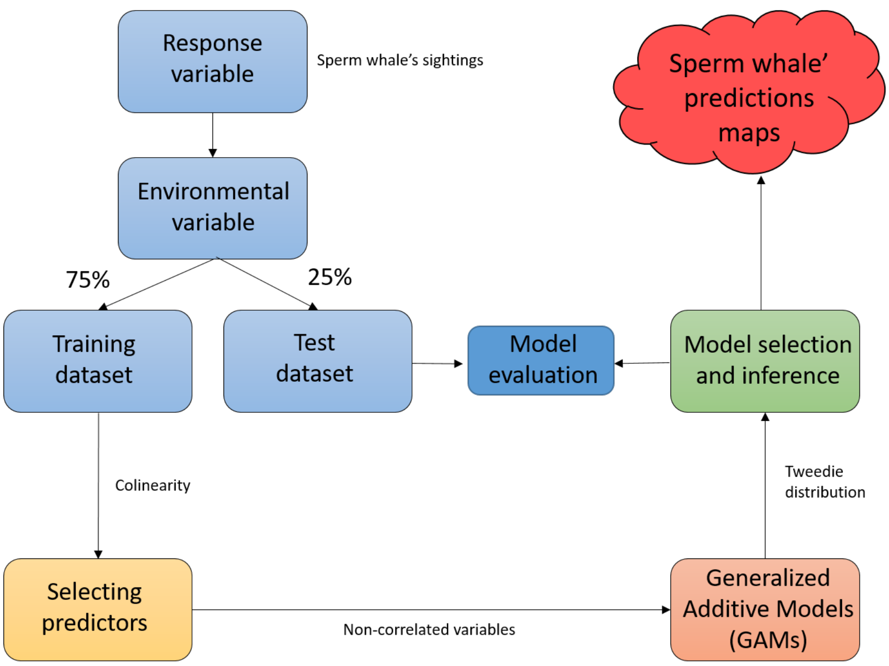

Figure 4: Workflow of the habitat modelling procedure applied to the sperm whale sighting data.

4.2 ENM 2 Results

The model including only environmental variables (ENM 2a) had a higher deviance explained and predictive performance than models with only prey-related variables (ENM2b) or a combination of both (ENM 2c; Table V). Trying to explain these results is outside the scope of this document and will be addressed in a future peer reviewed publication.

The best environmental model included three environmental variables and an interaction between longitude and latitude. This model explained 3.1 % of the total deviance (Table V).

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101076983. UK participants in Project OCEAN are supported by UKRI grant numbers 10038659 (Lloyds Register) and 10052942 (The Nautical Institute).

Table V. Summary of details and results of the three sperm whale GLM models (ENM 2a, b, c).

	Model		
	ENM 2a	ENM 2b	ENM 2c
Co-variate types	Environmental variables	Prey-related variables	Prey + environmental
Dynamic co-variates	Chlorophyll-a, sea surface temperature	SEAPODYM	Chlorophyll-a, sea surface temperature SEAPODYM
Deviance explained	3.1	<1	1.36
Predictive performance score (C-index train data)	0.66	0.51	0.61
Predictive performance score (C-index test data)	0.66	0.51	0.59
Predictive performance interpretation	Satisfactory	Low	Low
Data available at Copernicus Marine Service (CMEMS)	Yes	Yes	Yes
Best spatial resolution	0.083 degrees	0.083 degrees	0.083 degrees
Best temporal resolution	Daily	Daily	Daily
Hindcast data to build model	Yes	Yes	Yes
Near-real time data for daily predictions	Yes	No	No
Transferability of the variables	High	Low	Low

Sea surface temperature (SST) and chlorophyll the previous two months (log_chl_2m) had a positive relationship with the presence of whales in the study area (Figure 5). The highest probability of sperm whale's presence occurred at depths between 1000, and 2000 m. Model evaluation showed a satisfactory ability to discriminate between areas where sperm whales were present and absent (Table V).

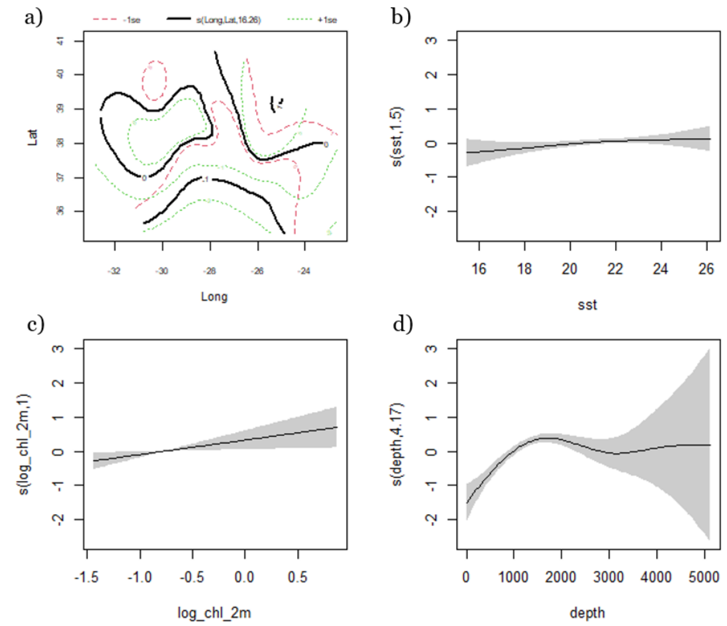

Figure 5: Functional relationships between sperm whale occurrence probability and environmental predictors: a) interaction Latitude vs Longitude, b) sea surface temperature (SST), chlorophyll from the previous second month (log_chl_2m) and depth.

The predictive performance of the best model (ENM 2a) was similar to previous studies for the sperm whales (Table VI). The predictive performance of these models ranged from 0.58 to 0.85, with a mean of 0.7 (SD: 0.06), close to the 0.66 obtained for the best GAM model of this study. The studies included a large variety of modelling approaches (GAM, ENFA, GLM, MARS, PCA and Maxent) from different study areas (Mediterranean Sea, Pacific Ocean and North Atlantic) (Fiedler et al., 2018; Guerra, Dawson, Somerford, Slooten, & Rayment, 2022; Pace et al., 2018; Pirotta, Matthiopoulos, MacKenzie, Scott-Hayward, & Rendell, 2011; Praca, Gannier, Das, & Laran, 2009).

Table VI. Comparison of studies predicting the distribution of sperm whales. GAM: Generalized Additive Model; PCA: Principal Component Analysis; GLM: Generalized Linear Model; ENFA: Ecological Niche Factor Analysis; MARS: Multivariate Adaptive Regression Splines; Maxent: Maximum Entropy.

Study (reference)	Model	Predictive performance	Study area
This study (**ENM 2a**)	GAM	0.66	Central North Atlantic
Tobeña et al. (2016b) (**ENM 1**)	Maxent	0.76	Central North Atlantic
Praca et al. (2009)	PCA	0.58	Mediterranean Sea
Praca et al. (2009)	GLM	0.7	Mediterranean Sea
Praca et al. (2009)	ENFA	0.66	Mediterranean Sea
Praca et al. (2009)	MARS	0.79	Mediterranean Sea
Pirotta et al. (2011)	GAM	0.77	Mediterranean Sea
Fiedler et al. (2018)	GAM	0.65	Eastern tropical Pacific Ocean

Fiedler et al. (2018)	Maxent	0.69	Eastern tropical Pacific Ocean
Guerra et al. (2022)	GAM	0.65-0.82	Southern Pacific Ocean
Pace et al. (2018)	Maxent	0.65-0.85	Mediterranean Sea

Some examples of predictions based on monthly averages of the predictive co-variates used in ENM 2a, are given in Figure 6. Similar maps for daily or weekly predictions are also possible, by changing the temporal resolution of the co-variate data. Some of the patterns seen in these maps are similar to those seen in the maps produced using ENM 1, namely higher probability of whale occurrence near the islands of the eastern group of the Azores, and southwest of the central group. The overall probability values for whale occurrence decrease along the season, in a similar manner as predicted for the habitat quality in ENM 1.

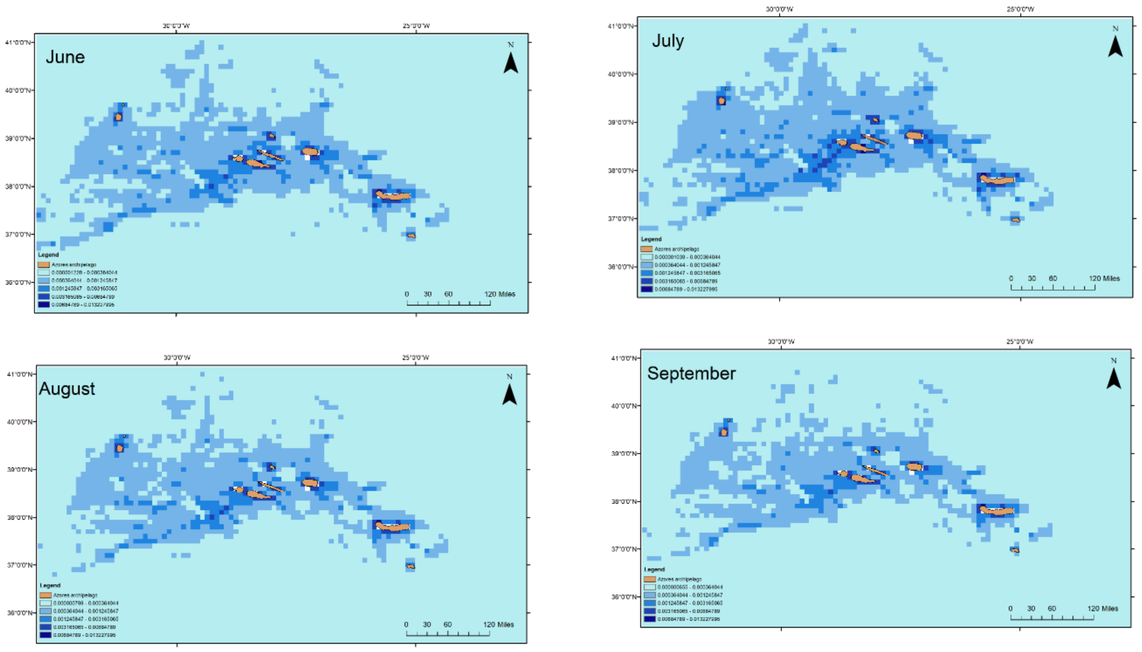

Figure 6: Predicted monthly sperm whale occurrence (occurrence probability per km²) from June to September based on ENM 2a.

4.3 ENM 2 Conclusions

Among the three GAM models that were fitted to the sperm whale sightings POPA data was a relatively simple model (ENM 2a) driven by geographic (latitude and longitude interaction), topographic (depth) and remote sensed co-variates (sea surface temperature and chlorophyll-a concentration).

Inclusion of modelled prey fields from SEAPODYM (Lehodey et al., 2014; Lehodey et al., 2010) did not, contrary to expected, improve model performance. On the contrary, these co-variates seem to degrade model performance (Table V). Discussing these seemingly perplexing results is out of the scope if this report. However, currently the prey co-variates are only available for past periods and, as such, not yet available for forecasting. Their inclusion in a predictive model at this time would be unfeasible, anyway. Testing their inclusion was an academic exercise with the goal of having a better model ready in case these co-variates proved to be informative and became available for forecasting in the near future.

As with any ENM model, ENM 2a can and should be improved with new data, and subject to regular performance testing if it is to be used as a predictive tool to inform management and operational decisions (Sofaer et al., 2019). However, in its current state, the model is at TLR 7 and can be operationalized for regular forecasting. Specifically, for the goas set in the OCEAN Grant Agreement, the model can be used to demonstrate the regular production predictions of areas with good conditions for whale aggregation (Task 4.3) and the communication of these predictions to mariners through the ENHI (WP6).

5 Final model selection

This section explains the reasoning for selection of one of ENM that will be used in Task 4.3 for operationalization of the regular forecasting of areas with conditions for whale aggregation. Reasoning for model selection took into consideration model performance, spatial and temporal resolution, and ease of implementation in the OCEAN framework.

5.1 Comparison of model results

5.1.1 Predictive power of models

Both ENM 1 and ENM 2a included depth, sea surface temperature and time-lagged Chlorophyll-a (2 months prior to sightings month) among the most informative co-variates, although ENM 1 had a higher number of retained covariates than ENM2.

For both ENMs, the highest presence of sperm whales was identified at depth ranging from 1000 to 3000 meters, with the maximum around 2000 meters. For SST and chl-a, some differences were found between the presence of the species and these variables which could be due to the differences in the periods covered by the two ENMs. Although there was a general pattern for these two variables, a high inter-annual variability has been observed in the Azores region (Amorim et al., 2017). The period covered by the ENM 1 (2004-2009) was characterized by warmer waters and less productive waters (higher SST and lower Chl-a) than the long-term mean (2001-2016) covered by the ENM 2. This effect may be further complicated because for ENM 1, the SST values were obtained from monthly averages. At the time of analysis, the daily data for SST (and other remote-sensed co-variates) was too patchy to enable inclusion in the model.

All three ENM 2 models had a low performance in explaining the variability, which was more drastic in ENM 2b and ENM 2c and is translated by the deviance explained values (Table V). The low deviance explained is common in GAMs used in cetacean distribution studies (Forney et al., 2012) and may be the consequence of modelling datasets with many absences (Welsh et al., 1996). Nevertheless, the deviance explained obtained for ENM 2a model is comparable with other studies predicting the distribution of the sperm whale (Table VI). Due to the method used, this metric is not available for ENM 1 and the models cannot be directly compared on that respect.

ENM 1 model had a moderately higher predictive performance than ENM 2a (Table VI). Differences on the performance metric could have been affected by the modelling approach (MaxEnt vs GAM model) and by the dataset used. However, ENM 2a model highlighted similar preference areas by sperm whales around the study area as those identified by the ENM 1 model.

Overall, both models have a predictive performance in line with other similar models for the same species in different regions (Table VI) and are, in that respect, analogous.

5.1.2 Spatio-temporal resolution

The dynamic co-variates (Chlorophyll-a and sea surface temperature) used to fit ENM 1 and ENM2a were obtained from distinct data products (Table III, Table IV), which resulted in some differences in spatio-temporal resolution.

For the ENM 1, the spatio-temporal resolution of these two variables were monthly averaged composites at a 0.04° obtained from oceanographic products from the National Aeronautics and Space Administration (NASA) Goddard Space Flight Center's Ocean Data Processing System (ODPS) (Campbell et al., 1995). For ENM 2a, the spatial resolution was 0.04º for Chlorophyll-a and 0.083º for sea surface temperature, both at a daily temporal scale, obtained from the Copernicus Marine Environment Monitoring Service (CMEMS).

The Copernicus Marine Service has oceanographic products for near-real time predictions for Chlorophyll-a and sea surface temperature. For Chlorophyll-a, the same product used to fit the models is still ongoing and available at a spatial resolution of 0.04º and two different temporal resolutions, daily and monthly. For SST, the product that was used to fit the GLM models was discontinued and replaced by a product with spatial resolution of 0.1º at the same daily and monthly temporal resolutions. The effect of this change is a coarser spatial resolution. However, there is no breach of model assumptions; as best practice, the resolution of model predictions should match that of the co-variate with lowest resolution to avoid an artificial increase in model resolution (Haining, 2003). However, in this case, there was a decrease in resolution and not the contrary, thus model assumptions are maintained.

ENM 1, on the other hand, used a very low temporal resolution (monthly) for the dynamic data products, for reasons already mentioned (see discussion in 5.1.1). These products are based on weighted means along the entire period (30 days). As a result, the variable values will be smoothed which, in turn, will be reflected in the accuracy of the estimation of habitat preferences (Haining, 2003). Although ENM 1 can produce predictions based on products with finer temporal resolutions, some caution should be exercised since this corresponds to an artificial increase in resolution.

5.1.3 Implementation in the OCEAN framework

Both models, ENM 1 and ENM 2a, are at TLR 7 and can be operationalized as a forecasting product. Operationalization will result from the work developed in Task 4.3 and must take into account not only model performance but the workflow that is necessary to create predictions at regular intervals (Figure 7).

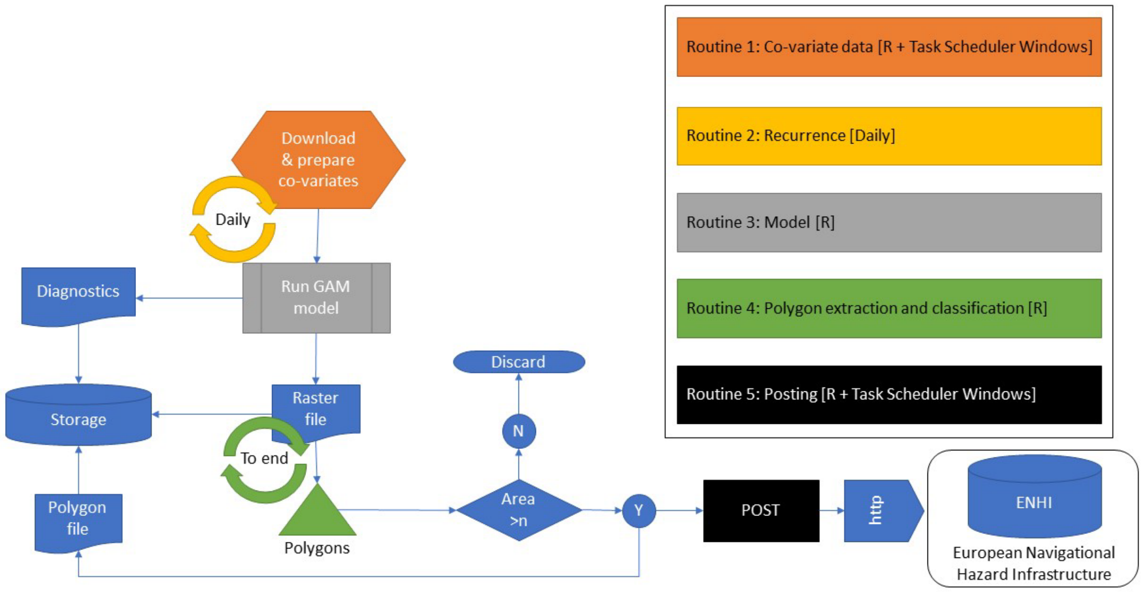

Figure 7: Framework for regular whale habitat predictions for distribution through the European Navigational Hazard Infrastructure.

The regularity of the forecasts must be tailored to the goals of the forecasts. For example, weather forecasts are usually updated daily or at infra-daily intervals, because users are interested in the weather conditions a few hours to days ahead. On the other hand, climatological forecasts (used for example in agriculture or civil protection preparation), based on climatological series, are tailored at giving an outlook of average conditions for more protracted periods and do not need to be produced at such finer temporal scale.

When *en route*, mariners are interested in forecasts with daily or nearly daily resolution, especially when dealing with safety issues, which is the case when avoiding ship strikes to marine mammals. Thus, ideally, the predictions based on the ENM chosen for operationalization in OCEAN should be daily, or nearly daily.

The CMEMS oceanographic products used to fit ENM 2a have the advantage of being distributed with a daily resolution, enabling the production of daily predictions. Another aspect to be taken into account is that CMEMS run by the European Commission, within the EU Earth Observation Programme.

In principle, it is possible to produce predictions using the same co-variates with ENM 1. However, there are some problems with that approach related to best practice. The oceanographic products that were used to fit ENM 1 were obtained from a different service (ODPS), and the way the products are created is not identical with CMEMS, raising the question of transferability of the model results when using alternative predictive co-variates to those used to fit the model. The best practice would be to fit a new model using the CMEMS variables with the approach used for ENM 1, but that would come at the cost of delays that are unnecessary and defeat the purpose of the work already done in Task 4.2.

Another issue, already brought up in Section 5.1.2, relates to the temporal resolution of the oceanographic data used to fit ENM 1. The oceanographic co-variates used to fit ENM 1 had a monthly resolution, and creating daily predictions can have effects that should be tested prior to implementation. This, again, would translate to unnecessary delays.

Finally, there are also considerations on processing requirements for both models. Model ENM 1 is somewhat more complex than ENM 2a in relation to the number and processing of co-variates.

ENM 1 is driven by six co-variates, with three dynamic co-variates that need to be downloaded daily prior to running the model. ENM 2a is driven by a slightly smaller set of co-variates, four, with two dynamic ones. Furthermore, the calculation of local variation of productivity (Table I) necessary to run ENM 1 is resource intensive and prone to computation crashes that require human intervention for resolution.

5.2 Selection of final model to be implemented in OCEAN

Taking into account the comparison made between ENM 1 and ENM 2a in Section 5.1, the choice about which model to carry on into Task 4.3 fell on **ENM 2a**, based on several aspects.

ENM 2b, and ENM 2c were not considered both due to very low predictive performance and because the some of the co-variates that they used are not available on a contemporary basis. Thus, they cannot be applied in regular prediction, even if the predictive performances of these models were good.

The sightings dataset used to fit ENM 2a is more robust than that used to fit ENM 1, comprising of a longer period and a larger number of sightings. Furthermore, by calculating sightings per unit effort, the predictions resulting from ENM 2a are calculated as a probability of occurrence. The predictions outputted by ENM 1 are calculated as a habitat quality score that is more complex to interpret. One of the processes developed in Task 4.3 will be responsible to create polygons from the prediction maps produced by the ENM, based on a threshold value (Figure 7). The interpretation of the probabilities produced by ENM 2a is more intuitive than the arbitrary values produced by ENM 1.

Despite having a larger dataset, ENM 2a had a worse predictive performance than ENM 1. However, as mentioned above, both models had predictive performances within what is usual for these type of models for the sperm whale (Table VI). Thus, the performance of both models can be considered analogous, in the sense that none performed neither exceptionally well or unfairly (which is the case for ENM 2b and ENM 2c).

ENM 2a allows for an appropriate temporal resolution of daily predictions. On the other hand, it would be necessary to test if that same temporal resolution is possible with ENM 1, since it fitted with a much different temporal resolution (monthly). Although that could be an interesting academic exercise, for the purposes of the project, that would be counterproductive.

The same principle applies to the oceanographic covariates to be used in the predictions. ENM 2a was originally fit with CMEMS co-variates, distributed by the EU-run Earth Observation Programme. ENM 1 was fit with ODPS co-variates distributed by US-run NASA. Producing predictions based on CMEMS co-variates with ENM 1 would require a time-consuming verification of transferability that would also be counterproductive.

Finally, ENM 2a is a simpler model unattended operation, requiring fewer resources, which is optimal for the following work developed in Task 4.3.

6 Conclusions

6.1 Main results

The goal of Task 4.2 was to enable forecasting areas with enhanced conditions for the occurrence of whale aggregations, based on ENM predictions. The work developed in Task 4.2 supports the subsequent work developed in Task 4.3 and the work of WP6, and required several sub-tasks:

- Identification of case study area(s);
- Identification of case study species;
- Informed decision on case study(ies) to develop;
- Identification of available ecologically relevant predictive co-variates;
- Selection of modelling approach;
- Data compilation, preparation;
- Model fitting;
- Model performance testing;
- Informed selection of modelling approach to be used in Task 4.3.

The process to choose the study area and species is detailed in Section 2.2. Discussions with specialists held during and after the *Marine Mammal Ship Strike Mitigation* held under Task 4.1 [43] helped in narrowing the group of species (large whales) and area (Azores). Subsequently, a systematic method for case study species selection was developed and applied, resulting in the selection of the sperm whale as the ideal case study (Table I).

The state-of-the-art ENM for the sperm whale in the Azores was published in 2016 by Tobeña et al. (2016b) (ENM 1; Section 3). Since then, more data became available, and some limitations of the model developed by Tobeña et al. (2016b) could be addressed. Thus, three new models (ENM 2a, b, c; Section 4) were fitted, using a larger and longer time-series sightings dataset and a different modelling approach, which more suited the goals set by the OCEAN Project.

After testing all models for predictive performance, one model stood out (ENM 2a). Although the predictive performance of model ENM 2a is not outstanding, it is *on par* with similar models for the sperm whale in other regions of the world and with the other similar model for the Azores (ENM 1; (Tobeña et al., 2016b)), as illustrated in Table VI.

Finally, analysing not only the predictive performance, but also the spatial and temporal resolution of the models, and the adequacy for the goals set by the OCEAN project, namely in Task 4.3, ENM 2a was considered to have some advantages over ENM 1. Specifically, ENM 2a can give a much better temporal resolution than ENM 1, translated by daily predictions, is relatively simple and not as resource intensive as ENM 1, and utilizes predictive co-variates distributed by the EU-run Earth Observation Programme, which was considered an advantage for an EU-funded project.

Overall, all sub-tasks were developed during the scheduled period, as reported here, and the work resulting from Task 4.2 will, as planned, serve as the basis to develop a framework for regular whale habitat predictions in Task 4.3.

6.2 Limitations

Every ecology student will eventually come across these words by George Box: "Essentially all models are wrong, but some are useful [...]" (Box & Draper, 1986). These inspired nine words have become a fundamental concept in evidence-based conservation policy. A little less known, however, is the continuation of Box's reasoning: "[...] the practical question is how wrong they do [the models] have to be to not be useful." (Box & Draper, 1986).

The sperm whale has proven to be a very challenging species when it comes to developing ENMs, as illustrated by the overall moderate predictive performance of the studies in Table VI (including ENM 1 and ENM 2a included here).

Specifically, for ENM 1 and ENM 2a, there are at least four possible reasons limiting the improvement on the predictive performance:

1. Very complex life history of the species, with different population segments having distinct distribution drivers, but occurring simultaneously in the same areas;
2. Lack of good data on prey distribution;
3. Lack of data on other ecologically relevant drivers (e.g.: predators; competitors; stressors);
4. Poor spatio-temporal resolution of some of the co-variates available for modelling.

We tried to address some of these issues, for example by using modelled prey-biomass from SEAPODYM to fit models ENM 2b, c. However, these models fared worse than the models using only environmental co-variates (ENM 1, ENM 2a), indicating that the SEAPODYM prey distribution based on macrozooplankton and micronekton (Lemarchand, 2016) do not reflect the distribution of the main prey of the sperm whale, and there is a mismatch between these distributions.

However, the utility of the models relates to the goals and how the results are presented. As demonstrated, the model selected to be used in the follow-up work in OCEAN has a predictive performance similar with other sperm whale models. The question, following George Box advice, is if these models are bad enough to not be useful.

The main goal in Task 4.2 is to have a functional and realistic ENM that is representative of other models that may eventually be integrated in the ENHI developed in WP6. Looking beyond the temporal horizon of the OCEAN project, when the fully deployed ENHI will receive data from different sources, including multiple ENMs. In that sense, ENM 2a fulfils the expectations, for being a model that is, in all aspects, similar to other future ENMs that may be part of the ENHI, including their limitations.

On the more practical matter of using this specific model (ENM 2a) to inform navigation, we would be more careful. Since the predictive performance for ENM 2a was not outstanding (as with other models for the species; Table VI), there is a degree of uncertainty in the model predictions that has to be dealt with for actual utilization when informing management and operation (Sofaer et al., 2019). One of the ways of dealing with that uncertainty is by tuning the thresholds for the designation of risk areas to be more conservative. For example, by lowering the designation of a risk area from a probability of 75% of whale occurrence to 50%, the risk of missing areas with whales decreases, however the area and number of risk areas will increase, with effects to the shipping industry.

We are confident that the model can give information on broad areas, a few tens of nautical miles across, that have a higher probability of having sperm whales present. This is informative, as ships can re-route in advance accordingly to avoid these areas. However, and depending on the areas, a very conservative threshold for designation of risk areas may not be practical.

Another issue to bear in mind is that environmental changes may be affecting the distribution and even the ecology of marine mammals (e.g. Peters, Stockin, & Saltré, 2022; Simmonds & Eliott, 2009; Simmonds & Isaac, 2007).

Thus, when informing management and operation, ENM models should never be considered definitive, and must be constantly improved and checked for predictive performance (Sofaer et al., 2019). That need is integrated in the design of the framework for regular whale habitat predictions (Task 4.3), by having a modular design, allowing for a seamless transition when models change (Figure 7).

6.3 Further work

The work predicted for Task 4.2 ends with the submission of this report and publication of Deliverable 4.2 (see Section 7). No further work is predicted in this task. As planned, the results from Task 4.2 support the work developed in Task 4.3 (framework for regular whale habitat predictions), which will be detailed in due time through its own Report.

7 Deliverable form and availability

The Deliverable for Task 4.2 (D4.2) is the model ENM2 (Section 4), which is defined by a single line of code run in the statistical programming language R (R Development Core Team, 2015):

```
gam(presence ~ s(sst)+s(log_chl_2m)+s(depth)+s(sqrt_dist_smnt),
data=[dataframe], family=binomial, method="REML")
```

The model itself and the data to reproduce the results is made publicly available as D4.2 through the GitHub repository in https://github.com/AzWhaleLab/OCEAN.

However, the process to select the appropriate model is lengthy and time-consuming, since it involves fitting multiple model variations and comparing their performance and ecological coherence, until selection of a final model.

The code for the multiple model fitting iterations used for model selection is also made available in the same repository. It must be stressed that statistical model fitting for prediction of animal distribution is a highly complex subject and the code is presented only as an illustrative example of the process adopted in this work. It is assumed that end-users are proficient in habitat modelling and R statistical programming language.

Tasks 4.2 and 4.3 and their respective deliverables (D4.2 and D4.3) are intrinsically related and all the code being developed in the scope of Task 4.3 is also stored in the same repository. Nevertheless, at the time of the submission of this report, work developed in Task 4.3 is underway, as scheduled in the Grant Agreement, and as such the code is not finalized. In due time, the finished code will be published there, under a GNU General Public License 3.0.

8 References

Amante, C., & Eakins, B. W. (2008). *ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis*. Retrieved from Boulder, Colorado, USA:

Amorim, P., Perán, A. D., Pham, C. K., Juliano, M., Cardigos, F., Tempera, F., & Morato, T. (2017). Overview of the ocean climatology and its variability in the Azores region of the North Atlantic including environmental characteristics at the seabed. *Frontiers in Marine Science, 4*(56). doi:10.3389/fmars.2017.00056

Aspinall, R. (1992). An inductive modelling procedure based on Bayes' theorem for analysis of pattern in spatial data. *International Journal of Geographical Information Systems, 6*(2), 105-121. doi:10.1080/02693799208901899

Austin, M. (2007). Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling, 200*(1–2), 1-19. doi:http://dx.doi.org/10.1016/j.ecolmodel.2006.07.005

Beaumont, L. J., Hughes, L., & Poulsen, M. (2005). Predicting species distributions: use of climatic parameters in BIOCLIM and its impact on predictions of species' current and future distributions. *Ecological Modelling, 186*(2), 251-270. doi:http://dx.doi.org/10.1016/j.ecolmodel.2005.01.030

Bivand, R., Keitt, T., & Rowlingson, B. (2019). rgdal: Bindings for the geospatial data abstraction library. R Package Version 1.4-7 (Version 1.4-7). http://CRAN.R-project.org/package=rgdal. Retrieved from http://CRAN.R-project.org/package=rgdal

Box, G. E. P., & Draper, N. R. (1986). *Empirical model-building and response surface*: John Wiley \& Sons, Inc.

Breiman, L. (2001a). Random forests. *Machine Learning, 45*(1), 5-32. doi:10.1023/a:1010933404324

Breiman, L. (2001b). Statistical modeling: the two cultures. *Statistical Science, 16*(3), 199-231.

Campbell, J. W., Blaisdell, J. M., & Darzi, M. (1995). Level-3 SeaWiFS data products: spatial and temporal binning algorithms. In S. B. Hooker, E. R. Firestone, & J. G. Acker (Eds.), *NASA Tech Memo 104566, vol. 32*. Greenbelt, MD, USA: NASA Goddard Space Flight Center.

Clark, J. S., & Gelfand, A. E. (2006). A future for models and data in environmental science. *Trends in Ecology & Evolution, 21*(7), 375-380. Retrieved from http://www.sciencedirect.com/science/article/pii/S0169534706001030

Conchon, A. (2016). *Modeling marine zooplankton and micronekton [Modélisation du zooplancton et du micronecton marins]*. Université de La Rochelle, Retrieved from https://theses.hal.science/tel-01661554 (2016LAROS011)

Costanza, R., d'Arge, R., de Groot, R., Farber, S., Grasso, M., Hannon, B., . . . van den Belt, M. (1997). The value of the world's ecosystem services and natural capital. *Nature, 387*, 253-260.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., . . . Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography, 36*(1), 027-046. doi:10.1111/j.1600-0587.2012.07348.x

Dudík, M., Phillips, S. J., & Schapire, R. E. (2007). Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *Journal of Machine Learning Research, 8*, 1217-1260.

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101076983. UK participants in Project OCEAN are supported by UKRI grant numbers 10038659 (Lloyds Register) and 10052942 (The Nautical Institute).

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., . . . Phillips, S. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography, 29*(2), 129-151.

Elith, J., & Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics, 40*(1), 677-697. doi:doi:10.1146/annurev.ecolsys.110308.120159

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology, 77*, 802-813.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions, 17*, 43-57.

Fiedler, P. C., Redfern, J. V., Forney, K. A., Palacios, D. M., Sheredy, C., Rasmussen, K., . . . Ballance, L. T. (2018). Prediction of large whale distributions: A comparison of presence–absence and presence-only modeling techniques. *Frontiers in Marine Science, 5*, 2018.00419. doi:10.3389/fmars.2018.00419

Fitzpatrick, M. C., Gotelli, N. J., & Ellison, A. M. (2013). MaxEnt versus MaxLike: empirical comparisons with ant species distributions. *Ecosphere, 4*(5), art55. doi:10.1890/es13-00066.1

Forney, K. A., Ferguson, M. C., Becker, E. A., Fiedler, P. C., Redfern, J. V., Barlow, J., . . . Ballance, L. T. (2012). Habitat-based spatial models of cetacean density in the eastern Pacific Ocean. *Endangered Species Research, 16*(2), 113-133. Retrieved from http://www.int-res.com/abstracts/esr/v16/n2/p113-133/

Fraser, D. A. S. (1968). A Black Box or a comprehensive model. *Technometrics, 10*(2), 219-229.

Grossi, F., Lahaye, E., Moulins, A., Borroni, A., Rosso, M., & Tepsich, P. (2021). Locating ship strike risk hotspots for fin whale (*Balaenoptera physalus*) and sperm whale (*Physeter macrocephalus*) along main shipping lanes in the North-Western Mediterranean Sea. *Ocean & Coastal Management, 212*, 105820. doi:https://doi.org/10.1016/j.ocecoaman.2021.105820

Grumbine, R. E. (1994). What is ecosystem management? *Conservation Biology, 8*(1), 27-38. doi:10.1046/j.1523-1739.1994.08010027.x

Guerra, M., Dawson, S. M., Somerford, T. R., Slooten, E., & Rayment, W. J. (2022). Fine-scale habitat use of foraging sperm whales is driven by seafloor topography and water column structure. *38*(2), 626-652. doi:https://doi.org/10.1111/mms.12881

Guisan, A., Edwards Jr, T. C., & Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling, 157*(2–3), 89-100. doi:http://dx.doi.org/10.1016/S0304-3800(02)00204-1

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters, 8*(9), 993-1009. doi:10.1111/j.1461-0248.2005.00792.x

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I. T., . . . Buckley, Y. M. (2013). Predicting species distributions for conservation decisions. *Ecology Letters, 16*(12), 1424-1435. doi:10.1111/ele.12189

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling, 135*(2-3), 147-186.

Haining, R. P. (2003). *Spatial data analysis : theory and practice*. Cambridge, UK ; New York: Cambridge University Press.

Harrell Jr, F. E. (2019). Hmisc: Harrell Miscellaneous. R Package Version 4.2 (Version 4.2). https://CRAN.R-project.org/package=Hmisc. Retrieved from https://CRAN.R-project.org/package=Hmisc

Hastie, T., & Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science, 1*(3), 297-318.

Hijmans, R. J., & van Ettern, J. (2012). raster: Geographic analysis and modeling with raster data. R Package Version 2.0-12 (Version 2.0-12). http://CRAN.R-project.org/package=raster. Retrieved from http://CRAN.R-project.org/package=raster

IUCN. (2023). The IUCN Red List of Threatened Species. Version 2024-1. Retrieved from https://www.iucnredlist.org

Kramer-Schadt, S., Niedballa, J., Pilgrim, J. D., Schröder, B., Lindenborn, J., Reinfelder, V., . . . Wilting, A. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions, 19*(11), 1366-1379. doi:10.1111/ddi.12096

Lawler, J. J., White, D., Neilson, R. P., & Blaustein, A. R. (2006). Predicting climate-induced range shifts: model differences and model reliability. *Global Change Biology, 12*(8), 1568-1584. doi:10.1111/j.1365-2486.2006.01191.x

Lehodey, P., Conchon, A., Senina, I., Domokos, R., Calmettes, B., Jouanno, J., . . . Kloser, R. (2014). Optimization of a micronekton model with acoustic data. *ICES Journal of Marine Science, 72*(5), 1399-1412. doi:10.1093/icesjms/fsu233 %J ICES Journal of Marine Science

Lehodey, P., Murtugudde, R., & Senina, I. (2010). Bridging the gap from ocean models to population dynamics of large marine predators: A model of mid-trophic functional groups. *Progress in Oceanography, 84*(1–2), 69-84. doi:http://dx.doi.org/10.1016/j.pocean.2009.09.008

Lemarchand, G. A. (2016). The scientific productivity and the dynamics of self-organizing networks: Ibero-American and Caribbean countries (1966–2013). In H. Horta, M. Heitor, & J. Salmi (Eds.), *Trends and Challenges in Science and Higher Education* (pp. 29-46): Springer International Publishing.

Mac Nally, R. (2000). Regression and model-building in conservation biology, biogeography and ecology: The distinction between – and reconciliation of – 'predictive' and 'explanatory' models. *Biodiversity and Conservation, 9*(5), 655-671. doi:10.1023/a:1008985925162

Mannocci, L., Laran, S., Monestiez, P., Dorémus, G., Van Canneyt, O., Watremez, P., & Ridoux, V. (2014). Predicting top predator habitats in the Southwest Indian Ocean. *Ecography, 37*(3), 261-278. doi:10.1111/j.1600-0587.2013.00317.x

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

Monk, J., Ierodiaconou, D., Harvey, E., Rattray, A., & Versace, V. L. (2012). Are we predicting the actual or apparent distribution of temperate marine fishes? *PloS One, 7*(4), e34558.

Morato, T., Machete, M., Kitchingman, A., Tempera, F., Lai, S., Menezes, G., . . . Santos, R. S. (2008). Abundance and distribution of seamounts in the Azores. *Marine Ecology Progress Series, 357*, 17-21.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General), 135*(3), 370-384.

Nix, H. A. (1986). A biogeographic analysis of Australian Elaphid snakes. In *Atlas of Elapid snakes of Australia* (pp. 4-15): Australian Government Publishing Service.

Olden, J. D. (2008). Machine learning withouth tears: a primer for ecologists. *The Quarterly Review of Biology, 83*(2), 171-193.

Pace, D. S., Arcangeli, A., Mussi, B., Vivaldi, C., Ledon, C., Lagorio, S., . . . Ardizzone, G. (2018). Habitat suitability modeling in different sperm whale social groups. *82*(5), 1062-1073. doi:https://doi.org/10.1002/jwmg.21453

Peters, K. J., Stockin, K. A., & Saltré, F. (2022). On the rise: Climate change in New Zealand will cause sperm and blue whales to seek higher latitudes. *Ecological Indicators, 142*, 109235. doi:https://doi.org/10.1016/j.ecolind.2022.109235

Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., & Araújo, M. B. (2011). *Ecological niches and geographic distributions*. Princeton and Oxford: Princeton University Press.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling, 190*, 231-259.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J. R., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications, 19*(1), 181-197.

Pierce, D. (2019). ncdf4: Interface to Unidata netCDF. R Package Version 1.17 (Version 1.17). https://CRAN.R-project.org/package=ncdf4. Retrieved from https://CRAN.R-project.org/package=ncdf4

Pirotta, E., Matthiopoulos, J., MacKenzie, M., Scott-Hayward, L., & Rendell, L. (2011). Modelling sperm whale habitat preference: a novel approach combining transect and follow data. *Marine Ecology Progress Series, 436*, 257-272.

Praca, E., Gannier, A., Das, K., & Laran, S. (2009). Modelling the habitat suitability of cetaceans: Example of the sperm whale in the northwestern Mediterranean Sea. *Deep Sea Research Part I: Oceanographic Research Papers, 56*(4), 648-657. doi:http://dx.doi.org/10.1016/j.dsr.2008.11.001

Prieto, R. (2023). *Workshop on Marine Mammal Ship Strike Mitigation Report*. Paper presented at the Workshop on Marine Mammal Ship Strike Mitigation. Horta, 24-26 January 2023., Horta; Portugal.

Pueyo, S. (2012). Comparison of two views of maximum entropy in biodiversity: Frank (2011) and Pueyo et al. (2007). *Ecology and Evolution, 2*(5), 988-993.

R Development Core Team. (2015). R: A language and environment for statistical computing. Retrieved from http://www.R-project.org

Richards, C. L., Carstens, B. C., & Lacey Knowles, L. (2007). Distribution modelling and statistical phylogeography: an integrative framework for generating and testing alternative biogeographical hypotheses. *Journal of Biogeography, 34*(11), 1833-1845. doi:10.1111/j.1365-2699.2007.01814.x

Ripley, B., & Venables, B. (2019). MASS: Support Functions and Datasets for Venables and Ripley's MASS. R Package Version 7.3-51.3. https://CRAN.R-project.org/package=MASS. Retrieved from https://CRAN.R-project.org/package=MASS

Ritter, F. (2012). Collisions of sailing vessels with cetaceans worldwide: First insights into a seemingly growing problem. *Journal of Cetacean Research and Management, 12*(1), 119-127.

Ritter, P. (1987). A vector-based slope and aspect generation algorithm. *Photogrammetric Engineering and Remote Sensing, 53*(8), 1109-1111.

Schoeman, R. P., Patterson-Abrolat, C., & Plön, S. (2020). A global review of vessel collisions with marine animals. *Frontiers in Marine Science, 7*(292). doi:10.3389/fmars.2020.00292

Schröder, B. (2008). Challenges of species distribution modeling belowground. *Journal of Plant Nutrition and Soil Science, 171*(3), 325-337. doi:10.1002/jpln.200700027

Silva, M. A., Feio, R., & Prieto, R. (2002). *Interaction between cetaceans and the Azorean tuna – fishery: is this a real problem?* . Paper presented at the 16th Annual Conference of the European Cetacean Society, Liège, Belgium. poster retrieved from

Silva, M. A., Feio, R., Prieto, R., Goncalves, J. M., & Santos, R. S. (2002). Interactions between cetaceans and the tuna fishery in the Azores. *Marine Mammal Science, 18*(4), 893-901.

Silva, M. A., Machete, M., Reis, D., Santos, M., Prieto, R., Dâmaso, C., . . . Santos, R. (2011). A review of interactions between cetaceans and fisheries in the Azores. *Aquatic Conservation: Marine and Freshwater Ecosystems, 21*(1), 17-27.

Silva, M. A., Prieto, R., Cascão, I., Seabra, M. I., Machete, M., Baumgartner, M. F., & Santos, R. S. (2014). Spatial and temporal distribution of cetaceans in the mid-Atlantic waters around the Azores. *Marine Biology Research, 10*(2), 123-137. doi:10.1080/17451000.2013.793814

Simmonds, M. P., & Eliott, W. J. (2009). Climate change and cetaceans: concerns and recent developments. *Journal of the Marine Biological Association of the United Kingdom, 89*(1), 203-210.

Simmonds, M. P., & Isaac, S. J. (2007). The impacts of climate change on marine mammals: early signs of significant problems. *Oryx, 41*(1), 19-26.

Sinclair, S. J., White, M. D., & Newell, G. R. (2010). How useful are species distribution models for managing biodiversity under future climates? *Ecology and Society, 15*(1), 8p. [online].

Soberón, J. (2007). Grinnellian and Eltonian niches and geographic distributions of species. *10*(12), 1115-1123. doi:https://doi.org/10.1111/j.1461-0248.2007.01107.x

Soberón, J., Osorio-Olvera, L., & Peterson, T. (2017). Diferencias conceptuales entre modelación de nichos y modelación de áreas de distribución. [Conceptual differences between ecological niche modeling and species distribution modeling]. *Revista Mexicana de Biodiversidad, 88*(2), 437-441. doi:https://doi.org/10.1016/j.rmb.2017.03.011

Sofaer, H. R., Jarnevich, C. S., Pearse, I. S., Smyth, R. L., Auer, S., Cook, G. L., . . . Hamilton, H. (2019). Development and delivery of species distribution models to inform decision-making. *Bioscience, 69*(7), 544-557. doi:10.1093/biosci/biz045 %J BioScience

Stockwell, D., & Peters, D. (1999). The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science, 13*(2), 143-158. doi:10.1080/136588199241391

Stockwell, D. R. B., & Noble, I. R. (1992). Induction of sets of rules from animal distribution data: A robust and informative method of data analysis. *Mathematics and Computers* *in Simulation, 33*(5–6), 385-390. doi:http://dx.doi.org/10.1016/0378-4754(92)90126-2

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *240*(4857), 1285-1293. doi:doi:10.1126/science.3287615

Taylor, B. L., Baird, R., Barlow, J., Dawson, S. M., Ford, J., Mead, J. G., . . . Pitman, R. L. (2019). *Physeter macrocephalus (amended version of 2008 assessment)*. Retrieved from https://dx.doi.org/10.2305/IUCN.UK.2008.RLTS.T41755A160983555.en

Tobeña, M., Prieto, R., Machete, M., & Silva, M. A. (2016a). *Mapped cetacean habitat suitability and richness in the Azores, links to ArcGIS files*. Retrieved from: https://doi.org/10.1594/PANGAEA.864511

Tobeña, M., Prieto, R., Machete, M., & Silva, M. A. (2016b). Modeling the potential distribution and richness of cetaceans in the Azores from fisheries observer program data. *Frontiers in Marine Science, 3,* Article 202. doi:10.3389/fmars.2016.00202

Uriarte, M., & Yackulic, C. B. (2009). Preaching to the unconverted: Ecological Applications. *Ecological Applications, 19*(3), 592-596. Retrieved from http://dx.doi.org/10.1890/08-0743.1

Welsh, A. H., Cunningham, R. B., Donnelly, C. F., & Lindenmayer, D. B. (1996). Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling, 88*(1), 297-308. doi:https://doi.org/10.1016/0304-3800(95)00113-1

Whitehead, H. (2003). *Sperm whales. Social evolution in the ocean*. Chicago; London: University of Chicago Press.

Winkler, C., Panigada, S., Murphy, S., & Ritter, F. (2020). *Global numbers of ship strikes: An assessment of collisions between vessels and cetaceans using available data in the IWC ship strike database* (IWC/68B/SC HIM09). Retrieved from https://archive.iwc.int/pages/view.php?ref=17562&k=c9f67e8c17

Wood, S. (2019). mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. R Package Version 1.8-29. https://CRAN.R-project.org/package=mgcv. Retrieved from https://CRAN.R-project.org/package=mgcv

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. New York: Chapman and Hall/CRC.

Wood, S. N. (2010). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 73*(1), 3-36. doi:10.1111/j.1467-9868.2010.00749.x %J Journal of the Royal Statistical Society Series B: Statistical Methodology

World Bank (Ed.) (2006). *Where is the wealth of nations? Measuring capital for the 21st century*. Washington: World Bank Publications.

Yackulic, C. B., Chandler, R., Zipkin, E. F., Royle, J. A., Nichols, J. D., Campbell Grant, E. H., & Veran, S. (2012). Presence-only modelling using MAXENT: when can we trust the inferences? *Methods in Ecology and Evolution*, n/a-n/a. doi:10.1111/2041-210x.12004

9 Annex 1: The Consortium

Høgskulen på Vestlandet	Inndalsveien 28 P.O Box 7030 5020 Bergen, Norway	https://www.hvl.no/en/
KYSTVERKET NORWEGIAN COASTAL ADMINISTRATION	Flathauggata 12 5523 Haugesund, Norway	www.kystverket.no
	Heroon Polytechniou 9 Zographou campus 15780 Athina, Greece	www.ntua.gr
CIMNE EXCELENCIA SEVERO OCHOA	C Gran Capitan, Edifici C1, Campus Nord UPC SN 08034 Barcelona, Spain	www.cimne.com
TELEDYNE RESON Everywhereyoulook	Fabriksvangen 13 3550 Slangerup, Denmark	www.teledynemarine.com/reson
OKEANOS	Rua Da Mae De Deus Reitoria Apatado 1422 9501 801, Ponta Delagada S Miguel Acores Portugal	www.uac.pt
KONGSBERG	Kongsberg Maritim: Strandpromenaden 50 Po box: 1009 3183 Horten, Norway	www.kongsberg.com
KONGSBERG	Kongsberg Seatex: Pirsenteret 7462 Trondheim, Norway	www.kongsberg.com
UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH	Calle Jordi Girona 31 08034 Barcelona, Spain	www.upc.edu
IRISH FERRIES	Alexandra Road Ferryport D01 W2F5 Dublin, Ireland	www.irishferries.com
IWDG	Merchants Quay V15 E762 Kilrush, Clare Ireland	www.iwdg.ie
The Nautical Institute	Lambeth road 202, SE 7LQ London, United Kingdom	www.nautinst.org
LR	Fenchurch street 71, EC3M 4BS, United Kingdom	www.lr.org

10 Annex 2: Project Summary

The OCEAN project is focused on enhancing operator awareness in navigation, to reduce the frequency of severe accidents like collision and grounding, to mitigate ship-strike risks to marine mammals, and to mitigate the risk presented by floating obstacles to ships.

The OCEAN project will contribute to an improved understanding of accident root causes, and will strive to reduce the resulting human, environmental and economic losses through socio-technical innovations supporting ship navigators.

The OCEAN consortium, coordinated by Western Norway University of Applied Sciences, includes 13 partner organizations across 7 different European countries from the industry, academia, NGOs and end users.

Around 3.000 maritime incidents occur every year in the European maritime fleet. 28% of these accidents are categorised as severe or very severe accidents, resulting in the loss of life onboard, pollution, fire, collisions or grounding. Navigational accidents are dominant in these statistics according to the European Maritime Safety Agency, be it for cargo, passenger or service ships.

The OCEAN project ambition is to contribute to the mitigation of navigational accidents by supporting the navigators to do an even better job than they do presently. The OCEAN consortium will address the most pertinent factors that may contribute to events becoming accidents: training, technical, human or organisational factors, operational constraints, processes and procedures, commercial pressures, and will recommend improvements and amendments to regulations, standards and bridge equipment design approaches.

OCEAN seeks to enhance navigational awareness “on the spot” and to improve the performance of evasive manoeuvring to avoid collision with near-field threats. The project will deliver and demonstrate several human centred innovations. For example, the 4D Situation Awareness Display which will be developed in the OCEAN project will improve the visualisation of navigational hazards, integrating current bridge information systems with marine mammal and lost floating containers detection and tracking capacity specifically developed by the project.

Going further, the project will design and implement a European navigational hazard data infrastructure to feed multi-source observations and hazard predictions relating to floating containers and large aggregations of marine mammals into the existing distributed maritime warning infrastructure. OCEAN seeks to transfer this data ecosystem to relevant European organisations for deployment and maintenance.

Co-funded by Horizon Europe, the European Union’s research and innovation programme, the consortium of 13 members represents 7 European countries, Norway, Greece, Spain, Denmark, Portugal, Ireland and UK, all located on major European coastal regions. Members include a coastal administration, a ship operator, maritime safety and transport researchers, marine mammal ecology and conservation experts, companies specialised in maritime information systems and sensors, a professional organisation, a risk and safety management organisation, as well as data infrastructure, data fusion and satellite imaging specialists.

UK participants are supported by UK Research and Innovation Grant Number 10038659 (Lloyd’s Register) and Grant Number 10052942 (The Nautical Institute).