



**HAL**  
open science

# Modelling distributions of rare marine species : the deep-diving cetaceans

Auriane Virgili

► **To cite this version:**

Auriane Virgili. Modelling distributions of rare marine species: the deep-diving cetaceans. Agricultural sciences. Université de La Rochelle, 2018. English. NNT : 2018LAROS003 . tel-02009798

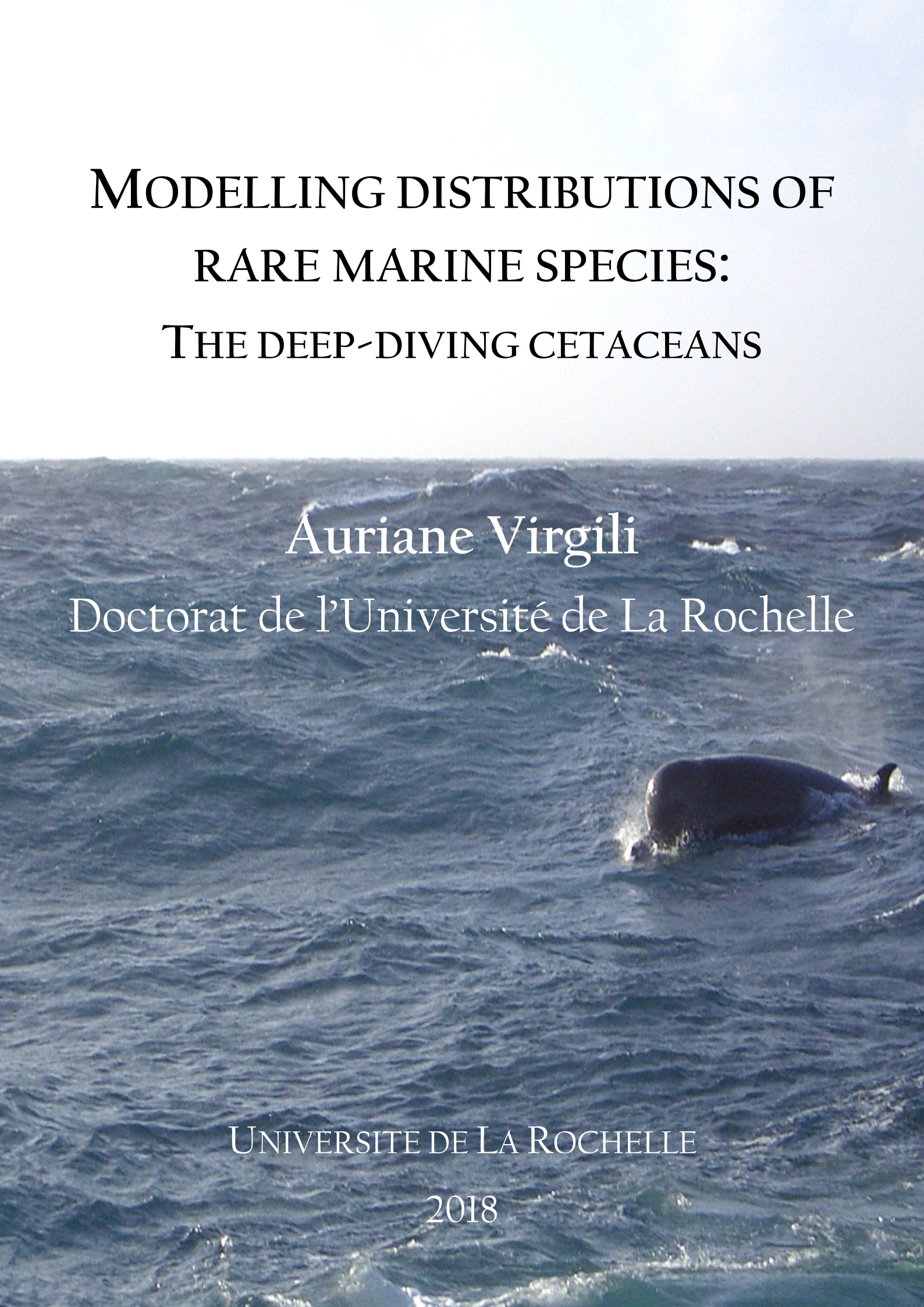
**HAL Id: tel-02009798**

**<https://theses.hal.science/tel-02009798>**

Submitted on 6 Feb 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



MODELLING DISTRIBUTIONS OF  
RARE MARINE SPECIES:  
THE DEEP-DIVING CETACEANS

*Auriane Virgili*

Doctorat de l'Université de La Rochelle

UNIVERSITE DE LA ROCHELLE

2018





UNIVERSITÉ DE LA ROCHELLE

ÉCOLE DOCTORALE GAY LUSSAC

Centre d'Etudes Biologiques de Chizé  
UMR 7372 Université de La Rochelle – CNRS

THÈSE présentée par :

**Auriane VIRGILI**

Soutenue publiquement le 11 Janvier 2018  
pour l'obtention du grade de Docteur de l'Université de La Rochelle

Spécialité : Biologie de l'Environnement, des Populations, Ecologie

**Modelling distributions of rare marine species:  
The deep-diving cetaceans**

JURY:

Ana RODRIGUES  
Len THOMAS  
Odile GERARD  
Etienne RIVOT  
Pascal MONESTIEZ  
Vincent RIDOUX

Directeur de Recherche, CNRS - CEFE, Rapporteur, Présidente  
Professeur, Université de St Andrews, Rapporteur  
Docteur, DGA Toulon Techniques navales, Examineur  
Docteur, AGROCAMPUS OUEST, Examineur  
Directeur de Recherche, INRA, Directeur de thèse  
Professeur, Université de La Rochelle, Directeur de thèse





# REMERCIEMENTS

« -C'est une bonne situation ça [doctorant] ?

- Vous savez, moi je ne crois pas qu'il y ait de bonne ou de mauvaise situation. Moi, si je devais résumer ma vie aujourd'hui avec vous, je dirais que c'est d'abord des rencontres. Des gens qui m'ont tendu la main... » .

« Astérix et Obélix : mission Cléopâtre » de Alain Chabat (2002)

Je vous l'accorde cette citation est extraite d'un de mes films préférés mais je trouve qu'elle résume exactement ce que je ressens en écrivant ces quelques mots. Au cours de ma thèse (et de mon master, je triche un peu), j'ai eu l'occasion de rencontrer de nombreuses personnes avec qui j'ai partagé beaucoup de moments et que je souhaite remercier.

Avant tout, je souhaite remercier sincèrement Vincent, mon directeur de thèse, qui m'a soutenue tout au long de ces années et ce depuis mon stage de fin d'étude. Merci d'avoir cru en moi jusqu'au bout et de m'avoir aidée à mener à bien ce projet. Tes conseils ont été d'une très grande aide quand je me sentais perdue au milieu de tous ces modèles et de toutes ces cartes.

Je tiens également à remercier Pascal, mon deuxième directeur de thèse. Même si nos échanges étaient moins fréquents, tu as su m'aiguiller dans mes choix méthodologiques et a été présent tout au long de cette expérience.

Je remercie la DGA (Direction Générale de l'Armement) qui a financé ce projet dans son intégralité, même lorsque celui-ci s'est vu retirer une partie de son financement. Merci, de m'avoir accordé des financements qui m'ont permis de participer à de fabuleuses conférences. Merci à vous Odile Gérard et Sophie Laran d'avoir été à l'origine du projet et de vous être battues pour qu'il se réalise. Et également, merci Odile de m'avoir suivie tout au long de ce projet.

Merci à Ana Rodrigues, Len Thomas, Etienne Rivot et Jean-Benoit Charrassin d'avoir accepté de participer à mon jury et d'avoir évalué cette thèse. Désolée Etienne d'avoir changé de matériel d'étude mais comme tu as pu le remarquer j'ai tout de même continué à faire des statistiques.

Je remercie chaleureusement toutes les personnes qui ont collecté et m'ont fourni les données que j'ai utilisées au cours de ma thèse, sans vous ce travail n'aurait jamais pu être réalisé. Merci de participer activement aux articles, votre aide m'est très précieuse. Un merci particulier à Ana Cañadas d'avoir participé à mon comité de thèse et de m'avoir conseillée à un moment clé de ma thèse.

Merci à Patrick Lehodey et Beatriz Calmettes de nous avoir accueilli à CLS et de nous avoir fourni les données de SEAPODYM qui laissent présager de futures collaborations.

Merci à toute l'équipe PELAGIS d'avoir participer de près ou de loin à cette thèse. Merci pour votre accueil, nos échanges et nos discussions, qui m'ont souvent valu plusieurs heures de boulot en plus, mais c'était pour la bonne cause ! Merci énormément à toi Matthieu, sans ton aide je ne serais pas arrivée là aujourd'hui, merci pour ta patience et ton implication qui a grandement aidée à la réalisation de ce travail.

Un énorme merci à l'équipe REMMOA, Sophie, Olivier et Ghislain qui m'ont offert la possibilité de participer à une expérience époustouflante. Survoler les lagons et l'océan et observer la mégafaune marine de Nouvelle-Calédonie est une expérience que je n'oublierai jamais ! Merci de m'avoir fait découvrir les réalités du terrain, derrière l'ordinateur, on oublie vite les fondamentaux. Merci aux pilotes

et aux observateurs d'avoir contribué à la réalisation de cette magnifique expérience et merci à toi Morgane d'avoir partagé mes délires et d'avoir rendu cette expérience exceptionnelle. *Tchou, tchou !*

Merci à mes deux stagiaires, Mélanie et Laura, de m'avoir aidée dans ce travail, vous m'avez boostée aux moments où j'en avais le plus besoin et avez contribué à l'aboutissement de ce travail. Un merci tout particulier à Laura pour ses admirables dessins qui m'ont permis d'illustrer mes chapitres de thèse, ton « bacalot » est encadré !

Merci à tous les doctorants, anciens doctorants, stagiaires, contractuels : Pierre, Henri, Laurent, Simon, David, Thomas, Anne, Cyrille, Aurélie, Alice S, Mathilde, Adrien, Kevin, Moussa, Isabel..., merci pour votre soutien et pour les rigolades, c'était un plaisir de se changer les idées avec vous. Bonne continuation à vous ! Je dois évidemment ajouter un merci ++ à mes coéquipiers de badminton, merci pour ces parties endiablées !

Ah, comment pourrais-je vous remercier toutes les quatre, en ne citant pas vos alias ? Un énorme merci à vous Amandine, Emeline, Julie et Ludivine, tous ces moments de rigolade ont augmenté mon espérance de vie d'au moins 10 ans ! Merci pour votre soutien, nos discussions et nos séances de psychanalyse qui m'ont fait un très grand bien. Vous avez toujours répondu présentes (même au point de s'embarquer dans une formation intense en statistiques !), et je vous en remercie. Désolée je ne m'épilouerais pas plus car j'ai aquaponey...

Merci également à Maud et Karine, pour votre bonne humeur à chaque fois que je me rendais au CCA.

Un merci tout particulier pour toi Alice, ma collègue-voisine. Merci pour ces moments de rire, de discussion et de procrastination, tous ces moments qui m'ont permis de me détacher de la thèse et de souffler un peu. Merci aussi pour nos échanges, sérieux cette fois-ci, qui nous ont permis d'avancer ensemble dans notre travail. Merci de m'avoir supportée pendant ces deux années et surtout pendant ces derniers mois, je me doute que la cohabitation n'a pas été facile tous les jours, alors merci. Merci encore pour Bubulle et Croki, leur place sur le bureau leur convient à ravir ! Je te souhaite une bonne continuation pour la suite, tu verras, la dernière année ce n'est pas si pire ! Même si je ne suis plus physiquement avec toi dans le bureau, nous pourrons continuer ces instants de brainstorming qui nous ont tant aidées.

Si je pense à Alice, je pense également à Nicolas ! Merci également à toi Nicolas pour tes incursions dans le bureau qui étaient toujours un plaisir !

A toi, ma petite Charlotte, je souhaite te dire un énorme merci pour tout ! Merci pour ta patience lorsque je n'étais encore que ta petite stagiaire, ton aide, ton soutien, tes conseils, nos rigolades, nos séances de décorations de bureau et sans oublier tes insinuations de chansons intempestives ! Merci d'avoir été présente tout au long de cette expérience. J'ai été ravie de partager tous ces moments avec toi, surtout durant notre périple en Afrique du Sud, plein de souvenirs qui resteront gravés, à quand la prochaine fois ? Merci à toi et Benjamin de m'avoir si souvent invitée et accueillie, c'est toujours un réel plaisir et ça l'est encore plus depuis que je suis tatée de vos adorables bouts de choux. Je ne sais pas ce que l'avenir nous réserve mais j'espère qu'il ne nous séparera pas.

Merci à mes loulous de l'halieute, Seb, Quentin, Juliette, Antoine, Romain, Erwan, Damien, Jules, Pierre, pour nos week-ends retrouvailles qui m'ont permis de décompresser et de passer des super moments avec vous. Bonne continuation à vous et à quand le prochain week-end ?

Un merci tout particulier à Flavie, Eglantine, Pauline L et Pauline M. Merci d'avoir été présentes malgré mon emploi du temps un peu compliqué. Vous ne vous en êtes probablement rendues compte

mais nos week-ends passés ensemble ont été d'une grande aide, ils m'ont permis de relâcher la pression et c'est toujours un plaisir de passer du temps avec vous.

Je souhaite également remercier mes nouveaux amis toulousains, Fanny, Mathilde, Laura, Stéphane, Estéban, Gui, Fred, je suis ravie de vous avoir rencontrés et d'avoir eu l'occasion de passer ces bons moments en votre compagnie.

Merci également à Martine et Yannick de m'avoir si chaleureusement accueillie dans votre famille, ces moments de détente à Grenade m'ont permis de décompresser et de réattaquer du bon pied les lundis.

Un immense merci ne suffirait pas à remercier ma famille, mes parents, mon frère (Romane et Louna aussi bien sûr !), qui ont toujours été là pour moi et qui ont toujours cru en moi. Même si vous ne comprenez pas toujours ce que je fais, vous êtes toujours là pour me soutenir, me remonter le moral dans les coups durs et me changer les idées et pour cela je vous en remercie. C'est grâce à vous si j'en suis là et grâce à vous si j'ai trouvé ma voie. Merci !

Mes derniers mots iront pour toi Brice, merci pour ton soutien sans faille. Merci d'avoir cru en moi à tout instant, d'avoir supporté mes sautes d'humeur, de rigoler de mes bêtises, d'avoir pris soin de moi mais surtout d'être là. Même si tu affirmes que j'y serais arrivé sans toi, tu m'as permis d'atteindre le sommet avec plus de facilité et je t'en remercie.

Merci à tous !



# SCIENTIFIC PRODUCTIONS

## PEER-REVIEWED PUBLICATIONS

### First author publications

- Virgili, A.**, Authier, M., Boisseau, O., Cañadas, A., Claridge, D., Cole, T., Corkeron, P. et al. (in prep.). Combining visual surveys to model habitat of deep-diving cetaceans at the basin scale.
- Virgili A.**, Authier M., Monestiez P., Ridoux V. (in revision). How many sightings to model rare marine species distributions. PLoS One.
- Virgili, A.**, Racine, M., Authier, M., Monestiez, P., Ridoux, V. (2017). Comparison of habitat models for scarcely detected species. *Ecological Modelling* 346: 88-98.
- Virgili, A.**, Lambert, C., Pettex, E., Dorémus, G., Van Canneyt, O., Ridoux, V. (2017). Predicting seasonal variations in coastal seabird habitats in the English Channel and the Bay of Biscay. *Deep Sea Research Part II: Topical Studies in Oceanography* 141: 212-223.

### Other

- Lambert C., **Virgili A.**, Pettex E., Delavenne J., Toison V., Blanck A., Ridoux V. (2017). Habitat modelling predictions highlight seasonal relevance of Marine Protected Areas for marine megafauna. *Deep Sea Research Part II: Topical Studies in Oceanography* 141: 262-274.
- Pettex E., David L., Authier M., Blanck A., Dorémus G., Falchetto H., Laran S., Monestiez P., Van Canneyt O., **Virgili A.**, Ridoux V. (2017). Using large scale survey to investigate seasonal variations in seabird distribution and abundance. Part I: the North Western Mediterranean sea. *Deep Sea Research Part II: Topical Studies in Oceanography* 141: 74-85.
- Delavenne, J., Lepareur, F., Witté, I., Touroult, J., Lambert, C., Pettex, E., **Virgili, A.**, Siblet, J-P. (2017). Spatial conservation prioritization for mobile top predators in French waters: comparing observation rates and predicted densities as input. *Deep Sea Research Part II: Topical Studies in Oceanography* 141: 275-284.

## CONFERENCE PAPERS

### Poster presentation

- Virgili A.**, Authier, M., Boisseau, O., Cañadas, A., Claridge, D., Cole, T., Corkeron, P., Dorémus, G., David, L., Di-Méglio, N., Dunn, C., Dunn, T.E., García Barón, I., Laran, S., Lewis, M., Louzao, M., Mannocci, L., Martínez-Cedeira, J., Palka, D., Panigada, S., Pettex, E., Roberts, J., Ruiz Sancho, L., Santos, M.B., Van Canneyt, O., Vázquez Bonales, J.A., Monestiez, P., Ridoux, V. (2017). Basin wide approach, combined datasets and gap analyses: options to overcome the lack of sighting data on rare cetacean species. In 31<sup>th</sup> European Cetacean Society Conference, Middlefart, Denmark.

## Oral presentations

**Virgili A.**, Authier, M., Boisseau, O., Cañadas, A., Claridge, D., Cole, T., Corkeron, P., Dorémus, G., David, L., Di-Méglio, N., Dunn, C., Dunn, T.E., García Barón, I., Laran, S., Lewis, M., Louzao, M., Mannocci, L., Martínez-Cedeira, J., Palka, D., Panigada, S., Pettex, E., Roberts, J., Ruiz Sancho, L., Santos, M.B., Van Canneyt, O., Vázquez Bonales, J.A., Monestiez, P., Ridoux, V. (2017). Large scale distribution of deep-diving cetaceans in the North Atlantic Ocean and the Mediterranean Sea. In 22<sup>nd</sup> Society for Marine Mammal Conference, Halifax, Canada.

**Virgili A.**, Authier M., Gérard O., Monestiez P., Ridoux V. (2016). Data degradation: An approach to simulate rare species distribution models. In 30<sup>th</sup> European Cetacean Society Conference, Funchal, Madeira.

**Virgili A.**, Lambert C., Pettex E., Ridoux V. (2015). Habitat modelling predictions: a tool for assessing the relevance of MPAs network. In World Seabird Conference II, Capetown, South Africa.

Pettex, E., Lambert, C., Laran, S., Ricart, A., **Virgili A.**, Falchetto, H., Authier, M., Monestiez, P., Van Canneyt, O., Dorémus, G., Blanck, A., Toison, V., Ridoux, V. (2014). Patrons saisonniers de la distribution des prédateurs supérieurs dans les eaux métropolitaines françaises - Apports pour la désignation des aires marines protégées. In Séminaire FREDD, La Rochelle.

## SCIENTIFIC REPORT

Pettex, E., Lambert, C., Laran, S., Ricart, A., **Virgili A.**, Falchetto, H., et al. (2014). Suivi Aérien de la Mégafaune Marine en France métropolitaine - Rapport Final. Technical Report, University of La Rochelle UMS 3462, 169 pp.

# ABBREVIATIONS

AIC	Akaike Information Criterion
ANN	Artificial Neural Network
AUC	Area Under the Curve
AUC <sub>mean</sub>	AUC averaged over the 100 experimental models
AUC <sub>ref</sub>	AUC of the reference model
CART	Classification And Regression Tree analysis
CITES	Convention on International Trade of Endangered Species
CMS	Conservation of Migratory Species of Wild Animals
CV	Coefficient of Variation
D*	Explained deviance
EKE	Eddy Kinetic Energy
EKE <sub>m</sub>	Mean of EKE
ENFA	Ecological Niche Factor Analysis
ESW	Effective Strip Width
GAM	Generalised Additive Model
GARP	Genetic Algorithm for Rule-set Production
GCV	Generalised Cross Validation score
GENERAL	Dataset of the whole study area
GLM	Generalised Linear Model
IUCN	International Union for Conservation of Nature
MaxEnt	Maximum Entropy
MED	Mediterranean Sea
MSE	Mean Squared Error
MSE <sub>mean</sub>	MSE averaged over the 100 experimental models
MSE <sub>ref</sub>	MSE of the reference model
N-ATL	Northeast and Northwest Atlantic Ocean
NB-GAM	GAM with a negative binomial distribution
NE-ATL	Northeast Atlantic Ocean
NPP	Net Primary Production
NW-ATL	Northwest Atlantic Ocean
R <sup>2</sup>	Coefficient of determination

RMSE	Root Mean Squared Error
PO-GAM	GAM with a Poisson distribution
RES	Relative Environmental Suitability
Sd	Standard error
SDM	Species Distribution Model
SSH	Sea Surface Height
SSHm	Mean of SSH
SST	Sea Surface Temperature
SSTgrad	Gradients of SST
SSTm	Mean of SST
TSS	True Skill Statistics
TW-GAM	GAM with a Tweedie distribution
ZIP-GAM	GAM with a zero-inflated Poisson distribution
ZIP-GLM	GLM with a zero-inflated Poisson distribution

# CONTENTS

---

REMERCIEMENTS .....	V
SCIENTIFIC PRODUCTIONS .....	VIII
ABBREVIATIONS .....	X
CONTENTS .....	XII
<b>1 GENERAL INTRODUCTION .....</b>	<b>1</b>
1.1 Rare species .....	2
1.2 How to model species distribution? .....	4
1.3 Objectives and outline of the thesis .....	8
<b>2 GENERAL METHODOLOGY: STUDY AREAS, SPECIES OF INTEREST &amp; STATISTICAL MODELS .....</b>	<b>13</b>
2.1 The North Atlantic Ocean and the Mediterranean Sea: two distinct but connected oceanographic regions .....	14
2.2 The deep-diving cetaceans .....	18
2.3 Presence-absence and count-based models versus Presence-only models .....	24
<b>3 RARE SPECIES: HOW TO MODEL THEIR DISTRIBUTION? .....</b>	<b>33</b>
3.1 Context and objectives .....	34
3.2 Methodology .....	35
3.3 Stage 1: Comparison of models for scarcely detected species .....	42
3.4 Stage 2: How many sightings to model rare species distributions? .....	46
3.5 General considerations .....	52
3.6 Predicting habitats of rare species .....	54
3.7 Recommendations for practitioners .....	56
<b>4 DEEP-DIVER HABITAT PREFERENCES .....</b>	<b>57</b>
4.1 Context and objectives .....	58
4.2 Methodology .....	58
4.3 Which distribution for the deep-divers? .....	64
4.4 A basin wide approach to model the distribution of rare marine species .....	69
<b>5 DATA-ASSEMBLING: A MATTER OF ECOSYSTEMS SIMILARITY .....</b>	<b>73</b>
5.1 Context and objectives .....	74
5.2 Methodology .....	75
5.3 Can predictions be extrapolated in different ecosystems? .....	78
5.4 Can different ecosystems be assembled? .....	83

5.5	What scale to consider for data-assembling? .....	84
<b>6</b>	<b>GENERAL DISCUSSION.....</b>	<b>85</b>
6.1	Overview of the thesis dissertation.....	86
6.2	Habitat preferences of deep-diving cetaceans .....	89
6.3	Predicting distributions of rare species: a particular framework.....	94
6.4	Perspectives .....	98
	<b>REFERENCES .....</b>	<b>101</b>
	<b>ANNEXES.....</b>	<b>121</b>
	ANNEX A: COMPARISON OF HABITAT MODELS FOR SCARCELY DETECTED SPECIES .....	123
	ANNEX B: HOW MANY SIGHTINGS TO MODEL RARE MARINE SPECIES DISTRIBUTIONS .....	153
	ANNEX C: COMBINING VISUAL SURVEYS TO MODEL HABITAT OF DEEP-DIVING CETACEANS AT THE BASIN SCALE .....	187
	ANNEX D: DATA-ASSEMBLING: A MATTER OF ECOSYSTEMS SIMILARITY – SUPPORTING INFORMATION.....	217
	ANNEX E: WOULD MODELS BE IMPROVED IF PREY DISTRIBUTIONS WERE INCLUDED?.....	221



# Chapter 1

---

## GENERAL INTRODUCTION

---



© Laura Hedon

## CONTENTS

---

1.1 RARE SPECIES .....	2
1.1.1 What is a rare species? .....	2
1.1.2 The role of rare species in ecosystem functioning .....	3
1.1.3 Rare species and Conservation .....	3
1.2 HOW TO MODEL SPECIES DISTRIBUTION? .....	4
1.2.1 Concepts .....	4
1.2.2 Temporal and spatial scales in ecology .....	6
1.2.3 Habitat modelling for rare species .....	7
1.3 OBJECTIVES AND OUTLINE OF THE THESIS .....	8
1.3.1 Context and objectives .....	8
1.3.2 Outline of the dissertation .....	9

**T**HE introduction presents the general framework of the thesis. Particularly, it aims to describe what is a rare species and what are the management implications for these species. A tool to help species conservation is the use of species distribution models and particularly habitat models. Consequently, I present the concept of habitat modelling and the issues related to the use of habitat models with rare species. Finally, I introduce the context and the objectives of the thesis work and outline the thesis dissertation.



## 1.1 RARE SPECIES

### 1.1.1 What is a rare species?

The rarity of a species is commonly characterised by its distribution and abundance relative to the distribution and abundance of taxonomically or ecologically comparable taxa (Reveal 1981; Gaston 1994). A quantile of the frequency distribution of abundance or geographic range size is commonly used to identify rare species. Gaston (1994) advised to define as rare 25% of species with the lowest abundance estimates. However, this definition is disputed and arbitrary. Reynoldson et al. (1997) suggested to include species distributions expected in pristine ecosystems (with limited disturbances) as reference conditions. In addition, this definition does not take into account the geographic scale of interest, a species can be rare at the regional scale but locally abundant because of local appropriate conditions. For example, the harbour porpoise (*Phocoena phocoena*) is abundant in the eastern English Channel in winter but rare in the entire English Channel (Lambert et al. 2017a). Consequently, the commonly accepted definition of species rarity appears to be somewhat simplistic.

However, depending on this definition, categories of rarity are defined. A widely known classification is Rabinowitz's classification (1981). In this framework, the rarity of a species is described in seven different ways depending on a combination of criteria describing the extent of the geographic range, the specificity of the habitat and the abundance of the population (Table 1.1; Rabinowitz 1981). According to these criteria, only species that are widely distributed, live in diversified habitats and are locally abundant, are considered common. Other species are defined as rare because they show restricted range, specific habitat and low abundance, or any combination of these criteria. In the marine environment, many species are considered as rare according to these criteria. For example, the leatherback sea turtle (*Dermochelys coriacea*) is rare because it shows a scarce population and a wide distribution in the open sea (Eckert 2002). The Galápagos penguin (*Spheniscus mendiculus*) shows another type of rarity with its population estimated between 4,000 and 8,000 individuals solely concentrated in the colder and nutrient-rich waters next to the Galápagos Islands (Boersma 1998).

**Table 1.1. The three characteristics that defined species rarity: habitat specificity, abundance and geographic range (from Rabinowitz 1981).** Each cell defines a form of species rarity except for the top left cell, which characterises a common species.

		Habitat specificity			
		Non-specialist		Specialist	
Abundance	High	Common species	Abundant but localised population in several habitats	Abundant and widespread population in specific habitats	Abundant and localised population in specific habitats
	Low	Scarce and widespread population in several habitats	Scarce and localised population in several habitats	Scarce and widespread population in specific habitats	Scarce and localised population in specific habitats
		Large	Limited	Large	Limited
		Geographic range			

The causes of species rarity are twofold, natural and anthropogenic (Pärtel et al. 2005; Flather and Sieg 2007). Natural causes are either inherent to the species and refer to its life history traits such as low growth rates, long generation time, low reproduction rates, high specialisation or higher trophic level (McKinney 1997) or inherent to the species habitats which might have a low carrying capacity or a low availability (Pärtel et al. 2005). Anthropogenic causes relate to human activities and are multiple. They include habitat loss and degradation, introduction of nonindigenous species, chemical or noise pollution or ecosystem exploitation (Flather and Sieg 2007). At different scales, these causes can lead to an increase in species rarity and ultimately species loss.

### 1.1.2 The role of rare species in ecosystem functioning

Few studies address the role of rare species in the ecosystem functioning. Lyons et al. (2005) reviewed these studies in order to assess the role that rare species may play in the ecosystems. This role turned out to be multiple. Ecosystems and trophic webs can be altered by the loss of some rare species, particularly the loss of top predators (Purvis et al. 2000; Duffy 2003). For example, in the past decades, a removal of deer predators in the United States have caused an overpopulation of these large herbivores and a degradation of their habitat by a loss of plant diversity (Anderson et al. 2001). In addition, rare plant species can limit the invasion of new species and play an important role in the nutrient cycle and retention (Theodose et al. 1996; Lyons and Schwartz 2001). Furthermore, under perturbations, rare species may also increase ecosystem resilience compared to abundant species (Walker et al. 1999). For example, after a fire episode, some rare plants quickly colonise the environment before their competitors recolonise the environment, which stabilises soils and maintains vegetation cover (Menges and Kimmich 1996). Consequently, rare species can be essential in ecosystem functioning and their conservation appears to be a major issue.

### 1.1.3 Rare species and Conservation

Due to their small populations, rare species have a greater risk of extinction than common species (Johnson 1998; Matthies et al. 2004). Indeed, if animals are dispersed, their reproductive success might be limited and they have difficulties to cope with changes in environmental conditions (*e.g.* diseases) because of genetic simplification (Lande 1995). In addition, rare species may face many threats (*cf.* 1.1.1) and anthropogenic causes of rarity may induce more extinction risks than natural causes because species are not evolutionary adapted to these threats (Flather and Sieg 2007). Consequently, rare species are often priority species for management plans to maintain or restore their populations and habitats (Lawler et al. 2003).

Rare species need to be classified to help conservation scientists prioritising species at risk of extinction. For that, Rabinowitz's classification (1981) is useful but not sufficient because it does not include natural and anthropogenic threats. According to Master et al. (2000), population viability also needs to be considered. It depends on threats and landscape connectivity, population size and number, conditions of occurrence and trends in these factors (Master et al. 2000).

In this context, the IUCN (International Union for Conservation of Nature) Red List of Threatened Species is a powerful tool to list threatened species (IUCN 2001). It assesses extinction risk of a species by examining population size and its trends, geographic range, degradation of habitat quality, level of habitat exploitation and effects of introduced elements (*e.g.* pathogens, pollutants, parasites...; IUCN

2001). Seven levels of extinction risk are defined in the IUCN Red List: 'least concern', 'near threatened', 'vulnerable', 'endangered', 'critically endangered', 'extinct in the wild' and 'extinct' but the main problem is the lack of data available to establish these levels. Indeed, many of the rare species are defined as 'insufficiently known' or 'data deficient' because population sizes or trends or geographic range are often unknown (IUCN 2001). Consequently, a better understanding of their distribution, population sizes and trends is necessary to improve their conservation. In this context, habitat modelling is a useful tool to describe habitat uses, density patterns and distributions of these species (Redfern et al. 2006; Hegel et al. 2010).

## 1.2 HOW TO MODEL SPECIES DISTRIBUTION?

### 1.2.1 Concepts

The ecological niche concept is central in ecology. Hutchinson (1957) described a niche as a multidimensional hyper-volume defined by the environmental conditions in which a species population can persist. In this ecosystem (or habitat), species have interactions with the environment which is defined by biotic and abiotic factors (Hutchinson 1957). Abiotic factors refer to non-living physical and chemical elements in the ecosystem (*e.g.* air, soil, sunlight) while biotic factors are living organisms in the ecosystem (*e.g.* animals, plants, fungi). The more an environment is able to produce suitable conditions for the survival, reproduction and persistence of a population, the better the quality of the habitat (Block and Brennan 1993). Based on the distribution of a species, the ecological niche concept allows to identify ecological drivers of species habitat preferences, *i.e.* higher densities are observed in the most suitable habitats.

Species distribution models (SDMs) are strongly based on the niche theory (Elith and Leathwick 2009; Franklin 2010). Elith and Leathwick (2009) defined a SDM as a "model that relates species distribution data (occurrence or abundance at known locations) with information on the environmental and/or spatial characteristics of those locations". Through statistical models, empirical correlations between species distribution and biotic or abiotic variables are described. In addition, by analysing the conditions where a species is sighted (*versus* non-sighted), SDMs estimate the similarity of the conditions at any site to these conditions and predict the potential geographic distribution of the species (Franklin 2010). Hence, the use of SDMs is based on two assumptions: environmental variables are the primary determinants of species distributions and species have reached an equilibrium with these variables (Guisan and Zimmermann 2000).

SDMs are based on three types of data (Anderson 2012). Presence-only data consist of a sample of locations where the species is observed without information about the sites where the species is absent; presence-absence data consist of a sample of locations where the species is observed or non-observed; count data consist of a sample of locations where the species is observed and the species abundance is recorded. Presence-only data are either tagging data in which the animal position is recorded (Edrén et al. 2010) or often recorded in non-dedicated surveys. These surveys can be platforms of opportunity (ferries, whale watching, fishing vessels; Zador et al. 2008; Cotté et al. 2010) or commercial whaling records (Kashner et al. 2006) in which survey design and effort are not known and observation reliability is variable (Redfern et al. 2006; Moura et al. 2012). Hence, the relative density of individuals cannot be inferred. Presence-absence and count data are recorded in dedicated surveys such as visual and

acoustic surveys in which the survey design is controlled and the presence of the animal is visually or acoustically detected (Redfern et al. 2006; Pirodda et al. 2011; Buckland et al. 2015).

Depending on the available data, presence-only models, that use presence-only data, presence-absence models, that use presence-absence data and count-based models, that use count data, are built (Guisan and Zimmermann 2000). These three types of models include different techniques that are described in Chapter 2 but presence-only models are generally used to predict habitat suitability of the species (*i.e.* the capacity of a given habitat to support a selected species; Thorne et al. 2012) while presence-absence and count-based models are used to infer probability of occurrence (probability that a species is present in a habitat; Panigada et al. 2008) or relative density (Lambert et al. 2017a) of the species.

Variables used to describe the relationships between a species observation and its environment can be divided into proximal and distal variables. Proximal variables are biological variables to which the species is assumed to react more directly than distal variables that are physical variables which describe the environment (Guisan and Zimmermann 2000). In the marine environment, the surface chlorophyll concentration is commonly used as proxy for the biomass of primary producers (Jaquet et al. 1996; Ferguson et al. 2006; Lambert et al. 2017b) but the distribution of low- (phytoplankton and zooplankton) and mid-trophic levels (micronekton) would be better predictors to describe top predator distributions (Redfern et al. 2006). Surface chlorophyll concentration data are commonly used because these satellite data are available at a global scale. In contrast, prey distribution data are not sampled at a large temporal and spatial scales and thus are unavailable at the spatial extent and resolution needed for modelling. To overcome these limitations, the use of output data provided by ecosystem models, which simulate the biomass and production of low- to mid-trophic levels at large spatial scale is emerging (Lehodey et al. 2010; Abecassis et al. 2013; Lambert et al. 2014). Distal variables encompass two types of variables; physiographic variables which are static descriptors that relate to the bathymetry (*e.g.* depth, slope) and oceanographic variables which are dynamic descriptors that describe the water masses (*e.g.* sea surface temperature, sea surface height, eddy kinetic energy, winds, surface chlorophyll). These variables are more largely available than proximal data (from satellite or numerical models) and are often used to describe marine top predator distributions (Yen et al. 2004; Redfern et al. 2006; Lambert et al. 2017a).

Habitat models are categories of SDMs. They aim to explain relationships established between the species distribution and its environment but also to predict its distribution. As habitat modellers, we also aim to predict species distribution in non-surveyed areas, either because a study area cannot be uniformly sampled or because some areas cannot be sampled; predicting in unsurveyed area is a geographical extrapolation (Elith and Leathwick 2009; Franklin 2010). However, predictions cannot be extrapolated beyond the range of sampled environmental conditions (environmental extrapolation) and models are considered not transferable from one region to another (Randin et al. 2006; Redfern et al. 2017), except between similar ecoregions (Vanreusel et al. 2007). Consequently, precautions have to be made when extrapolating predictions beyond surveyed areas, we can geographically extrapolate predictions only in environmental interpolation areas. Consequently, the scale of the study may have an impact on these predictions.

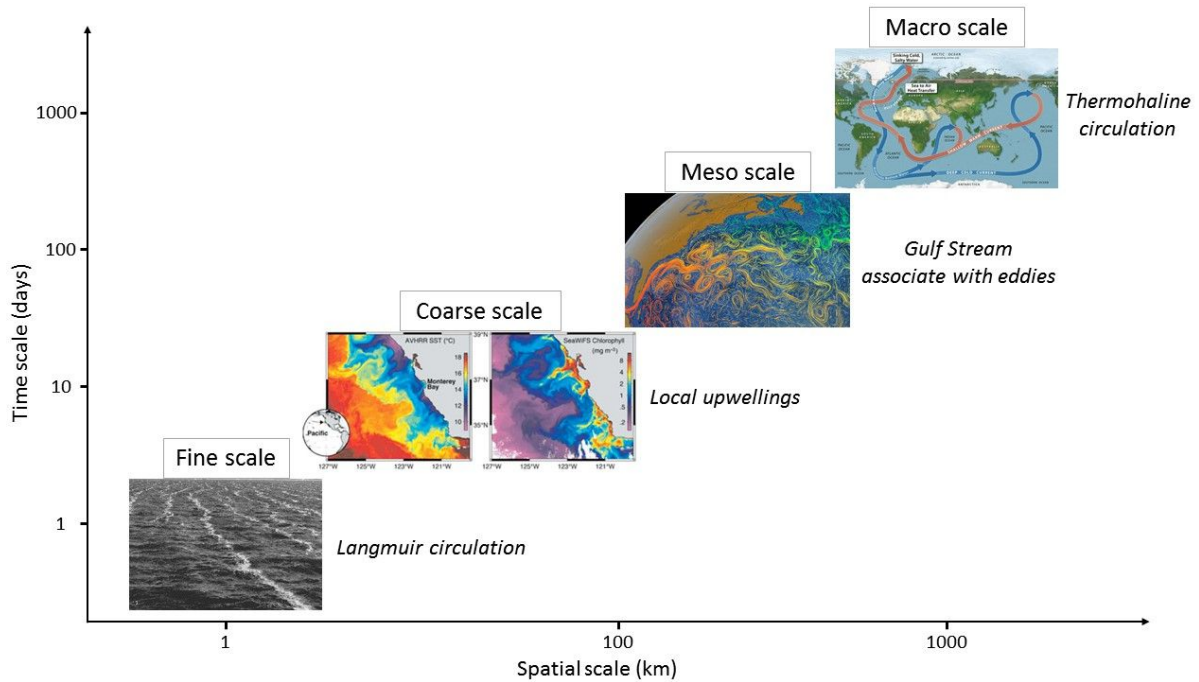
### 1.2.2 Temporal and spatial scales in ecology

The scale plays a central role in ecology. As Levin (1992) wrote, “there is no single natural scale at which ecological phenomena should be studied; systems generally show characteristic variability on a range of spatial, temporal, and organizational scales”. However, the word ‘scale’ has multiple meanings. In ecology, scale mostly refers to the extent (*i.e.* the spatial domain over which the system is studied and for which data are available) relative to the grain of a variable (*i.e.* the minimum spatial resolution of the data or the size of the individual units of observation) indexed by time or space (Schneider 1994; Wu and Li 2006). Depending on the perspective, the grain and the extent may differ. For an organism, the grain would be the finest component of the environment that would be differentiated in a range which defined the extent. For a management objective, the grain would be the smallest unit of management while the extent would be the total area under management consideration.

Wu and Li (2006) identified three dimensions of scale: space, time and integrative levels, the latter being a conceptual construction from the observer. Spatial and temporal scales are closely linked and small events are faster and more frequent than large events which have lower frequencies and a longer longevity. For example, in the marine environment, winds induce at fine scale local and fleeting vortices that concentrate phytoplankton (Evans and Taylor 1980) while at a meso scale, large and long-lived eddies bring up nutrients from deep waters and enhance primary production (Oschlies and Garçon 1998). Both in marine or terrestrial environments, it has been shown that temporal and spatial scales of physical or biological processes drive species distributions and movements at multiple scales (Naugle et al. 1999; Apps et al. 2001; Pinaud and Weimerskirch 2005; Cotté et al. 2009). At a fine scale, top predator distributions are generally driven by the presence of their prey (Fauchald et al. 2000; Goldbogen et al. 2008; Elmhagen et al. 2010) while at a broader scale, species generally respond to more persistent processes such as the presence of watering-places or current systems (Jaquet et al. 1996; Atwood et al. 2011; Reygondeau et al. 2012).

In the marine environment, which defines the framework of this study, Hunt and Schneider (1987) have identified four spatial scales in which hydrodynamic processes occurred (Fig. 1.1): the fine scale (1 m-1 km) in which vortices and Langmuir circulation occur; the coarse scale (1-100 km) in which upwellings and oceanic frontal zones occur; the meso scale (100-1000 km) in which rings, eddies, jets occur and the macro scale (>1000 km) in which surface currents (*e.g.* equatorial circulation in the central Pacific, the central gyres of the North Atlantic and South Pacific) and the global ocean circulation occur. Considering scale is consequently crucial. The interactions between the species and its environment, the involved ecological processes and the statistical relationships may vary depending on the scale (Wu et al. 2002; Mannocci et al. 2017a). Marine top predators are generally linked to persistent oceanographic features described at coarse to macro scales. Indeed, to optimise their foraging success, they would select favourable habitats associated with persistent features based on their experience (Davoren 2003) such as shelf edges or frontal zones because productivity is on average locally higher (Balance et al. 2006; Weimerskirch 2007). For example, from tagging data, Chilvers (2008) showed year-to-year foraging site fidelity of New Zealand sea lions (*Phocarctos hookeri*). In addition, Scott et al. (2002) showed that climatic regulators generally control the large scale species distribution while fine scale patchy species distribution result from patchy distribution of the resource. In this context, Mannocci et al. (2017a) recommended to define the resolution of the used variables so as to match the ecological question of the study. For behavioural processes, instantaneous variables should be preferentially used while when seasonal or annual processes are dominant, contemporaneous (daily or monthly) or

climatological (multi years) variables should be used. Considering that, the use of persistent oceanographic features at meso or macroscale in habitat modelling of top predators is more recommended than the use of ephemeral features at fine scale to determine long-term distribution patterns.



**Fig. 1.1. Hierarchical structure of the marine environment.** Each scale is defined by a time and spatial scale. The larger the event, the greater its time span. An example of oceanographic feature is provided for each scale. At a fine scale, the Langmuir circulation consists of a series of surface vortices induced by the wind (source image: <http://homepages.cae.wisc.edu>). At a coarse scale, upwellings are induced by the action of the winds which cause rise of cold waters and nutrients (source image: <http://www.clivar.org>). At a meso scale, the Gulf Stream is a northward accelerating hot current flowing off the east coast of North America and generating multiple eddies (source photo: NASA/Goddard Space Flight Centre Scientific Visualization Studio). At a macro scale, the thermohaline circulation is driven by global density gradients created by surface heat and freshwater fluxes (source image: <http://www.emse.fr/>).

### 1.2.3 Habitat modelling for rare species

As developed in 1.1.2, rare species, particularly species threatened by human activities, generally correspond to high priority conservation challenges. To describe their distribution and habitat uses through habitat models (similarly for common species), the highest data quality (*e.g.* sighting and effort data) is required (Redfern et al. 2006). Indeed, sighting data with associated effort data (*i.e.* the ship/aircraft GPS tracks which allows to attribute an absence of the species when the species is not observed) are needed to estimate the relative density of individuals in a study area and identify density hotspots (Redfern et al. 2006), which can help delineating marine protected areas (Cañadas et al. 2005). However, rare species usually result in a low number of sightings per unit effort (Cunningham and Lindenmayer 2005). This scarcity of sighting data makes it difficult to fit habitat models because the reliability of the predictions largely depends on the number of sightings on which the models are fitted (Welsh et al. 1996; Barry and Welsh 2002; Cunningham and Lindenmayer 2005).

Some studies have addressed the use of habitat models for rare species datasets. Welsh et al. (1996) and Cunningham and Lindenmayer (2005) showed that count data associated with extra zeros should be treated in two steps, firstly by only considering the species presence pattern and then,

conditionally on the presence, by modelling the number of individuals. In addition, Welsh et al. (1996) added that the number of individuals should not be considered as a Poisson distribution but as a truncated Poisson distribution. Due to the lack of absence data, Engler et al. (2004) simulated species pseudo-absences and suggested that the quality of prediction maps provided by generalised linear models which used pseudo-absence data was higher than predictions maps provided by a presence-only model (an ecological niche factor analysis). However, the reliability of the predictions produced by these various models and the uncertainty associated with these predictions remain pending issues.

To provide answers to these issues, one option would be to test if the performance of a species distribution model is maintained when the amount of input data decreases, which would assess the reliability of the models using small datasets of rare species. Following Winiarski et al. (2014), a second option, would be to merge different datasets collected by multiple surveys in order to increase the number of data implemented in the habitat models. Indeed, taken separately, surveys cannot in some cases provide a sufficient number of rare species data to model their habitat (Waring et al. 2001; Barlow et al. 2006; Kiska et al. 2007). For the rarest cetacean species, the number of sightings rarely exceeds a few tens, which is generally insufficient to estimate relationships with the environment by fitting habitat models (Stockwell and Peterson 2002). Consequently, over the past few years, data-assembling has been increasingly used for the study of cetacean distribution (Paxton et al. 2016; Roberts et al. 2016; Mannocci et al. 2017b; Rogan et al. 2017). This consists in merging datasets collected in different surveys, that follow or not the same observation protocols, to increase the amount of data available for the models. For example, by merging four surveys, Rogan et al. (2017) were able to predict abundances of sperm whales and beaked whales in the northeast Atlantic Ocean. Similarly Roberts et al. (2016) predicted habitats of multiple cetacean species in the northwest Atlantic Ocean by assembling 23 years of aerial and shipboard surveys. This data-assembling is fairly easy when the protocols used in the surveys are similar, *i.e.* the same parameters are recorded (observation conditions, number of individuals, sighting distance) or the platforms are identical (*e.g.* same observation heights, same speed). In contrast, when surveys do not use the same protocols or platforms, assembling datasets implies that the species detection capacity and data quality are taken into account for each survey. In addition, each survey may not collect the same ancillary data, particularly regarding observation conditions; some surveys record only Beaufort seastate while other surveys also record other parameters of potential importance for the species detection, such as sun glare, cloud coverage or wave height. Consequently, the homogenisation of these different data may require levelling to the coarsest commonalities across datasets, which lead to some level of data degradation. However, data-assembling can be an effective solution to overcome limitations associated with the study of rare species.

## 1.3 OBJECTIVES AND OUTLINE OF THE THESIS

### 1.3.1 Context and objectives

In the marine environment, most cetaceans can be considered as rare according to Rabinowitz's criteria (Rabinowitz 1981). Among them, the deep-diving cetaceans, defined here as beaked whales (family *Ziphiidae*; *e.g.* *Ziphius*, *Hyperoodon* spp. and *Mesoplodon* spp.) and sperm whales (families *Physeteridae* and *Kogiidae*), are a good example of rare species. They are oceanic species distributed worldwide and frequently associated with steep slope habitats where they feed in deep waters during

long dives (often over an hour; Perrin et al. 2009). Due to their offshore habitat, short time at the surface and therefore low availability to sightings, little is known about their population trends and densities within their distributional range (especially for kogiids and ziphiids). These species are mostly listed as 'vulnerable' and 'data deficient' by the IUCN list (Taylor et al. 2008a, b; 2012a, b).

Deep-divers are threatened by a variety of anthropogenic activities. Historically, while most beaked whales and kogiids have not been hunted, sperm whales have been massively harvested from the early nineteenth century onward (Whitehead 2003). Today, deep-divers are impacted by bycatch and entanglement, debris ingestion and ship collisions (Carrillo and Ritter 2010; Madsen et al. 2014; Unger et al. 2016). The best known threat relates to human activities that produce high intensity acoustic signals (e.g. military sonars, seismic guns or techniques used in large maritime construction projects; Stone and Tasker 2006). Recent studies have demonstrated the sensitivity of deep-diving cetaceans, and particularly beaked whales, to underwater noise pollution. Certain sounds can cause death or a diversity of sub-lethal traumas to these whales and several unusual stranding events have occurred in connection with the use of military sonars (Frantzis 1998; Balcomb and Claridge 2001; Brownell et al. 2004; Fernández et al. 2005; D'Amico et al. 2009).

In France, deep-divers are protected by a ministerial order (*Arrêté du 1<sup>er</sup> juillet 2011*). The French Navy and the Ministry of Defense (*Direction Générale de l'Armement - DGA*) aim to limit the impact of military activities, in particular training operations, on these species. To do that, a better understanding of their distribution is needed. Indeed, to mitigate the impact of anthropogenic activities, good knowledge of the distribution and density hotspots of deep-diving cetaceans is crucial for Marine Spatial Planning and to inform management measures (Douve 2008).

In order to meet the conservation objectives for deep-divers, considering that these species are rare, the objectives of the thesis were: (i) to find a suitable method of analysis adapted to datasets containing a large number of zeros; (ii) once this method was adopted, to determine how environmental and biological variables influence the distribution and habitat selection of deep-diving cetaceans in order to (iii) predict the potential areas used by these species in the North Atlantic Ocean and the Mediterranean Sea and estimate the uncertainties associated with these predictions.

### 1.3.2 Outline of the dissertation

Besides its ecological and conservation oriented objectives, this thesis has a strong methodological content. Indeed, I aim to propose a methodological approach for the study of rare species habitats. Even if this thesis is focused on the marine environment, some analytical strategies developed here could be used in terrestrial environments as well. The thesis is divided into six chapters. Each chapter is based on the results of the previous one to develop a methodological approach that help modelling the habitat preferences of rare species (Fig. 1.2).

**Chapter 1** states the problems associated with the study of rare species but also the issues related to their conservation, with a focus on deep-diving cetaceans.

**Chapter 2** describes the ecology of the species of interest to the study (beaked whales, sperm whales and kogiids) but also the study area, namely the North Atlantic Ocean and the Mediterranean Sea, and the habitat model techniques commonly used in marine predator ecology.

Based on the different available statistical models, **Chapter 3** aims to find a model that would be suitable for the study of rare species by comparing different models applied to small datasets. This



chapter is based on two papers (**Annexes A and B**), one being published (Virgili et al. 2017a) while the other is in revision at the time of printing the present PhD report (Virgili et al. in revision).

The model selected in Chapter 3 was then used in **Chapter 4** to model the distribution of deep-diving cetaceans. However, even for a model adapted to a large number of zeros, the number of deep-diver data available for the analysis is generally insufficient in each survey when considered separately. Thus, data from different surveys were assembled to model the large-scale distributions of the three taxa of interest. This chapter is based on an article that is currently in preparation and is about to be submitted (**Annex C**, Virgili et al. in prep.).

In Chapter 4, to implement the maximum number of available data in the models, I hypothesised that the habitat drivers of deep-divers were identical throughout the study area, but in reality, they may vary. Therefore, in **Chapter 5**, through a model transferability analysis, I highlight that habitat drivers vary between contrasted large ecosystems and consequently habitat models fitted at a geographic scale that encompasses several ecosystems show lower performance than when fitted across an ecologically more homogeneous area. This chapter is planned to be published as a separate stand-alone paper.

Finally, in **Chapter 6**, the strategy adopted during the thesis, its main outcomes and the potential effect of using more proximal variables in the models (*e.g.* prey distributions) are assessed. The challenges related to the statistical modelling of rare species habitats and the management implications of this thesis are finally discussed.

Five **annexes** are given after the six main chapters. The first three annexes are the articles currently published, in revision or to be submitted and each annex has its own appendices. Annex D is the supporting information of Chapter 5. Annex E describes a work in progress which attempts to explore how models would be improved if the distributions of deep-diver preys were included as explanatory variables. In the main body of the manuscript, references to the annexes are labelled from A to E (*e.g.* 'Annex A', for the first annex). References to the appendices are labelled as 'Appendix A1 of Annex A', for the first appendix of the first annex.

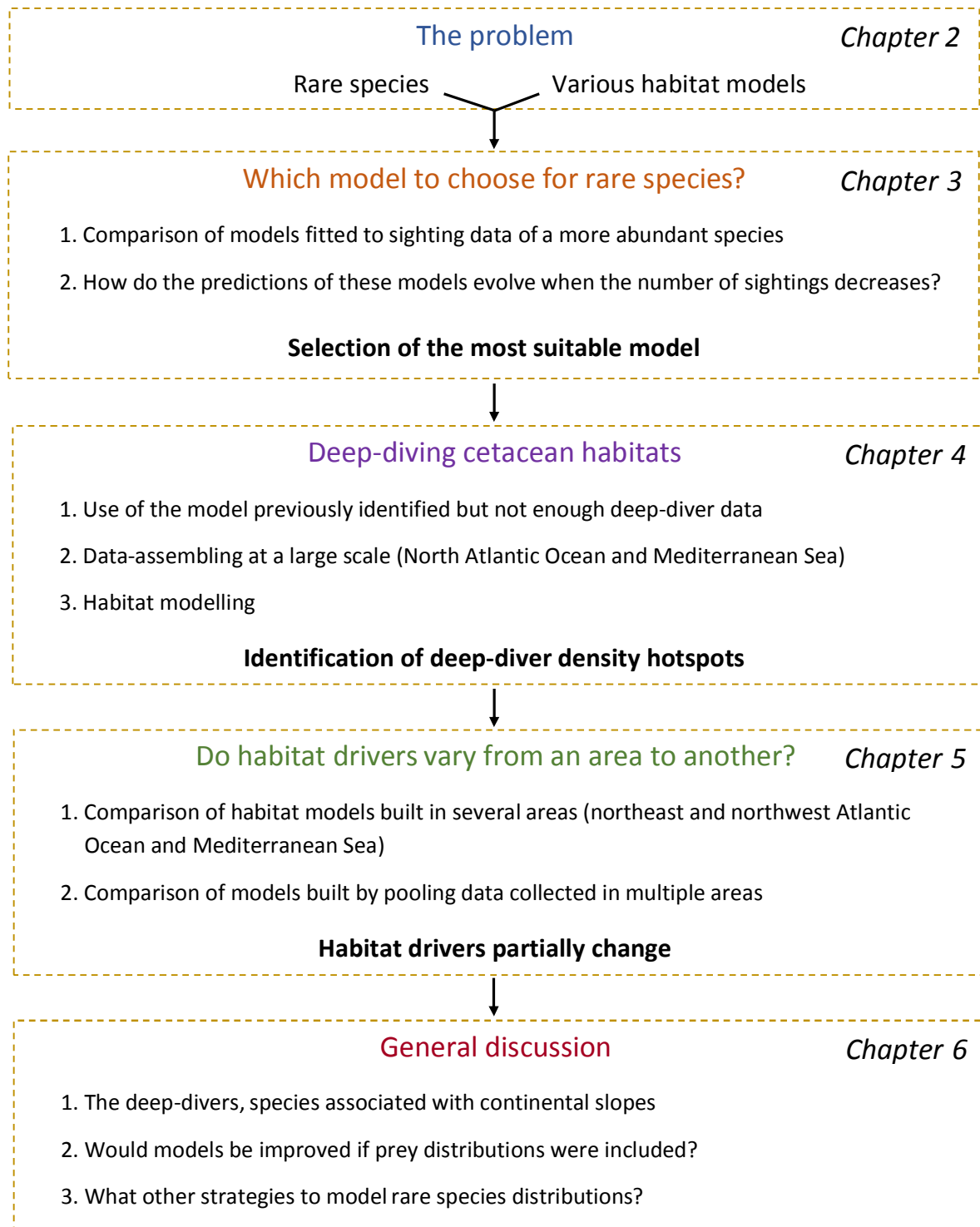


Fig. 1.2. Flowchart of the PhD strategy.



# Chapter 2

---

## GENERAL METHODOLOGY: STUDY AREAS, SPECIES OF INTEREST & STATISTICAL MODELS

---



© Laura Hedon

### CONTENTS

---

2.1 THE NORTH ATLANTIC OCEAN AND THE MEDITERRANEAN SEA: TWO DISTINCT BUT CONNECTED OCEANOGRAPHIC REGIONS .....	14
2.1.1 The North Atlantic Ocean .....	14
2.1.2 The Mediterranean Sea .....	16
2.2 THE DEEP-DIVING CETACEANS.....	18
2.2.1 The beaked whales .....	18
2.2.2 The sperm whales.....	21
2.2.3 The kogiids .....	23
2.3 PRESENCE-ABSENCE AND COUNT-BASED MODELS VERSUS PRESENCE-ONLY MODELS .....	24
2.3.1 Presence-only models.....	25
2.3.2 Presence-absence and count-based models.....	28

**T**HIS chapter aims to describe the oceanographic characteristics of the study area, the North Atlantic Ocean and the Mediterranean Sea but also the ecology of the studied species, the beaked whales, sperm whales and kogiids. In addition, an overview of the habitat modelling methods commonly used to infer marine top predator distribution is given in order to delineate the study framework.

## 2.1 THE NORTH ATLANTIC OCEAN AND THE MEDITERRANEAN SEA: TWO DISTINCT BUT CONNECTED OCEANOGRAPHIC REGIONS

The study area of the thesis encompassed two oceanographic regions: the North Atlantic Ocean from the Guiana Plateau to Iceland (approximately from 1-65°N), and the Mediterranean Sea, excluding the Baltic, Red and Black Seas, the Gulf of Mexico and Hudson Bay.

### 2.1.1 The North Atlantic Ocean

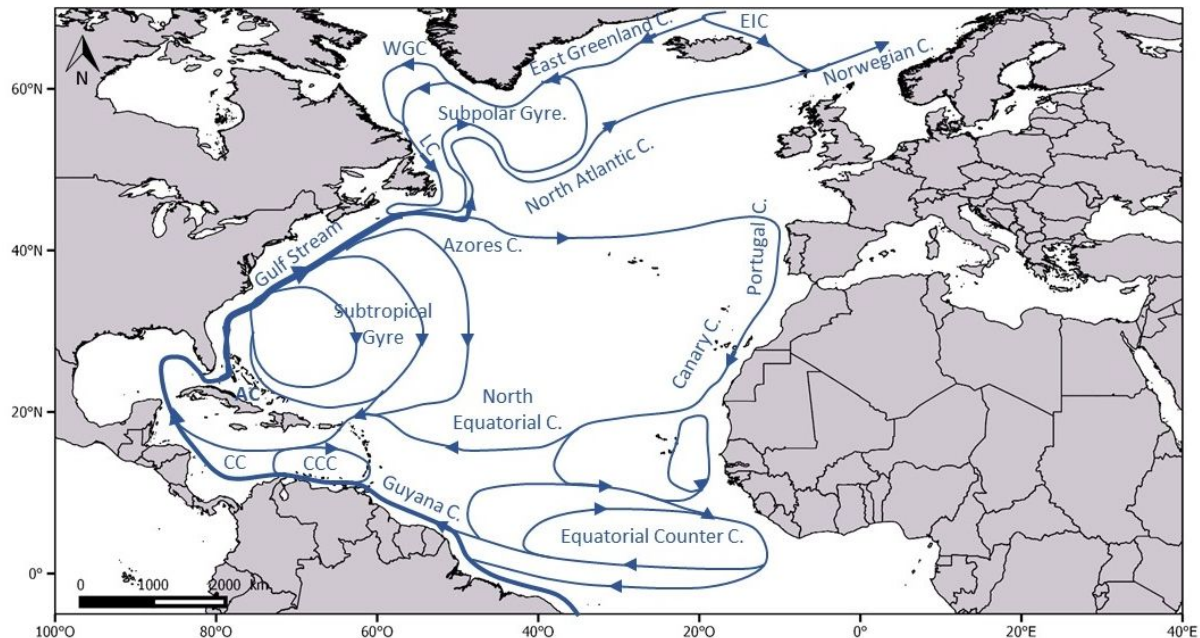
The North Atlantic Ocean, with a surface of 41,490,000 km<sup>2</sup> (Eakins and Sharman 2010), is divided longitudinally by the Mid-Atlantic Ridge into an eastern and western sub-basin (Levin and Gooday 2003). Its average depth is about 3,500 m but abyssal basins are deeper than 5,000 m and its maximum depth is 8,486 m in the Puerto Rico Trench (Eakins and Sharman 2010). The North Atlantic Ocean includes numerous submarine canyons particularly on the shelf break off the United States and western Europe (Levin and Gooday 2003) and fracture zones (*e.g.* the Romanche Fracture Zone near the Equator or the Gibbs Fracture Zone near 53°N; Fig. 2.1; Tomczak and Godfrey 2003).



Fig. 2.1. Topography of the North Atlantic Ocean. From Broadus et al. (2009).

The general circulation in the Atlantic Ocean is characterised by the presence of various currents (Fig. 2.2; Levin and Gooday 2003; Tomczak and Godfrey 2003). The North Equatorial Current, Antilles Current, Caribbean Current, Florida Current and Gulf Stream bring warm waters from the equatorial Atlantic Ocean and create a general northward movement of surface waters. The North Atlantic Current and Azores Current create an eastward transport of warm waters creating a gradient of temperatures from west to east (at around 40°N, water being approximately 8°C warmer in the west than in the east). The Azores Current loops gently into the Portugal Current and Canary Current that transport colder waters southwards. Trade Winds blowing almost parallel to the northwestern African coasts lead to cool

and nutrient-rich deep waters being upwelled to the surface. Cold waters from the north merge with the warm waters of the North Equatorial Current (at around 20°N), thus completing the North Atlantic Ocean circulation pattern (Schmitz and McCartney 1993, Levin and Gooday 2003; Tomczak and Godfrey 2003). Due to evaporation, precipitation and inflows from adjacent seas (*e.g.* Mediterranean Sea), salinity in the North Atlantic Ocean is higher than in other oceans (more than 35 ‰; Tomczak and Godfrey 2003).



**Fig. 2.2. Surface currents of the Atlantic Ocean.** C: Current; EIC: East Iceland Current; IC: Irminger Current; WGC: West Greenland Current; CC: Caribbean Current; AC: Antilles Current; CCC: Caribbean Counter Current. Adapted from Tomczak and Godfrey (2003).

Within the North Atlantic Ocean, primary production is quite low compared to other oceans, particularly in the tropical zone, but varies seasonally with maximum productivity in winter in the subtropical zone, a spring bloom and summer oligotrophic conditions at mid-latitudes and maximum productivity in summer in the subpolar zone (Fig. 2.3; Campbell and Aarup 1992). Subtropical regions are characterised by a low phytoplankton biomass and a low productivity, except in upwelling areas, with lower productivity in the west than in the east (Pérez et al. 2005). The northern part of the Atlantic Ocean is more productive. The very active vorticity of the Gulf Stream induces strong mesoscale variability by creating eddies characterised by strong horizontal gradients of temperature and causing local enhancement of either primary production or biomass accumulation (Fernandez and Pingree 1996; Longhurst 2007). In winter, permanent stratification in depth increases nutrient concentrations and allow phytoplankton growth with a bloom in spring. Numerous seamounts induce uplifting of isotherms, upwelling of nutrients, interaction with diel vertical migrant zooplankton and leading to locally high densities of pelagic fishes (Saltzman and Wishner 1997). The maximum biomass of zooplankton is observed in the epipelagic zone. Similarly, the highest diversity of cephalopods is observed in the epipelagic zone and along continental shelves but decreases with depth (Rosa et al. 2008). Fish diversity is lower than in other oceans with 589 species of pelagic fishes and 505 species of demersal fishes (Merrett 1994). Thirty-seven species of seabirds are seen regularly in the North Atlantic Ocean, 5 species of sea turtles and 32 species of whales are recorded (OPSAR 2000).

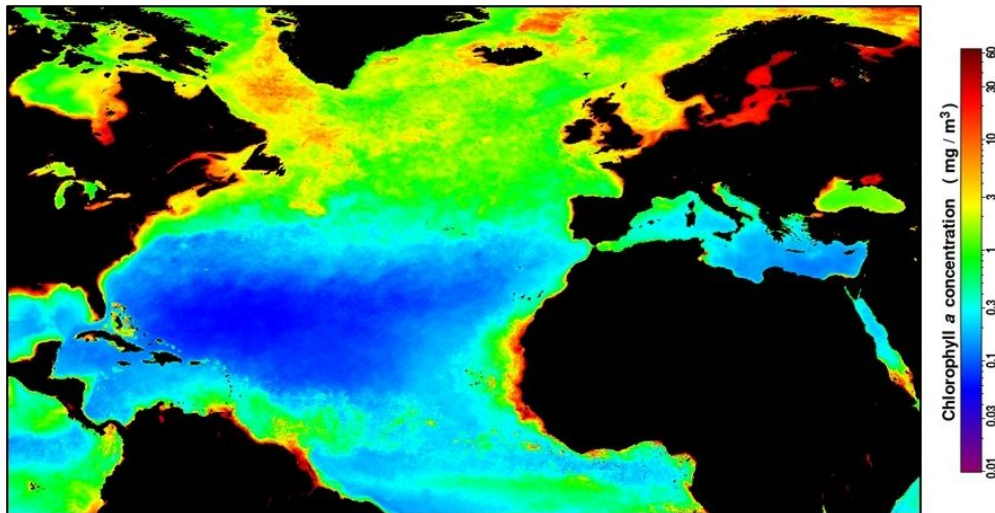


Fig. 2.3. Mean Chlorophyll *a* concentration in the North Atlantic Ocean in 2016 (Aqua MODIS satellite, <https://oceancolor.gsfc.nasa.gov>).

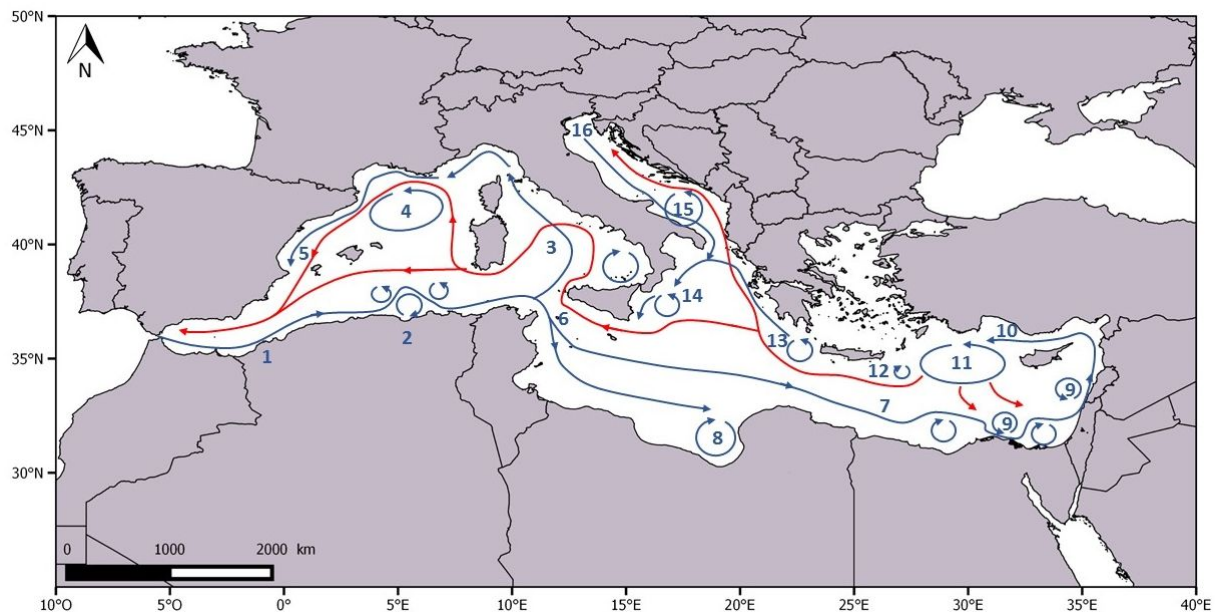
### 2.1.2 The Mediterranean Sea

The Mediterranean Sea is a semi-enclosed sea connected to the Atlantic Ocean through the Strait of Gibraltar, to the Black Sea through the Bosphorus Strait and, since 1869, to the Red Sea through the man-made Suez Canal (Fig. 2.4). A submarine ridge between Sicily and Tunisia divides the Mediterranean Sea into a western and eastern sub-basin, themselves divided into smaller basins (Fig. 2.4). The surface of the Mediterranean Sea is 2,967,000 km<sup>2</sup>, its mean depth is 1,500 m and its maximum depth is 5,139 m (Eakins and Sharman 2010). Narrow continental shelves (20% of total surface area), steep slopes, numerous submarine canyons and seamounts are characteristics of the Mediterranean Sea and particularly of its northern regions (Sarda et al. 2004).



Fig. 2.4. Topography of the Mediterranean Sea. From Salah and Boxer (2009).

Through the Strait of Gibraltar, constant inflows of low-density, comparatively less saline, Atlantic waters in the upper layer, and outflows of denser Mediterranean waters in the lower layer, generate anticyclonic gyres and eddies along the African coasts (Fig. 2.5; Pinardi and Masetti 2000; Sarda et al. 2004; Tanhua et al. 2013). Northward flows induce persistent cyclonic gyres in the northern part of the Mediterranean Sea, creating upwelling of nutrient-rich waters (Pinardi and Masetti 2000). Along the southern and eastern coasts of the basin, the Algerian Current, Atlantic Ionian Stream and Mid-Mediterranean Jet induce anticyclonic gyres (Pinardi and Masetti 2000). Due to an eastward gradual surface water evaporation, surface waters become saltier and warmer from west to east and sink to initiate the deep water circulation in the Levantine basin. A westward and northward water movement returns dense deep waters through the Tyrrhenian Sea and the Gulf of Lions to the Atlantic Ocean (Pinardi and Masetti 2000; Tanhua et al. 2013).



**Fig. 2.5. Main currents in the Mediterranean Sea.** Blue arrows represent the surface currents and red arrows represent the deep water circulation. 1: Inflows of the Atlantic Ocean; 2: Algerian Current and eddies; 3: Tyrrhenian cyclonic circulation; 4: Lions Gyre; 5: Ligurian-provencal Current; 6: Atlantic Ionian Stream; 7: Mid-Mediterranean Jet; 8: Anticyclone in the Gulf of Syrte; 9: Shikmona and Mers a-Matruh gyres; 10: Cilician and Asia Minor Current; 11: Rhodes Gyre; 12: Iera-Petra Gyre; 13: Pelops Gyre; 14: Western Ionian Gyre; 15: Southern Adriatic Gyre; 16: Western Adriatic Coastal Current. Adapted from Pinardi and Masetti (2000).

The Mediterranean Sea is an oligotrophic sea characterised by gradients of salinity and temperature from west to east, resulting in a gradual decrease in primary production from west to east (Fig.2.3; Bethoux et al. 1999; Longhurst 2007; Pujo-Pay et al. 2011). In addition, nutrient-rich deep waters are exported and nutrient-poor surface waters from the Atlantic Ocean are imported, causing low concentrations of nutrients in the basin except at the largest river plumes (*e.g.* Gulf of Lions, Nile Delta, northern Adriatic; Tanhua et al. 2013). Primary production shows seasonal cycles with winter and spring blooms of phytoplankton, while stratification limits primary production in the summer (Bosc et al. 2004; Longhurst 2007). Despite a low biological productivity, the Mediterranean Sea is characterised by a high biodiversity that decreases from west to east. Coll et al. (2010) estimated that about 17,000 marine species would live in the Mediterranean Sea, including 2,100 molluscs, 2,200 crustaceans, 650 fishes (of which 80 elasmobranchs), 5 sea turtles, 15 seabirds and 23 cetaceans.



## 2.2 THE DEEP-DIVING CETACEANS

Among all cetacean taxa recorded in the North Atlantic Ocean and the Mediterranean Sea, I focused on the beaked whales, sperm whales and kogiids. The species that belong to these groups are all deep-diving cetaceans, they are rare and threatened species with a cryptic behaviour at the surface and they forage on meso- to bathy-pelagic organisms.

### 2.2.1 The beaked whales

Beaked whales are toothed whales (order *Odontoceti*) that belong to the *Ziphiidae* family. Ziphiids contain 6 genera (*Berardius*, *Hyperoodon*, *Indopacetus*, *Mesoplodon*, *Tasmacetus* and *Ziphius*) divided into approximately 21 species (Table 2.1; Fig. 2.6; Mead 2009); this total is likely to be revised by current taxonomic studies.

**Table 2.1. Species of beaked whales belonging to the *Ziphiidae* family (6 genera and 21 species). Species reported from the North Atlantic Ocean are denoted by an asterisk.**

<b>Ziphiidae family</b>	
<b>Genus <i>Berardius</i></b>	
<i>Berardius arnuxii</i>	Arnoux's beaked whales
<i>Berardius bairdii</i>	Blaird's beaked whale
<b>Genus <i>Hyperoodon</i></b>	
<i>Hyperoodon ampullatus</i>	Northern bottlenose whale *
<i>Hyperoodon planifrons</i>	Southern bottlenose whale
<b>Genus <i>Indopacetus</i></b>	
<i>Indopacetus pacificus</i>	Longman's beaked whale
<b>Genus <i>Mesoplodon</i></b>	
<i>Mesoplodon bidens</i>	Sowerby's beaked whale * (Fig. 2.6)
<i>Mesoplodon bowdoini</i>	Andrew's beaked whale
<i>Mesoplodon carlhubbsi</i>	Hubbs' beaked whale
<i>Mesoplodon densirostris</i>	Blainville's beaked whale *
<i>Mesoplodon europaeus</i>	Gervais' beaked whale *
<i>Mesoplodon ginkgodens</i>	Ginkgotoothed beaked whale
<i>Mesoplodon grayi</i>	Gray's beaked whale
<i>Mesoplodon hectori</i>	Hector's beaked whale
<i>Mesoplodon layardi</i>	Straptoothed whale
<i>Mesoplodon mirus</i>	True's beaked whale *
<i>Mesoplodon perrini</i>	Perrin's beaked whale
<i>Mesoplodon peruvianus</i>	Peruvian beaked whale
<i>Mesoplodon stejnegeri</i>	Stejneger's beaked whale
<i>Mesoplodon traversii</i>	Spade-toothed whale
<b>Genus <i>Tasmacetus</i></b>	
<i>Tasmacetus shepherdi</i>	Shepherd's beaked whales
<b>Genus <i>Ziphius</i></b>	
<i>Ziphius cavirostris</i>	Cuviers' beaked whale *



Fig. 2.6. The Sowerby's beaked whale (*Mesoplodon bidens*). Illustration from Laura Hedon.

Beaked whales are medium-sized cetaceans with adult body lengths ranging from 3 to 13 m. They are characterised by a long rostrum (beak), a small triangular dorsal fin placed far back on the body, small pectoral fins, a tail fluke with no central notch and a generally (except for the *Berardius* and *Tasmacetus* genera) reduced dentition with a single pair of teeth for the males used for male-male interactions (in females and immatures males, these teeth are vestigial and usually do not erupt from the mandible; Mead 2009).

Beaked whales are pelagic species widely distributed in open oceans, except in the highest latitude Polar Regions. The Cuvier's beaked whale is the only species regularly recorded in the Mediterranean Sea (MacLeod et al. 2006). Considering the North Atlantic Ocean and the Mediterranean Sea as study area, the species recorded would be the northern bottlenose whale (*Hyperoodon ampullatus*), Sowerby's beaked whale (*Mesoplodon bidens*), Blainville's beaked whale (*Mesoplodon densirostris*), Gervais' beaked whale (*Mesoplodon europaeus*), True's beaked whale (*Mesoplodon mirus*) and Cuvier's beaked whale (*Ziphius cavirostris*; Fig. 2.7; MacLeod et al. 2006). Beaked whales are usually associated with steep slope and deep-water habitats characterised by the presence of submarine canyons and seamounts (Waring et al. 2001; Perrin et al. 2009; MacLeod et al. 2011; Whitehead 2013). They have a cryptic behaviour at the surface which make them difficult to sight, and for many species most available information is from stranded animals. Thus, it is difficult to determine the social structure. Group size ranges from 1-15 for Cuvier's beaked whales, 1-22 for northern bottlenose whales, 1-8 for Blainville's beaked whales and 1-15 for *Mesoplodon* spp. (MacLeod and D'Amico 2006).

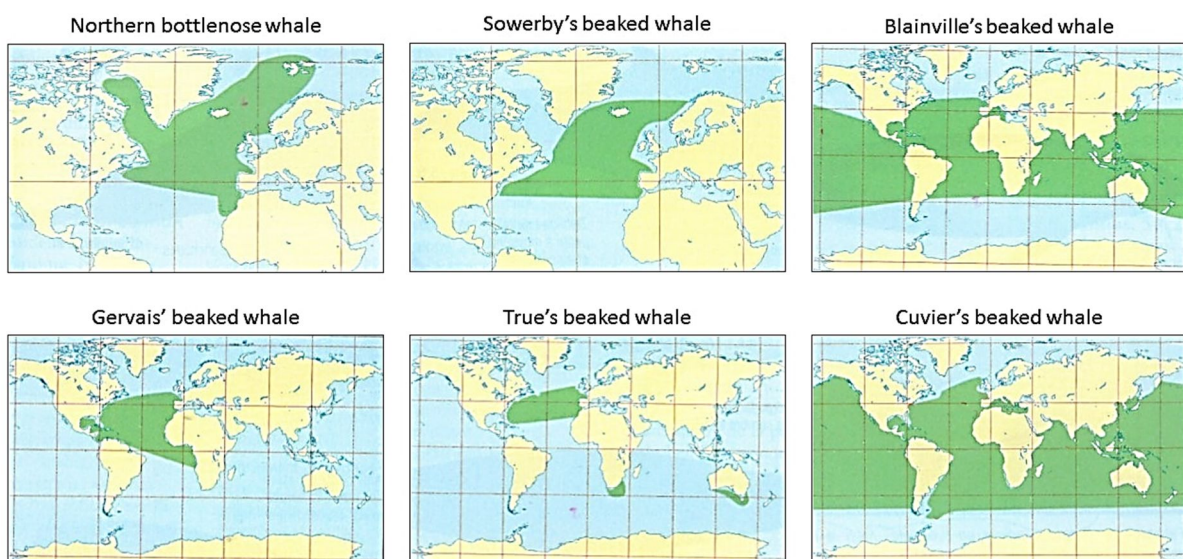
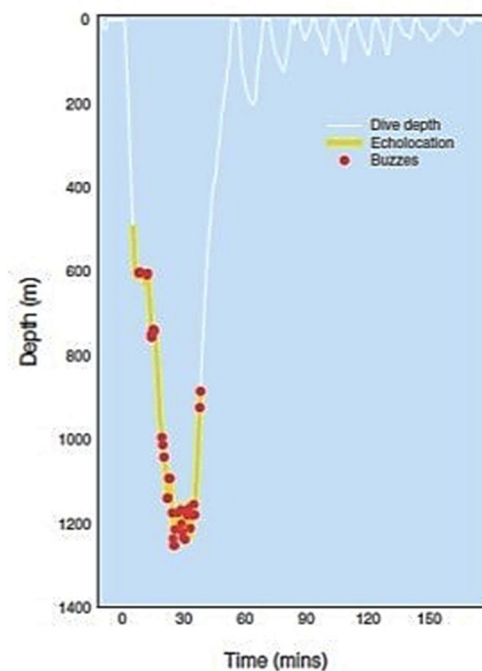


Fig. 2.7. Distribution areas of the beaked whales recorded in the North Atlantic Ocean and the Mediterranean Sea (other beaked whales are not present in the study area so they were not represented). Adapted from Shirihai (2006).

Similarly, limited information is available about their life history. Maximum recorded ages were between 27 and 60 years according to species (Mead 1984; MacLeod and D'Amico 2006). The age at sexual maturity is estimated between 7 and 15 years and the gestation is 17 months for Baird's beaked whale and 12 months for the northern bottlenose whale (Mead 1984).

Cephalopods are the main prey for most beaked whale species but fishes and crustaceans can occasionally be consumed (MacLeod et al. 2003; Spitz et al. 2011). Stomach contents reveal the presence of deep-water species living below 200 m depth including pelagic squid species such as cranchids (*Teuthowenia megalops* and *G. armata*), histioteuthids (*H. reversa* and *H. bonnellii bonnellii*) and giant octopod (*Haliphron atlanticus*) as well as benthic or benthopelagic fishes and salps (Spitz et al. 2011). Indeed, dives deeper than 1000 m and longer than one hour are recorded for these species with echolocation clicks and foraging buzzes observed below 600 m (Fig. 2.8; Madsen et al. 2014). In general, the prey of *Ziphius* and *Hyperoodon* spp appear larger than the prey of *Mesoplodon* spp, even given that body length of individual predators may influence prey size (MacLeod et al. 2003; MacLeod and D'Amico 2006).



**Fig. 2.8.** Dive profile of a Blainville's beaked whale. Red dots represent foraging buzzes and the yellow line represents the echolocation clicks. From Madsen et al. (2014).

Most beaked whales are listed in Appendix II of CITES (Convention on International Trade of Endangered Species) and as "data deficient" by the IUCN (International Union for the Conservation of Nature) except the Cuvier's beaked whale which is listed as "Least Concern" since 2008 (Taylor et al. 2008a). Due to the difficulty to estimate the abundance of beaked whales (cryptic behaviour, lack of data, difficult identification at species level), population trends are unknown for these species (Read and Wade 2000). Beaked whales are occasionally taken in pelagic driftnets (Read and Wade 2000) and bottlenose whales and *Berardius* have been hunted in the nineteenth century (Mead 2009). However, we know almost nothing about contaminant, toxin, shipping noise or bycatch impacts (Madsen et al. 2014). Recently, studies have demonstrated the sensitivity of beaked whales to underwater noise pollution (Frantzis 1998; Balcomb and Claridge 2001; Brownell et al. 2004; Fernández et al. 2005; D'Amico et al. 2009). Powerful mid- to low-frequency impulse sounds can have impacts on beaked

whales (Frantzis 1998). Several unusual stranding events have occurred in connection with the use of military sonars (Balcomb and Claridge 2001; D'Amico et al. 2009). In other cases these sounds may lead to permanent or temporary loss of hearing and more often to behavioural disturbances (Fernández et al. 2005; DeRuiter et al. 2013). Besides, noise pollution is considered to result in an acoustic masking effect, which would considerably limit the cetaceans' acoustic perception of the environment and consequently their foraging performance and social functioning (Frantzis 1998; Brownell et al. 2004).

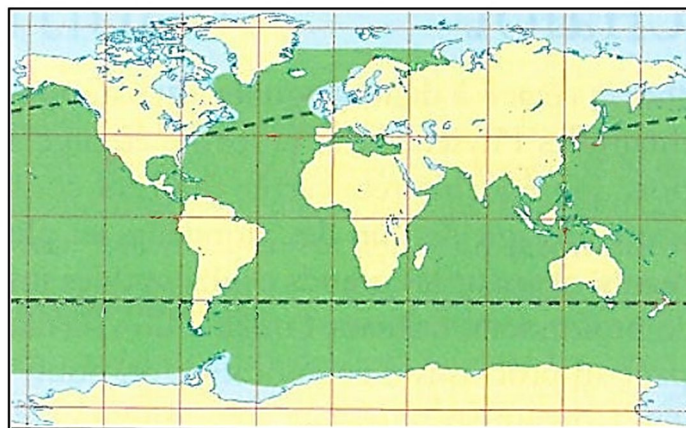
### 2.2.2 The sperm whales

The *Physeteridae* family contains only one species, the sperm whale (*Physeter macrocephalus*). Sperm whales are the largest representatives of the odontocetes (toothed whales) and show an important sexual dimorphism with male adult size reaching 16 m and 45 tons and female adult size reaching 11 m for 15 tons (Whitehead 2009). Sperm whales are characterised by a very large squared head (25% to 35% of its body length) which contains the spermaceti organ. The lower jaw contains from 20-26 large conical teeth, seemingly not involved in feeding (Whitehead 2009). Sperm whales have a single asymmetrically located blowhole on the left side. They have a triangular fluke, small pectoral fins and a small dorsal fin (Fig. 2.9).



**Fig. 2.9.** The sperm whale (*Physeter macrocephalus*). Illustration from Laura Hedon.

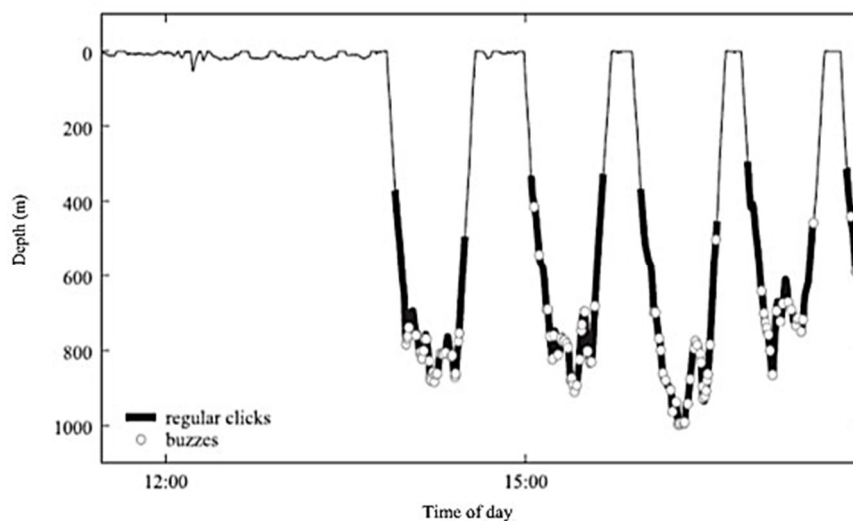
The distribution of sperm whales is one of the most extended among all marine mammal distributions, ranging from the equator to the edge of the pack ice in both hemispheres (from 60°S to 60°N; Fig. 2.10; Rice 1989; Shirihai 2006). Females and immature males are often observed in tropical and subtropical waters while adult males migrate towards the pack ice edge (Engelhaupt et al. 2009). As beaked whales, sperm whales are associated with deep-water and steep slope habitats (Rice 1989; Jaquet and Whitehead 1996; Praca et al. 2009).



**Fig. 2.10.** Distribution area of the sperm whale. Dotted lines represent the limits of the female distribution area. Adapted from Shirihai (2006).

Female sperm whales usually gather at low-latitudes in family units with on average 12 females and their youngs. Young males leave this maternal unit at between 4 and 21 years to join male groups of the same age (bachelor groups); largest mature males migrate alone towards polar waters and occasionally return to tropical waters to mate (Whitehead 2003). Females are sexually mature at around 9 years old and give birth to a single 4 m calf approximately every five years, after a 14-16 month gestation. Males are sexually mature at around 18 years but play an active role in reproduction only once they leave the bachelor group (Whitehead 2003). Sperm whale longevity is estimated at 60-70 years, but knowledge is scarce and this estimation is probably negatively biased (Whitehead 2003).

Sperm whales mainly feed on a variety of cephalopods, from mesopelagic to bathypelagic species (Santos et al. 1999; Santos and Pierce 2002), ranging from small chiroteuthids ( $\approx 400$  g) to giant squids ( $\approx 400$  kg; Clarke et al. 1993; Spitz et al. 2011). They, notably males, also feed on medium-sized fishes (Kawakami 1980; Clarke et al. 1993) depending on prey availability (Whitehead et al. 2003; Evans and Hindell 2004). Similarly to beaked whales, sperm whales forage at deep depth, from 400 m to over 1000 m and for over an hour (Fig. 2.11; Watwood et al. 2006).



**Fig. 2.11. Dive profile of a sperm whale.** White dots represent presence of foraging buzzes and the bold line represents the echolocation clicks. From Watwood et al. (2006).

Sperm whales are listed in Appendix I of CITES, in Appendices I and II of CMS (Conservation of Migratory Species of Wild Animals) and as “vulnerable” by the IUCN (Taylor et al. 2008b). As for beaked whales, population trends are difficult to estimate, and the worldwide population is estimated at 360,000 individuals with a noticeable decline of population in density (relative to other areas) in the Mediterranean Sea (Whitehead 2002). Sperm whales have been hunted for two centuries (until the end of the 1980s), removing up to 25,000 whales per year; today large-scale commercial harvesting has ceased and only few very small-scale fisheries in Japan and Indonesia persist. Sperm whales are subject to entanglement in fishing gear, in gillnets and driftnet fisheries (Barlow and Cameron 2003; Reeves and Notarbartolo di Sciara 2006) and they are increasingly reported to cause depredation in some fisheries (Hucke-Gaete et al. 2004). Despite high levels of contaminants being reported in this species, the effects on its health status are still unknown (Nielsen et al. 2000). Effects of noise pollution is also uncertain, because mortality has not been documented (no evidence of stranding associated with military sonars) but short-term effects, such as avoidance of sonars or seismic surveys, have been noticed (Bowles et al. 1994; Madsen et al. 2002). Sperm whales are also affected by ship collisions (Jensen and Silber 2003; Carrillo and Ritter 2010) and ingestion of marine debris (Jacobsen et al. 2010; Unger et al. 2016).

### 2.2.3 The kogiids

The *Kogiidae* family represents two species, the dwarf sperm whale (*Kogia sima*) and pygmy sperm whale (*Kogia breviceps*). The kogiids are the least sighted deep-divers, resulting in a limited knowledge of the genus. Pygmy and dwarf sperm whales are medium-sized odontocetes reaching an adult size of <4m with no apparent sexual dimorphism (Fig. 2.12; Leatherwood and Reeves 1983). As sperm whales, they have a left single blowhole and a spermaceti organ (Price et al. 1984; McAlpine 2002). They have a sharp rostrum, short lower jaw and a line of clear pigments behind the eye recalling the shape of fish opercula (Leatherwood and Reeves 1983; Caldwell and Caldwell 1989; McAlpine 2002). The distinction between the two species is difficult: they have a large body, a dark grey back and a white belly (Yamada 1954) but the dwarf sperm whale is slightly smaller than its congener. The dorsal fin of the dwarf sperm whale is located more in the middle of the back and is sharper and more vertical than the dorsal fin of the pygmy sperm whale (Caldwell and Caldwell 1989; McAlpine 2002). Their teeth are sharp and thin, between 10-16 pairs, but the teeth of the pygmy sperm whale are larger and longer (Caldwell and Caldwell 1989).



Fig. 2.12. The pygmy sperm whale (*Kogia breviceps*). Illustration from Laura Hedon.

Kogiids are rarely seen at sea and most of the information comes from strandings. Both species live in deep tropical and warm-temperate oceanic waters beyond the continental shelf, with no evidence of migrations (McAlpine 2002). They are both widely distributed, but the dwarf sperm whale seems to live closer to the continental slope in warmer waters while the pygmy sperm whale would prefer more temperate waters (Fig. 2.13; Caldwell and Caldwell 1989; McAlpine 2002).

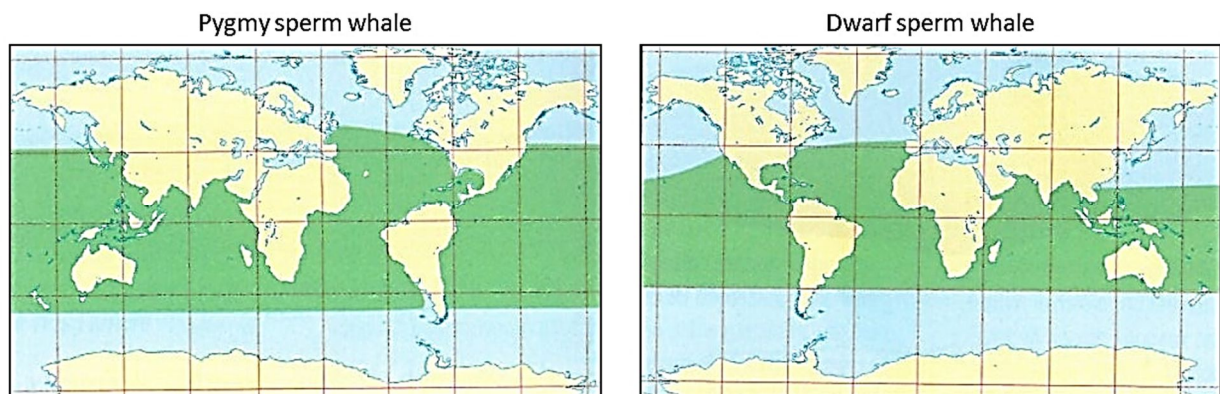


Fig. 2.13. Distribution area of the pygmy and dwarf sperm whales. Adapted from Shirihai (2006).

The two *Kogia* are generally undemonstrative and often float under the sea surface. When they dive, they do not lift their fluke and may emit a reddish-brown fluid when they are afraid (McAlpine 2002). Little is known about their ecology; they usually gather in small groups, often less than 6

individuals. Sexual maturity is reached between 4 and 5 years and gestation lasts 9 to 11 months. The maximum recorded longevity is 23 years (McAlpine 2002; Shirihai 2006).

Staudinger et al. (2014) showed similarities in feeding ecologies of the two species but also a trophic niche overlap in the Atlantic Ocean off the United States, although larger prey sizes were found in the pygmy sperm whale. Both kogiids are mainly teuthophagous, they feed primarily on oceanic cephalopods (*Histioteuthididae*, *Ommastrephidae* and *Cranchiidae*) but also on neritic cephalopods (*Sepiolidae*), mesopelagic fishes and crustaceans. The feeding area concentrates on the deeper shelf and slope, particularly in the epi- and meso-pelagic zones (Santos et al. 2006; Beatson 2007; Spitz et al. 2011; Staudinger et al. 2014).

Kogiids are listed in Appendix II of CITES and as “data deficient” by the IUCN. Their population trends are unknown and there are no global abundance estimates (Taylor et al. 2012a; 2012b). They have never been hunted commercially, but are sometimes caught by coastal whaling operations or harpoon fisheries in Japan, Antilles, Indonesia or Sri Lanka (Caldwell and Caldwell 1989). Rare bycatch in gillnets, driftnets and purse-seine fisheries have been reported but would not represent a cause for concern regarding their conservation (Baird et al 1996; Barlow et al. 1997; Zerbini and Kotas 1998; Perez et al. 2001). Ingestion of plastic debris was noticed from stranded animals and would be common (Caldwell and Caldwell 1989; Laist et al. 1999). In addition, anthropogenic sounds seem to have an impact on these species but, even if gas bubble lesions have been reported in some strandings, no clear link with military or seismic sounds was established (Hohn et al. 2006; Wang and Yang 2006; Yang et al. 2008).

To sum up, the three species groups, beaked whales, sperm whales and kogiids, can all be considered as rare, data deficient to different degrees and threatened by anthropogenic activities, particularly noise pollution. As a result, the aim of this thesis which is to map the habitat of these species is of particular interest to military fleets and authorities as well as to several civil activities such as seismic prospecting and major construction programmes at sea. Indeed, a precautionary management strategy would firstly favour avoidance of deep-diver distribution hotspots and secondly attempt to reduce potential exposure.

## 2.3 PRESENCE-ABSENCE AND COUNT-BASED MODELS VERSUS PRESENCE-ONLY MODELS

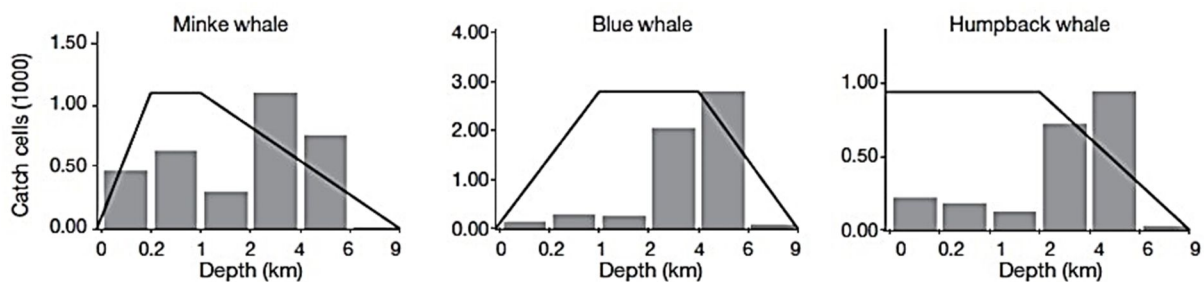
As described in Chapter 1, species habitat preferences are determined by using statistical models gathered under the name ‘habitat models’ (Elith and Leathwick 2009). These models allow modelling of the spatial distributions of animal sightings as a function of environmental and/or spatial characteristics at these locations in order to promote understanding of ecological factors driving species distribution or to provide spatial predictions in a specific area (Austin 2002; Guisan and Thuiller 2005; Redfern et al. 2006; Elith and Leathwick 2009). Habitat models are performed in four steps: calibration, prediction, evaluation and uncertainty quantification (Redfern et al. 2006). There are generally three categories of habitat models depending on whether they require presence-absence, abundance and presence-only data, *i.e.* whether effort data are recorded in parallel to sighting data or not (Guisan and Zimmermann 2000). Throughout this thesis, the three categories of habitat models were used and here, without being exhaustive, I describe some techniques used to model marine species distributions.

### 2.3.1 Presence-only models

Presence-only models only require detection data, such as opportunistic data, where the absence data are missing because effort data and non-detection data are not recorded (Hirzel et al. 2002). They allow for the identification of potentially suitable sites by displaying the environmental conditions that are similar to the sites where animals were recorded (Elith et al. 2006). The accuracy of presence-only model outputs is conditional on random or representative sampling of the habitat at the data collection stage (Yackulic et al. 2013) but presence-only models can be helpful when absence data (that is effort data) are not available (Zaniewski et al. 2002).

#### Environmental envelope models

Envelope models are the simplest available models; they allow to quantify relationships between animal distributions and environmental variables. This consists in overlaying the sightings and the maps of habitat variables to determine the species range and generate predictive maps (Redfern et al. 2006). The environmental envelope is defined by minimum and maximum values of habitat variables that contain a certain percentage of species occurrences. These environmental envelopes are multidimensional and projected into a two dimensional space (Redfern et al. 2006). Envelope models were, for example, used to map the global distribution of various marine mammals based on species habitat usage described by experts (Kaschner et al., 2006). Species were assigned to niche categories at a broad scale and, by linking these niche categories to local environmental conditions, they generated an index of relative environmental suitability and made predictions over the entire map (Fig. 2.14; Kaschner et al. 2006). Environmental envelope models appear relevant to map the large scale distribution of species, particularly for data-poor rare species, but the broad and static nature of environmental envelopes does not allow a detailed understanding of the processes underlying species habitat selection and to make predictions at a finer scale because there is no formal fitting mechanism and no model (Redfern et al. 2006).



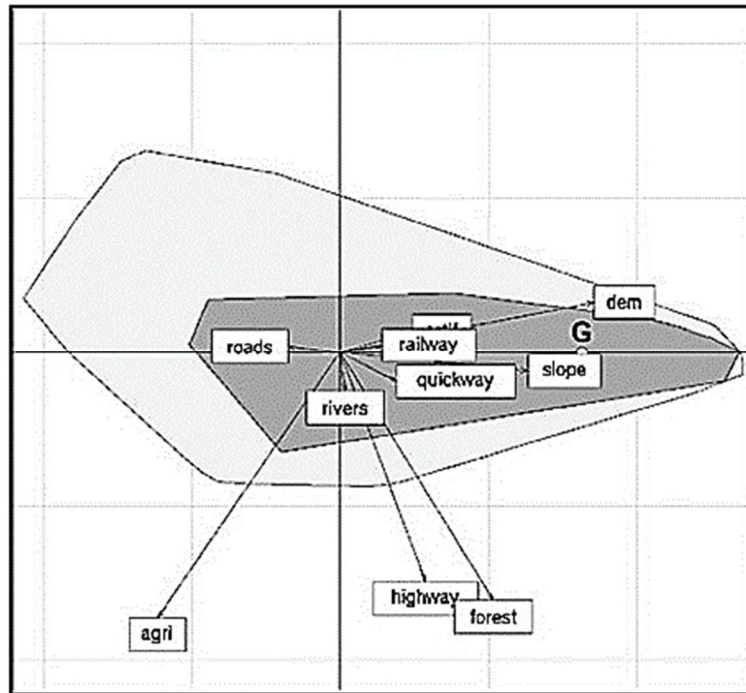
**Fig. 2.14. Example of environmental envelope model.** Each species have been assigned to broad-scale habitat categories defined by the depth (trapezoidal probability distributions) based on expert knowledge. Each barplot represent the frequency distributions of cells where the species was present. This suggest that minke whales and blue whales occurred mainly on the continental slope while the humpback whales occurred on the continental shelf and the continental slope. From Kaschner et al. (2006).

#### Ecological niche factor analysis

Ecological niche factor analysis (ENFA) is based on the Hutchinson's ecological niche concept, which describes a niche as a multidimensional hyper-volume defined by the environmental conditions in which a species population can persist (Hutchinson 1957). The method involves a factor analysis that measures habitat suitability by comparing, within this multidimensional space of ecological variables, the distribution of species occurrences to a reference set that describes the whole study area (Hirzel et al. 2002). This analysis uses various factors: a factor that maximises marginality, *i.e.* the niche position of



the species compared to the mean habitat in the reference area, and factors that maximise specialisation, *i.e.* the variance of the species distribution width with respect to the overall distribution in the whole reference area (Hirzel et al. 2002). ENFA appears to be useful to estimate the suitability for cryptic and rare species with small datasets (Fig. 2.15; Hirzel et al. 2002; Basille et al. 2008; Praca et al. 2009) but Brotons et al. (2004) showed a better performance of GLMs, compared to ENFA, to predict the species habitat suitability.



**Fig. 2.15.** Example of an ENFA biplot describing habitat selection of the Lynx (*Lynx lynx*) in the Vosges Mountains. The x-axis represents marginality and the y-axis axis represents the first specialisation. Grey polygons correspond to the minimum convex polygon enclosing all the projections of the available (light grey) and used points (dark grey). The white dot G corresponds to the centroid of the used habitat. The arrows are the projections of the environmental variables. Adapted from Basille et al. (2008).

### Genetic algorithm for rule-set production

Genetic algorithm for rule-set production (GARP) is a machine-learning genetic algorithm used to delineate ecological niches of a species, *i.e.* where environmental conditions are favourable to maintain the species population (Stockwell and Peters 1999). A GARP model is composed of mathematical rules included in an iterative process that searches for non-random correlations between occurrences and environmental parameter values (Guinan et al. 2009). At each step, there is a rule selection, a test of the rule, an evaluation and an incorporation or rejection of the rule. For example, one rule is based on a single value of a variable (*e.g.* if the species is recorded at depth of 2000 m and slope of 2.5° then the species is predicted to be present for these particular environmental conditions), another rule is based on the range of variable values (*e.g.* if the species is recorded at depth between 1000 m and 2000 m and slope between 1° and 2.5° then the species is predicted to be present in this environmental envelope). At each step, the predictive accuracy of the model (*i.e.* the proportion of species location data predicted correctly by the model) is evaluated and the change in predictive accuracy from the previous step is used to evaluate whether or not a particular rule is incorporated or rejected into the model, allowing to predict habitat suitability (Stockwell and Peters 1999; Tsoar et al. 2007; Guinan et al. 2009). According to Tsoar et al. (2007), GARP showed a better predictive accuracy than ENFA (Fig. 2.16).

For example, GARP has been used in marine ecology to model the distribution of habitats of a data-poor cold-water coral species (Guinan et al. 2009).

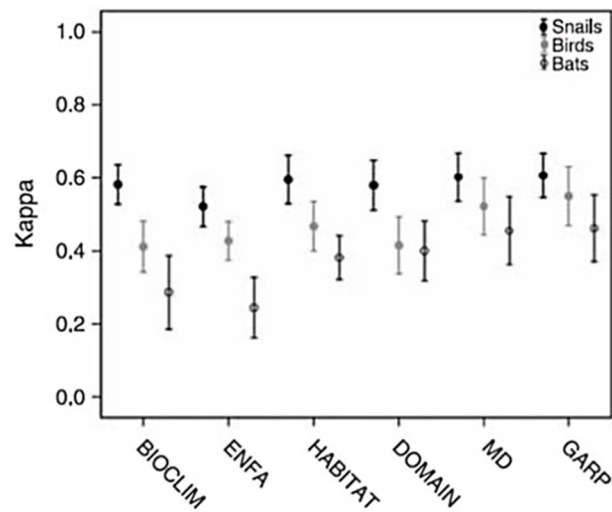


Fig. 2.16. Comparison of six presence-only modelling methods (BIOCLIM, ENFA, HABITAT, DOMAIN, MID, GARP) using the Kappa criterion. The six methods were tested on three animal groups (snails, birds and bats). From Tsoar et al. (2007).

### Maximum entropy modelling

The maximum entropy model (MaxEnt) is a machine-learning method that uses an optimisation procedure which compares the presence of the species with the characteristics of the environment based on the principle of maximum entropy (Phillips et al. 2006). This principle states that when fitting a probability distribution to data, the best distribution is the one which maximises entropy or the “uncertainty” (Jaynes 1957) by respecting the constraint that the expected value for each environmental variable under this estimated distribution matches its empirical average (Phillips et al. 2006). Generally, MaxEnt is equivalent to a generalised linear model (cf. below), more specifically, to a Poisson regression. MaxEnt is a commonly used presence-only model in the literature and was used to model habitat suitability of cetaceans (Fig. 2.17; Edrén et al. 2010; Thorne et al. 2012).

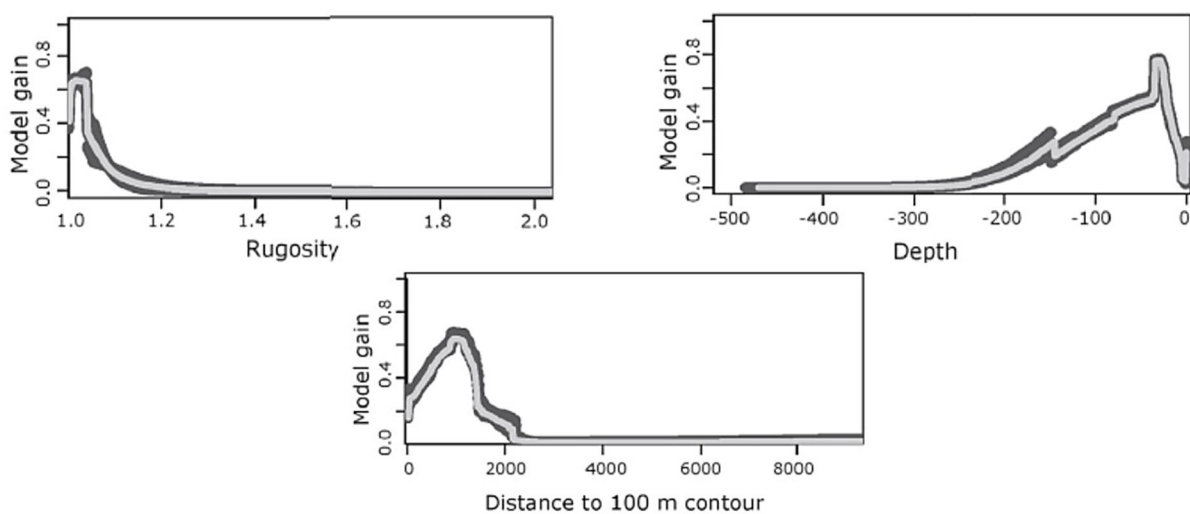


Fig. 2.17. Example of response curves of a MaxEnt model used to predict the distribution of the spinner dolphin (*Stenella longirostris*) in the Hawaiian Archipelago. Light grey lines represent the response curves and shaded regions represent the standard deviation. Adapted from Thorne et al. (2012).

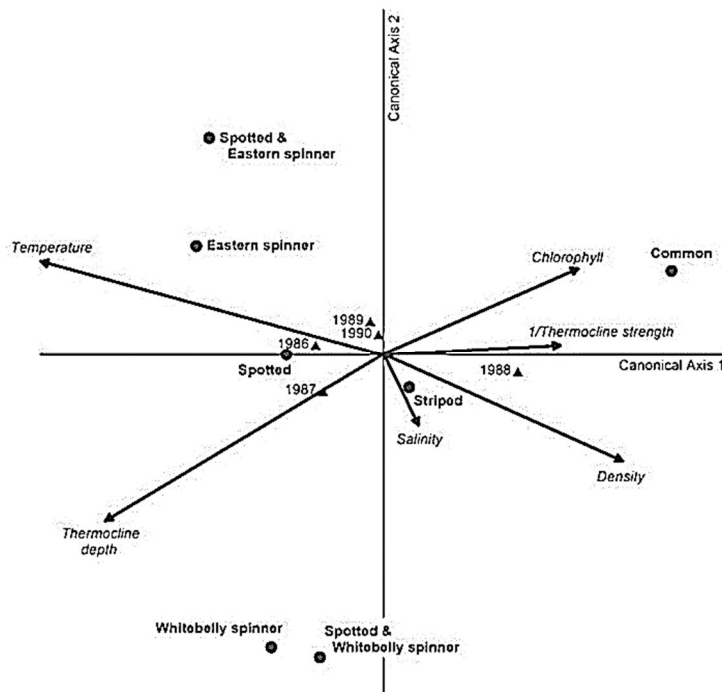
### 2.3.2 Presence-absence and count-based models

Presence-absence models require presence and effort data that are recorded during planned surveys (or at least surveys during which effort is recorded), where each on-effort sighting represents a detection of the target species. Count-based models require count data instead of presence. Such models include ordination methods, classification and regression trees, regression models and artificial neural networks (Guisan and Zimmermann 2000; MacKenzie et al. 2002; Brotons et al. 2004; Franklin 2010).

#### Ordination techniques

Ordination or gradient analyses are multivariate analyses that order species along environmental gradients to summarise and reduce the dimensions of complex multivariate datasets (Jongman et al. 1995). These analyses describe species-environment relationships by dividing the variance of the original dataset into various orthogonal and independent axes which are linear combinations of environmental variables. The data points are projected onto a two dimensional space, defined by two axes, in which similar data points are plotted close together, and dissimilar data points are placed far apart; the farther from the origin a point is found, the more informative the corresponding variable is (Pielou 1984).

There are many ordination techniques, *e.g.* principal components analysis, which assumes linear relationships between species and their environment, or canonical correspondence analysis which assume unimodal relationships. Ordination analyses are frequently used in community ecology and genetics to relate community composition to the environment (Dollhopf et al. 2001) or to discriminate species according to different habitats (Fig. 2.18; Ballance et al. 2006).

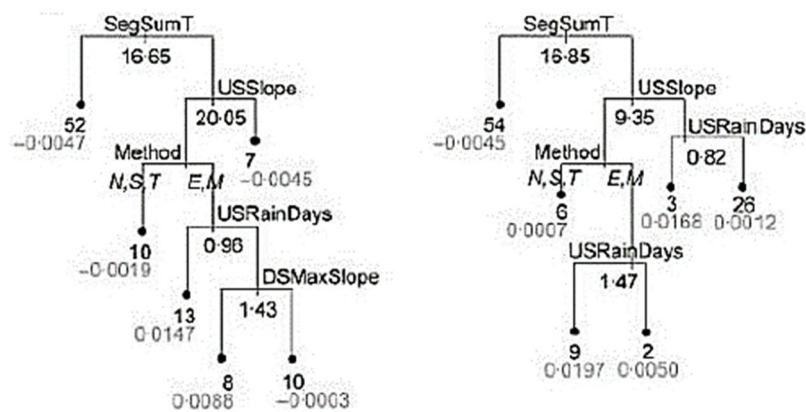


**Fig. 2.18.** Illustration of an ordination technique, a canonical correspondence analysis. Combinations of environmental predictors that explain the greatest proportion of variance in density of seven dolphin species are represented by the two Canonical Axes. Arrows represent the direction and degree of influence of each environmental variable. Points represent the location of the species in the habitat space identified by the two axes. A habitat segregation is apparent between common, spotted & eastern spinner, and spotted & whitebelly spinner dolphins. From Ballance et al. (2006).

### Classification and regression tree analysis

Classification and regression tree analysis (CART) constitute a non-parametric method used to determine the most significant variables in a dataset by binarily and recursively partitioning the dataset into smaller homogeneous sub-datasets with minimum prediction errors (Elith et al. 2008). If the response variable is categorical, a classification tree is used while with a continuous response variable, a regression tree is used (De'ath and Fabricius 2000). With these techniques, it is possible to highlight complex interactions with variables and predict patterns (Marmion et al. 2009) but misclassification errors can be generated, which can be minimised using a boosted regression trees that assemble various simple trees (Fig. 2.19; Elith et al. 2008).

Despite their use in ecology to explore habitat preferences of species (MacLeod et al. 2007; 2008) or determine the past distribution of various whale species (Monsarrat et al. 2015), decision trees are less efficient than regression models because the tree structure is mainly influenced by the sample data (the splits can greatly vary depending on the sample data) and, contrary to regressions, they can hardly model smooth functions, which limit their predictive performance (Hastie et al. 2001).



**Fig. 2.19.** An example of boosted regression tree used to model short-finned eels (*Anguilla australis*) occurrence. Each name (SegSumT – summer air temperature; USSlope – average slope in the upstream catchment; Method – Fishing method; USRainDays – days per month with rain >25 mm; DSMaxSlope – maximum downstream slope) represent an environmental variable that allowed to split the parent dataset (node). Split values are displayed under the split, and terminal nodes show percentage of sites in that node (black) and prediction in logit space (grey). From Elith et al. (2008).

### Artificial neural network

Artificial neural network (ANN) is being increasingly used, notably to model biological functions thanks to a complex non-parametric process. ANN is based on the same functioning as the animal brain with connected artificial neurons that learn from experience. The network is composed of three layers (input, hidden and output layers), each being composed of independent neurons. Each neuron of a layer is linked to the neurons of the following layer through multivariate linear functions (Fig. 2.20; Lek and Guégan 1999; Liu et al. 2010).

Due to their leaning ability and highly flexible functions, ANNs have great capacity in predictive modelling, especially when the underlying data relationships are unknown (Lek and Guégan 1999). However, it is a “black box” approach with a slow learning ability that needs abundant data (Liu et al. 2010) but was used, for example, to predict the probability of sperm whale occurrence in the central Mediterranean Sea (Aïssi et al. 2014).

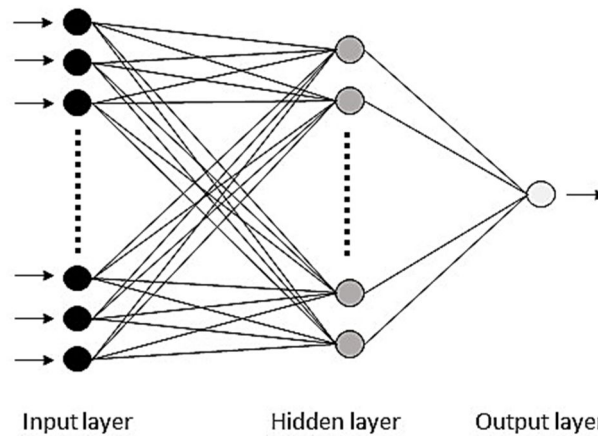


Fig. 2.20. Schematic illustration of a neural network. Adapted from Lek and Guégan (1999).

### Regression models

The principle of regression models is to model the relationships between a response variable, abundance or probability of occurrence of a given species, and a single (simple regression) or a combination (multiple regression) of explanatory variables, such as environmental. It is probably the most used method in habitat modelling (Guisan and Zimmermann 2000; Redfern et al. 2006). Regression models are conditioned by the use of non-collinear explanatory variables and independent observations. Regression models encompass different types of models depending on the form of their functional relationships. The most basic version assumes a linear relationship between response variable and explanatory variables, and assumes the response variable has a Gaussian distribution (Guisan and Zimmermann 2000). However, linear functions can hardly represent ecosystem complexity and Generalised Linear Models (GLMs) or Generalised Additive Models (GAMs) are often preferred.

GLMs can model nonlinear relationships between the response variable and the explanatory variables by using different types of statistical distributions (*e.g.* normal, Poisson, binomial, negative binomial, Gamma probability distributions). A link function relates the linear predictor to the mean of the response variable allowing a transformation to linearity and a constraint of predictions within the range of the response variable values (McCullagh and Nelder 1989; Guisan and Zimmermann 2000). GLMs allow to consider an over-dispersion and a non-constant variance structure in the data (McCullagh and Nelder 1989; Guisan and Zimmermann 2000; Guisan et al. 2002). They have commonly been used to model habitat preferences of marine mammals (Fig. 2.21; Cañadas et al. 2002; Praca et al. 2009; Bailey and Thompson 2009, Becker et al. 2010).

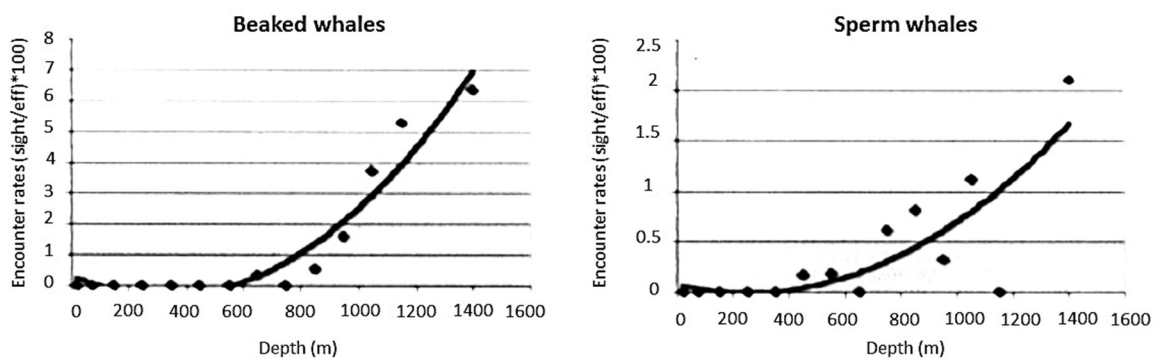


Fig. 2.21. Linear functions of a GLM used to relate beaked whale (*Ziphiidae*) and sperm whale (*Physeter macrocephalus*) encounter rates to depth in the Mediterranean Sea. Adapted from Cañadas et al. (2002).

GAMs are semi-parametric extensions of GLMs allowing nonlinear relationships to be modelled (Hastie and Tibshirani 1986). Like GLMs, they use link functions and various distributions for the response variable but predictors are additive and modelled with a smooth function. As available data mainly determine the nature of the relationships between the response and the predictors, GAMs are sometimes described as data driven (Yee and Mitchell 1991; Guisan et al. 2002). In addition, over-fitting may be an issue with GAMs and a selection of an appropriate degree of smoothness (trade-off between the number of observations and the number of degrees of freedom) is necessary to avoid over-fitting of the data (Guisan and Zimmermann 2000; Guisan et al. 2002; Wood 2006b). Due to their ability to deal with nonlinear and non-monotonic relationships, GAMs are relevant to model ecological relationships. Various studies have used GAMs to predict the distribution of cetacean species with remotely sensed and *in situ* data (Becker et al. 2010), in an energetic guild approach (Mannocci et al. 2014a; 2014b) or in a model transferability approach (Fig. 2.22; Redfern et al. 2017). GAMs are the most widely used techniques in marine mammal studies (Guisan and Zimmermann 2000; Redfern et al. 2006; Mannocci et al. 2014; Redfern et al. 2017).

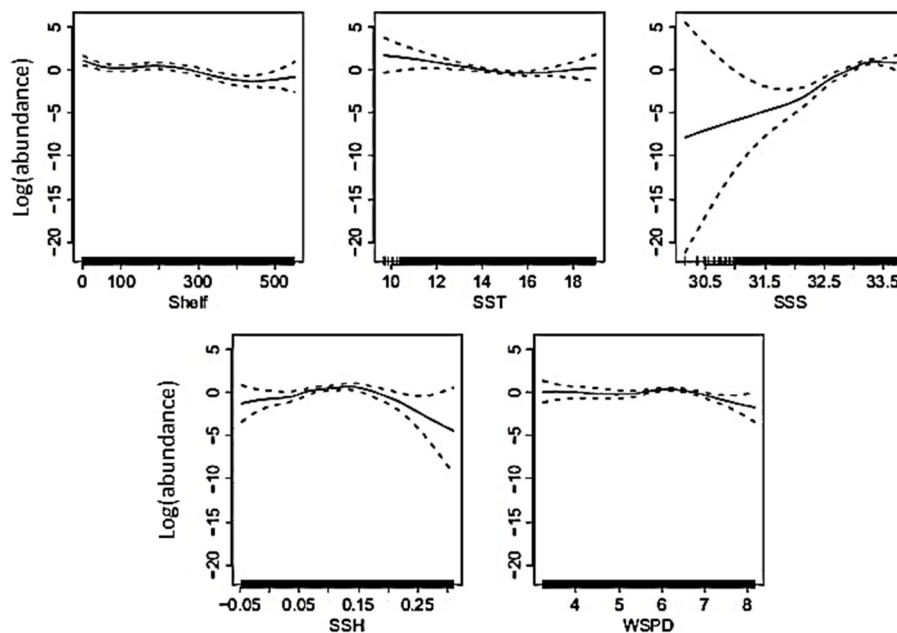


Fig. 2.22. Smooth functions of a GAM used to predict the distribution of blue whales (*Balaenoptera musculus*) in the California Current. Each plot relate the logarithm of the number of individuals to environmental predictors (Shelf – distance to shelf edge; SST – sea surface temperature; SSS – sea surface salinity; SSH – sea surface height and WSPD – wind speed). Adapted from Redfern et al. (2017).

To sum up, many species distribution models are available and based on a variety of principles. Each model described here has been used in the study of marine ecosystems but some of them seem to be more suitable to model habitat of rare species, my research topic. However, these models have rarely been tested on rare species data and the objective of the following chapter was to test some of these models when using datasets of rare species.

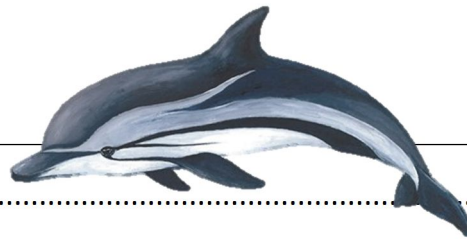


# Chapter 3

---

## RARE SPECIES: HOW TO MODEL THEIR DISTRIBUTION?

---



© Laura Hedon

### CONTENTS

3.1 CONTEXT AND OBJECTIVES .....	34
3.2 METHODOLOGY .....	35
3.2.1 Data collection .....	35
3.2.2 Environmental predictors .....	36
3.2.3 General analytical strategy .....	36
3.2.4 Reference and Baseline models .....	39
3.2.5 Thinning-out of the sightings .....	40
3.2.6 Assessment of the predictive performance of models.....	40
3.3 STAGE 1: COMPARISON OF MODELS FOR SCARCELY DETECTED SPECIES.....	42
3.3.1 Model selection and predictions of the dolphin models.....	42
3.3.2 Evaluation and comparison of the dolphin models.....	46
3.4 STAGE 2: HOW MANY SIGHTINGS TO MODEL RARE SPECIES DISTRIBUTIONS? .....	46
3.4.1 Model selection and predictions of the auk baseline models .....	47
3.4.2 Predictive performance of the experimental thinned models.....	48
3.5 GENERAL CONSIDERATIONS.....	52
3.5.1 Biological systems.....	52
3.5.2 Reference and Baseline models .....	53
3.5.3 Thinning-out sighting data.....	54
3.6 PREDICTING HABITATS OF RARE SPECIES .....	54
3.7 RECOMMENDATIONS FOR PRACTITIONERS.....	56

**T**HIS chapter aims to identify the most suitable methodology to model habitat preferences and distribution of rare species. Firstly, I compared the performance of several models and then tested them with a decreasing number of sightings, simulating a rare species dataset. This chapter compiles two papers, one of which has been published in *Ecological modelling* and the other one was submitted to *PLoS ONE* and is currently in revision (**Annexes A and B**).



### 3.1 CONTEXT AND OBJECTIVES

Identifying habitats needed and used by species is important for wildlife management and conservation (Cañadas et al. 2005; Bailey and Thompson 2009). However, as described in Chapter 2, a large number of habitat models exists with specific characteristics, some of them use presence-only data while others use presence-absence or count data. Except for presence-only models, which do not handle absence data, *i.e.* zeros, choosing among presence-absence or count-based models might be difficult depending on the studied species, particularly when focusing on rare species, because of the inherent difficulty of models to accommodate a large number of absences.

As mentioned earlier, rare species usually result in a low number of sightings per unit effort (Cunningham and Lindenmayer 2005) and this scarcity of sighting data makes it difficult to fit species distribution models (Welsh et al. 1996; Barry and Welsh 2002; Cunningham and Lindenmayer 2005), particularly because of data over-dispersion (variance greater than the mean). Hence, habitat modellers face two main issues. Firstly, they have to define if their data are under-, equi- or over-dispersed, and secondly, depending on their data, they have to find an appropriate model for the dispersion, particularly in the case of rare species (Redfern et al. 2006).

Although some studies have addressed the use of models for rare species datasets (Welsh et al. 1996; Engler et al. 2004; Cunningham and Lindenmayer 2005), the reliability of the predictions produced by these models and the uncertainty associated with these predictions remain pending issues. To address these issues, one option is to test if the performance of a species distribution model is maintained when the amount of input data decreases, which would assess the reliability of the models when handling small datasets of rare species.

Consequently, the aim of this chapter was to find an appropriate model to handle datasets with many zeros. To do that, I conducted an analytical procedure in two main stages. In a first stage, I compared the predictive performance of different count-based models and presence-only models and tested their ability to address an apparently zero-inflated dataset. In a second stage, with the best performing models, I determined a threshold for the number of sightings to be used to model the habitats of rare species and to assess how the threshold evolves depending on the type of model being used, in order to choose the most suitable one for the study of deep-divers. By using a small delphinid dataset in the first stage (approximately 92% zeros), I tested different types of models: GAMs with a Poisson, a Negative Binomial, a Tweedie and a zero-inflated Poisson distribution; a GLM with a zero-inflated Poisson distribution and a presence-only model, the MaxEnt model. Due to their ability to model separately the absences and presences (Lambert 1992), I assumed *a priori* that a zero-inflated Poisson model would perform best. However, the Negative Binomial and Tweedie distributions proved to provide good fits (Warton 2005; Dunn and Smyth 2005; Lindén and Mantyniemi 2011). In addition, with its multiple applications (Yackulic et al. 2013), including those by managers, and its ability to take into account the complex interactions between response and predictor variables (Elith et al. 2006; 2011; Phillips et al. 2004; Phillips and Dudik 2008), the MaxEnt model appeared to be another relevant tool for modelling habitats of rare species (Wisz et al. 2008). The first stage aimed to identify the best performing model while the second stage aimed to determine the minimum threshold of sightings needed to reliably fit these models. This threshold was expected to be lower for species with a narrow habitat than for more generalist species. Indeed, even with few sightings, the distribution pattern of a species with a narrow or specialised habitat would *a priori* require fewer sightings to be modelled

because these sightings would be sufficient to describe almost entirely the habitat used by the species, while for a widely spread distribution, a low number of sightings would not describe all the habitats used by the species.

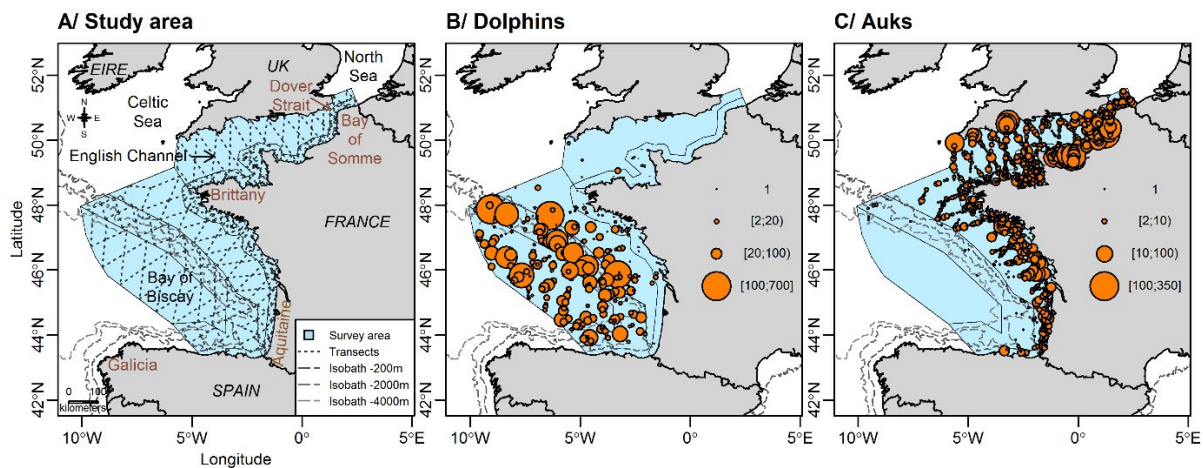
To determine this threshold, I conducted a sighting thinning experiment using two large datasets of marine megafauna: a small delphinid (as for the previous stage) and an auk dataset. These two taxa had different distribution patterns. The small delphinids group (hereafter called “dolphins”) included the common (*Delphinus delphis*) and striped (*Stenella coeruleoalba*) dolphins, which both have an offshore distribution and range at depth between 50-5000 m, thus representing a generalist taxon. The auk group (hereafter called “auks”) mostly consisted of the common guillemot (*Uria aalge*) and, to a much less extent, the razorbill (*Alca torda*) and the Atlantic puffin (*Fratercula arctica*), which all show a more coastal distribution (Fig. 3.1), ranging from 10 to 150 m deep and representing a more specialised taxon regarding depth, which is an environmental descriptor of major importance for marine top predators. Hence, the thinning of the dolphin dataset would generate datasets of a rare, non-specialist species living in a large habitat range, and the thinning of the auk dataset would simulate a rare, more specialised species living in a more restricted habitat range. Hence, these datasets represented two forms of rare species defined by Rabinowitz (1981; Chapter 1). This study helped me to determine the more flexible model that would be applicable to my data-poor deep-diver datasets and to define a precautionary threshold below which inferences from habitat modelling might be too fragile.

## 3.2 METHODOLOGY

### 3.2.1 Data collection

The sightings of auks and dolphins were collected during the two aerial surveys SAMM (*Suivi Aérien de la Mégafaune Marine – Aerial Census of Marine Megafauna*) conducted in the English Channel and the Bay of Biscay (Fig. 3.1). The surveys were conducted during the winter 2011-2012 (from mid-November to early February; 28,068 km of transects) and the summer 2012 (from mid-May to early August; 31,427 km of transects; Lambert et al. 2017a). A standard methodology for cetacean surveys was applied (Hammond et al. 2013) using twin-engine high-wing aircrafts equipped with bubble windows. The flights followed a zig-zag pattern, at a speed of 167 km/h and an altitude of 183 m. Observation conditions (Beaufort seastate, turbidity, cloud cover and glare severity) and sightings with group size were recorded following a taxon-specific methodology (Buckland et al. 2015): a line-transect methodology was used to record the cetacean sightings (Buckland et al. 2015), while seabird sightings were recorded using a strip-transect methodology (Certain and Bretagnolle 2008). In the line-transect methodology, the angle between the sighting and the track line was measured to determine the effective strip width (ESW; see the detection functions and estimated ESW in Laran et al. 2017a) on each side of the plane. In the strip-transect methodology, the sightings were gathered from a 200 m strip on each side of the plane, and it was assumed that all animals were detected. Effort data were split into 10km segments and sightings joined to the segments.

For the analyses, I only used the data recorded in summer for dolphins and in winter for auks. The sightings during these seasons represented the most abundant datasets, which allowed for the sighting thinning approach to be implemented. A total of 277 sightings (*i.e.* 14,477 individuals, Fig. 3.1) of dolphins and 1,455 sightings (*i.e.* 16,658 individuals, Fig. 3.1) of auks were recorded in good observation conditions (seastate <4 and medium to excellent observation conditions).



**Fig. 3.1.** The study area (A), and the dolphin (B) and auk (C) sightings recorded during the survey. The study area expands through the Bay of Biscay and the English Channel. The surveys were carried out along transects (dotted lines) following a zig-zag pattern across bathymetric strata. The sightings were classified by group sizes (1; 2-20; 20-100 and 100-700 individuals for dolphins and 1; 2-10; 10-100 and 100-350 individuals for auks), with each point representing one group of individuals.

### 3.2.2 Environmental predictors

Two types of environmental predictors at a 10 km resolution were used to model the habitats of the two taxa (Table 3.1). Static (or physiographic) predictors relate to the bathymetry and included depth and slope, and dynamic (or oceanographic) predictors describe the water masses and included mean, variance and gradient of sea surface temperature (SST), mean and standard deviation of sea surface height (SSH), and maximum current velocity (mostly referring to tidal currents in the study area (Appendix A.1 of Annex A). To avoid gaps in the remotely sensed oceanographic variables, a 7-day resolution was used. All available data were averaged over the 6 days prior to each sampled day (for more details, refer to Lambert et al. 2017a and Virgili et al. 2017b).

### 3.2.3 General analytical strategy

#### First stage: comparison of models for scarcely detected species

In the first stage, I arbitrarily chose a GAM with a Poisson distribution, appropriate for equi-dispersed data, as reference model (hereafter labelled as 'reference model') for comparison with the other models (Fig. 3.2). For this reference model, relationships between the densities of small delphinids and environmental variables were investigated by selecting the variables that best describe dolphin distribution. Next, I fitted GAMs with a negative binomial and a Tweedie distribution, which are suitable for over-dispersed data, a GAM and a GLM with a zero-inflated Poisson distribution, which are suitable when over-dispersion is due to zero-inflation, and a MaxEnt model, which is specific to presence-only data. Using these models, I applied the variables previously selected in the reference model. To finish, I compared all models by using various criteria such as the Akaike Information Criterion (AIC), deviance, rootograms (Kleiber and Zeileis 2016) and predicted density map to evaluate the predictive performance of each model and to select the best performing models for the second stage.

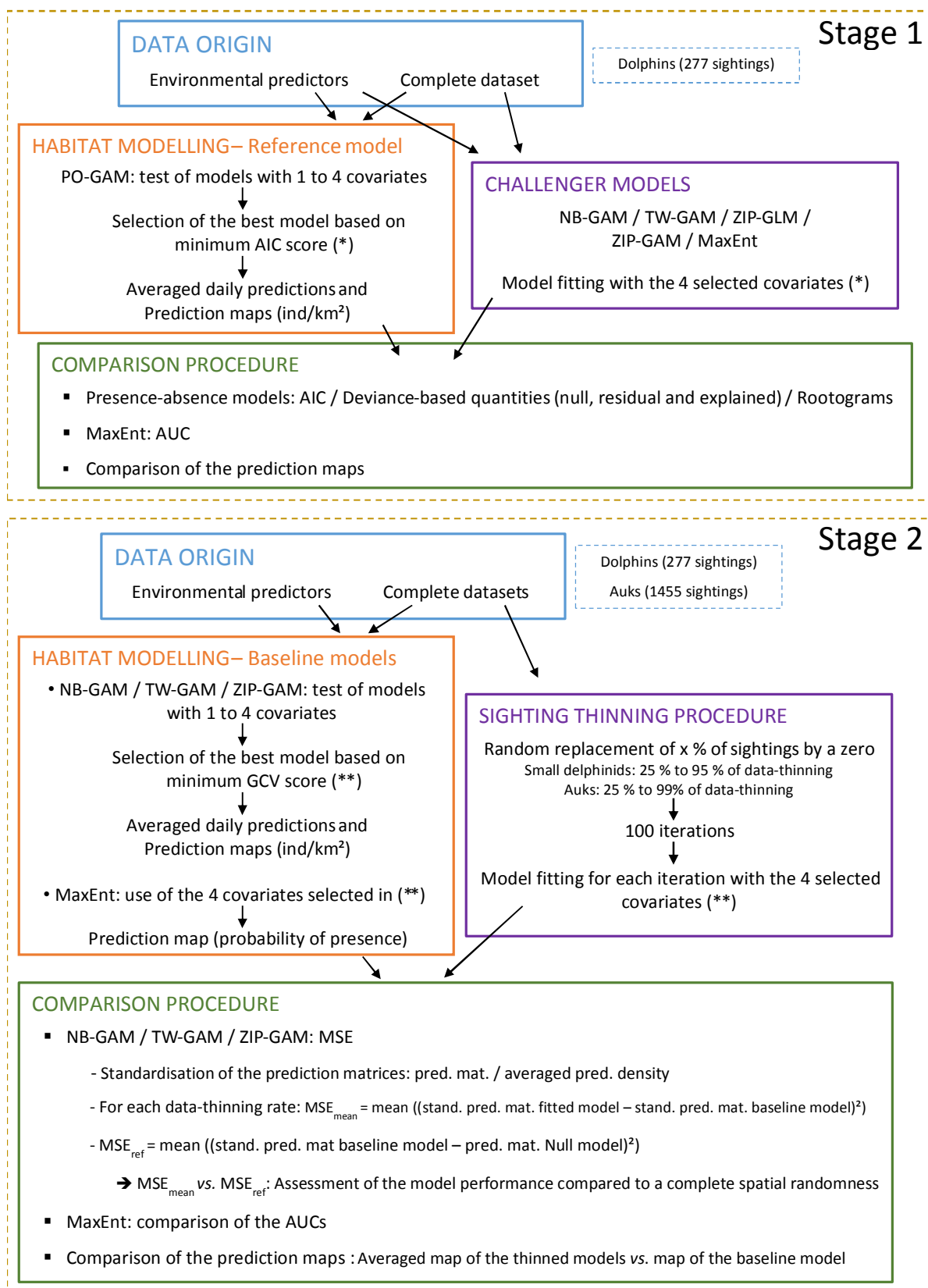
**Table 3.1. Environmental variables used for habitat modelling.** A: Depth and slope were computed from the GEBCO-08 30 arc-second database (<http://www.gebco.net/>). B: The mean, variance and gradient of the sea surface temperature (SST) were calculated from the ODYSSEA product (from My Ocean project <http://www.myocean.eu/>). C: The MARS 3D model from Previmer ([www.previmer.org](http://www.previmer.org)) was used to compute the mean and standard deviation of the sea surface height (SSH). D: The daily maximum intensity of the currents was computed from the MARS 2D model ([www.previmer.org](http://www.previmer.org)).

Environmental predictors	Sources	Effects on pelagic ecosystems of potential interest to top predators
<b>Physiographic</b>		
Depth (m)	A	Shallow waters could be associated with high primary production
Slope (°)	A	Associated with currents, high slope induce prey aggregation and/or primary production increasing
<b>Oceanographic</b>		
Mean of SST (°C)	B	Variability over time and horizontal gradients of SST reveal front locations, potentially associated to prey aggregations
Variance of SST (°C)	B	
Mean gradient of SST (°C)	B	
Mean of SSH (m)	C	High SSH is associated with high mesoscale activity and prey aggregation and/or primary production increase
Standard deviation of SSH (m)	C	
Daily maximum intensity of the currents (m.s <sup>-1</sup> )	D	High currents induce water mixing and prey aggregation

### Second stage: how many sightings to model rare species distributions?

In the second stage, I wanted to test the predictive capacity of the previous selected models when they were used with rare species datasets (Fig. 3.2). I tested a GAM with a negative binomial distribution, a GAM with a Tweedie distribution, a GAM with a zero-inflated Poisson distribution, and a MaxEnt model. These SDMs were first fitted to the original dolphin and auk datasets in order to select the four most important predictors for each taxon (hereafter referred to as ‘baseline models’). These baseline models served as a reference to compare models fitted to the thinned-out datasets (hereafter referred to as the ‘experimental models’). The original datasets were thinned of sightings by randomly removing 25-99% of the sightings to reduce the number of sightings used in the fitted models at a constant survey effort. For each thinning-out level, the four SDMs were fitted with the same explanatory variables as in the baseline model. Finally, predictions from experimental models were compared to those of the baseline models to determine the minimum number of sightings to reliably predict rare species distribution.

Although model performance is largely determined by its selected variables (Syphard and Franklin 2009), I used the same specification for each experimental model in this study to assess how the results were affected by the sighting thinning alone.



**Fig. 3.2. Flowchart of the methods used in the study.** PO-GAM: generalised additive model (GAM) with a Poisson distribution; NB-GAM: GAM with a negative binomial distribution; TW-GAM: GAM with a Tweedie distribution; ZIP-GLM: generalised linear model with a zero-inflated Poisson distribution; ZIP-GAM: GAM with a zero-inflated Poisson distribution; MaxEnt: maximum entropy model; AIC: Akaike information criterion; AUC: area under the curve; GCV: generalised cross-validation score; ind: individuals; MSE: mean squared error; stand.: standardised; pred.: prediction; mat.: matrix; ref: reference.

### 3.2.4 Reference and Baseline models

To fit GLM and GAMs, I used the ‘gam’ function with ‘poisson’, ‘nb,’ ‘tw’ and ‘zip’ family distributions within the ‘mgcv’ R package (R-3.1.2 version; Wood 2006a; 2013). A log function linked the response variable to the predictors; smooths were restricted to three degrees of smoothness for the GAMs (Ferguson et al. 2006), and an offset that considered the variation of effort per segment (Hastie and Tibshirani 1986) was included and calculated as the segment length multiplied by twice the ESW (see Laran et al. 2017a). After removing the combinations of variables with correlation coefficients higher than |0.7|, all models with combinations of 1 to 4 variables were tested. A maximum of four covariates was implemented to avoid excessive complexity and difficulty of interpretation (Mannocci et al. 2014a; 2014b).

In the first stage, the selection procedure, using the Akaike Information Criterion (AIC; Akaike 1974), was only applied on the reference model, a GAM with a Poisson distribution, called hereafter PO-GAM (only dolphin data). For the challenger models, I used the variables associated with the best PO-GAM to fit the models, there was no variable selection procedure. Hereafter, the fitted models will be called NB-GAM (for GAM with a negative binomial distribution), TW-GAM (for GAM with a Tweedie distribution), ZIP-GAM (for GAM with a zero-inflated Poisson distribution) and ZIP-GLM (for GLM with a zero-inflated Poisson distribution). The explained, null and residual deviances were extracted to assess the goodness-of-fit of each models.

In the second stage, the variable selection procedure using the lowest generalised cross-validation score (GCV; Wood 2006b), which estimates the mean prediction error using a leave-one-out cross-validation process (Clark 2013), was used on all baseline models (NB-GAM, TW-GAM, ZIP-GAM) and for the two taxa (dolphins and auks), to select the best models. For each taxon, the selected variables for NB-GAM, TW-GAM and ZIP-GAM were identical so that it was possible to compare the different models.

In addition, a presence-only model, the Maximum Entropy model (MaxEnt), was fitted (MaxEnt version 3.3.3; <http://www.cs.princeton.edu/~schapire/maxent/>; Phillips et al. 2006). The input file was the same as the reference and baseline models, but I removed all absences; hence, each line corresponded to one observation of dolphins or auks. For the environmental predictors, I used the four covariates selected by the reference and baseline models. Regarding model parameters, I used the “hinge” feature to generate models with smooth functions similar to GAM’s, with a default prevalence of 0.5 and a logistic output format to compare it to the relative probability of presence (Phillips and Dudík 2008; Elith et al. 2011; Merow et al. 2013).

For each fitted model, except for MaxEnt, predicted densities (in individuals per km<sup>2</sup>) were mapped on a 0.05° x 0.05° resolution grid. I computed the predictions for each day of the surveys and averaged the predictions over the entire survey period. To limit extrapolation, the covariates were constrained within the range of the covariate values used when fitting the models. Finally, I provided uncertainty maps by computing the variance around the predictions as the sum of the variance around the mean prediction and the mean of the daily variances. Then, the percentage coefficient of variation was calculated as

$$CV = 100 \times \sqrt{(\text{variance over the survey period}) / \text{mean over the survey period}}.$$

In addition, in the first stage, I assessed whether a prediction was an extrapolation or an interpolation by using the non-parametric Gower’s distance (King and Zeng 2007). An extrapolation is a

prediction for a combination of covariate values that falls outside the convex hull that is defined by the covariate data used to calibrate the model (King and Zeng 2007; Authier et al. 2016). However, even if a prediction falls outside this convex hull, this extrapolation can nevertheless be informed by calibration data lying in its neighbourhood. The neighbourhood of a prediction was defined as the calibration covariate data within a radius of one geometric mean Gower’s distance of the prediction (King and Zeng 2007). The geometric mean was computed from all pairs of calibration data point. The results from this extrapolation analysis were mapped to visually assess how trustworthy the predictions were.

### 3.2.5 Thinning-out of the sightings

To generate datasets of rare species, I thinned the original auk and dolphin sightings at different rates. I aimed to obtain a decreasing number of sightings, simulating thereby an increasing rarity of the two taxa. In the dolphin dataset, I randomly replaced 25, 50, 75, 90, 92 and 95% of the sightings with zeros, and in the auk dataset, I randomly replaced 25, 50, 75, 90, 92, 95, 97 and 99% of the sightings with zeros (Table B. 3.2; Appendices B.3 and B.4 show examples of thinning-out). For each thinning rate, sightings to be replaced with zero were randomly sampled without replacement, and the procedure was iterated 100 times, hence producing 100 randomly thinned or experimental datasets for each thinning rate. This procedure simulates different levels of species rarity as observed under a constant sampling effort. Removing part of survey effort (*e.g.* whole transects) would not have generated a greater rarity of the species but only a lower sighting effort; and would have led to similar results to the baseline models because encounter rates would have remained similar on average.

**Table 3.2. Number of sightings contained in the “new” datasets for each thinning rate and each species group.**  $n_{\text{sigh}}$ : number of sightings;  $n_z$ : number of segments with a zero;  $\%_z$ : proportion of zeros; “-”: sighting thinning was not performed; “Original”: initial (and complete) datasets.

Species groups		Sighting thinning rates								
		Original	25%	50%	75%	90%	92%	95%	97%	99%
Dolphins	$n_{\text{sigh}}$	277	208	139	69	28	23	14	-	-
	$n_z$	3043	3112	3181	3250	3292	3297	3306	-	-
	$\%_z$	91.7	93.7	95.8	97.9	99.2	99.3	99.6	-	-
Auks	$n_{\text{sigh}}$	1455	1091	728	364	146	116	73	44	15
	$n_z$	2046	2409	2773	3137	3355	3384	3428	3457	3486
	$\%_z$	56	66	76	85.9	91.9	92.7	94	94.7	95.5

### 3.2.6 Assessment of the predictive performance of models

Evaluating the predictive performance of a model requires demonstrating its consistency with raw observation data and comparing the outputs of several models (Pearce and Ferrier 2000). Each assessment criterion quantifies a particular aspect of a model performance and several criteria must be used in combination (Elith and Graham 2009). I calculated different selection measures to improve the relevance of model comparison.

#### First stage

The Akaike Information Criterion (AIC) was computed for each model to assess model relative fit: the lower the AIC, the better the model (Akaike 1974). I also examined several deviance-based

quantities (null, residual and explained) as a proxy of the model reliability to predict the frequencies of species occurrence (Elith and Graham 2009). A high explained deviance can indicate a good fit, whereas a high null deviance and a high residual deviance can indicate a bad one. Finally, to evaluate the absolute goodness-of-fit of the models and how they handled the excess of zeros, I plotted rootograms that compared, with histograms, the raw data frequencies to the frequencies fitted with the models (Kleiber and Zeileis 2016).

These methods cannot be readily applied with presence-only models, which leads to some difficulty for model comparison. To evaluate the predictive performance of MaxEnt, I used the Area Under the receiver operating characteristic Curve (AUC; Elith et al. 2006). This method works only on binary data (not on count data) and measures how a model can differentiate the sites where the species is present and the sites where it is absent. A perfect discrimination of the sites is revealed by a score of 1, a discrimination equivalent to a random distribution is indicated by a score of 0.5 and for a score lower than 0.5, the model performance is worse than a random guess (Elith et al. 2006). This AUC is directly provided by the MaxEnt software. However, with this method, I cannot compare the model performance to the fitted reference model. I thus transformed the PO-GAM prediction maps (only this one) to probability of presence with the formula:

$$\textit{presence probability} = 1 - e^{(-\textit{predicted density})}.$$

### Second stage

The baseline SDMs were selected using the minimum GCV score, a leave-one-out cross-validation process used to estimate mean prediction error and explained deviances (Wood 2006b; Clark 2013). However, for experimental models, I based the assessment of the predictive performance of the count-based models on two criteria: mean squared error (MSE; Wallach and Goffinet 1989; Harvey et al. 1997) and maps of the predicted densities. The MSE directly compared the prediction matrices of the experimental models to the prediction matrix of the baseline model. Each cell of the matrices provides the densities predicted by the model over the entire prediction area. The MSE is given by  $MSE = \text{mean} \left( \sum (\hat{Y}_{\text{exp}} - \hat{Y}_{\text{baseline}})^2 \right)$  (Wallach and Goffinet 1989; Harvey et al. 1997). Here, " $\hat{Y}_{\text{exp}}$ " represents the prediction matrix of an experimental model, and " $\hat{Y}_{\text{baseline}}$ " represents the prediction matrix of the baseline model. For each type of model and thinning rate, I averaged the MSEs of all the experimental models to obtain an averaged MSE (called  $MSE_{\text{mean}}$ ). Then, I investigated whether the predictions provided by the models fitted to sighting thinned-out datasets were better than those from a homogeneous process. For this purpose, I compared the MSE of each fitted model and the  $MSE_{\text{mean}}$  to a reference threshold, called the  $MSE_{\text{ref}}$ , which was calculated as the MSE between the prediction matrix of the baseline model (NB-GAM, TW-GAM or ZIP-GAM) and the prediction matrix of a null NB-GAM, TW-GAM or ZIP-GAM (which described a homogeneous spatial distribution). I assumed that if the MSE was higher than the  $MSE_{\text{ref}}$ , it was more appropriate to consider a homogeneous spatial distribution rather than taking into account the predictions provided by the experimental model.

Like in the first stage, to assess the predictive performance of the MaxEnt models, I used the AUC (Elith et al. 2006). I compared the AUC of each fitted model and the  $AUC_{\text{mean}}$  (averaged over the 100 fitted models) to the AUC of the baseline model and used a threshold value of 0.5 to assess the performance of the experimental models.

Finally, I compared the prediction maps of the models fitted to the thinned datasets to the prediction maps of the baseline models in order to determine the lowest sample size that did not change predicted distribution patterns. For each model type and each thinning rate, I averaged the predictions

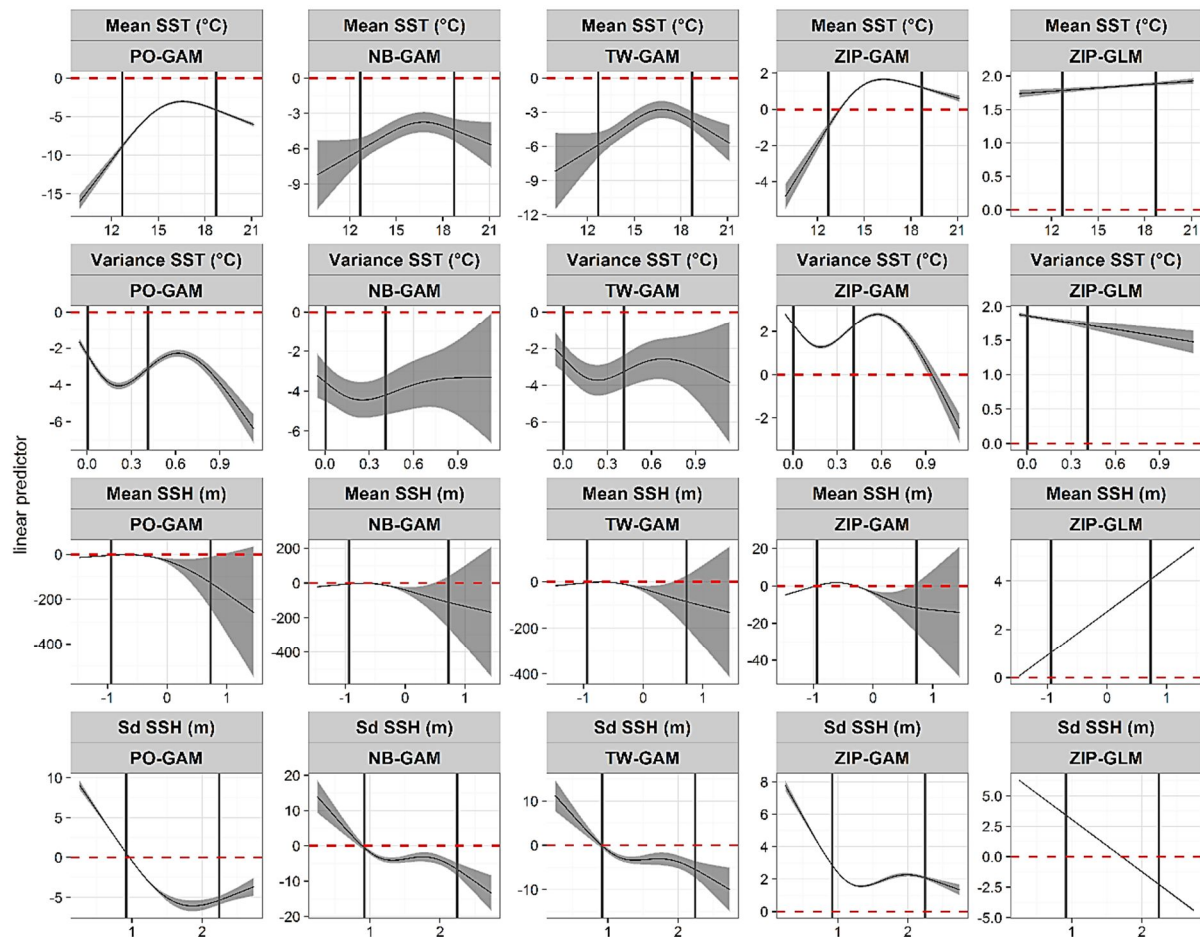


over the 100 models fitted to the thinned datasets and produced averaged prediction maps that I compared to the prediction maps of the baseline models. I averaged the predictions over the 100 fitted models to ensure a uniform data deletion throughout the area. In practice, habitat modellers only have one real dataset (not 100); hence, I compared the MSE or AUC of each fitted model to the  $MSE_{ref}$  or  $AUC_{ref}$  to determine the proportion of the model that provided good predictions.

### 3.3 STAGE 1: COMPARISON OF MODELS FOR SCARCELY DETECTED SPECIES

#### 3.3.1 Model selection and predictions of the dolphin models

Among the eight environmental predictors, the variables selected by the best PO-GAM, defined as the reference model, were the mean and variance of SST and the mean and standard deviation of SSH (Fig. 3.3). The highest densities of dolphins were predicted for stable temperatures at approximately 16°C (variance around 0°C), and a rather stable low average altimetry (SSH, around -0.5 m, standard error around 0.5 m).



**Fig. 3.3.** Functional relationships for the selected covariates of each count-based model of stage 1. The solid line in each plot is the smooth function estimate and shaded regions represent approximate 95% confidence intervals. The y-axis indicates the logarithm of the abundance in individual/km<sup>2</sup>. The x-axis indicates the values of the covariates and zero on the x-axis indicates no effect of the covariate. Best model fits are between the vertical lines indicating the 10<sup>th</sup> and 90<sup>th</sup> quantiles of the data.

The NB-GAM ( $k=0.028$ ), TW-GAM ( $p=1.573$ ) and ZIP-GAM ( $\vartheta=(-4.861, 0.26)$ ) showed fairly similar smooth functions compared to the PO-GAM, except for sd SSH (Fig. 3.3). However, confidence intervals around the functional relationships were significantly smaller and the predicted densities were higher in the case of the ZIP-GAM. The smooth functions of the ZIP-GLM showed increasing small delphinids densities with increasing SST mean and SSH mean and decreasing densities with decreasing SST variance and sd SSH, which was an opposite trend compared to the other models for SSH mean.

To complete the comparison of the models, we analysed the residuals of each fitted model (Appendix A.3 of Annex A). In all cases, there was an accumulation of residuals at zero and an over-dispersion of positive values, but it was less important for the TW-GAM. In addition, we calculated Cook's distances (Appendix A.4 of Annex A) to determine if some values highly influenced the fitted models (Cook's distance  $> 1$ ). It appeared that some values greatly influence the PO-GAM and ZIP-GLM. However, for NB-GAM, TW-GAM and ZIP-GAM, no value appeared to affect the models (Cook's distance  $< 1$ ) and the only values that could influence them correspond to non-extreme values of covariates. Consequently, that strengthened the results provided by the fitted models, especially for NB-GAM, TW-GAM and ZIP-GAM.

Finally, with an AUC of 0.822, the MaxEnt model predicted delphinids presence probabilities much better than a random prediction would do (AUC of 0.5).

Prediction maps of the PO-GAM showed a concentration of delphinids in offshore waters, from the continental shelf to the oceanic waters, with the highest densities over the slope (Fig. 3.4). The highest densities, which reached 30 individuals.km<sup>-2</sup>, were predicted in the north of Galicia, which is outside the survey area. In addition, we noticed a good match between observations and predictions of the model (Fig. 3.1). Within the survey area, predictions were associated with low uncertainties (Appendix A.6 of Annex A), which strengthened the results. In contrast, outside the survey area, patches of high densities were associated with higher uncertainties and needed to be considered with caution.

TW-GAM and NB-GAM predicted the same distribution as the reference model but with higher densities, maximum at 35 individuals.km<sup>-2</sup> for TW-GAM (Fig. 3.4; Appendix A.5 of Annex A) and maximum at 73 individuals.km<sup>-2</sup> for NB-GAM (Fig. 3.4; Appendix A.5 of Annex A). ZIP-GAM showed the same distribution patterns in the Bay of Biscay but with lower densities (max at 11 individuals.km<sup>-2</sup>, Fig. 3.4; Appendix A.5 of Annex A) and more individuals predicted near the coasts. However, contrary to PO-GAM, this model predicted dolphins in the western English Channel, with a concentration of individuals around the Channel Islands (Appendix A.5 of Annex A). Regarding ZIP-GLM, densities were also predicted in offshore waters, approximately 5 and 10 individuals.km<sup>-2</sup> and similar to the other models, but a larger patch was identified and located west of the Isle of Wight with prediction of up over 2,000 individuals.km<sup>-2</sup> (Fig. 3.4; Appendix A.5 of Annex A). Similarly to PO-GAM, high predicted densities of NB-GAM and TW-GAM were associated with high uncertainties outside the surveyed area but low uncertainties within the survey area. For ZIP-GAM and ZIP-GLM, patches of high densities in the survey area were associated with high uncertainties, making the predictions less reliable (Appendix A.6 of Annex A).

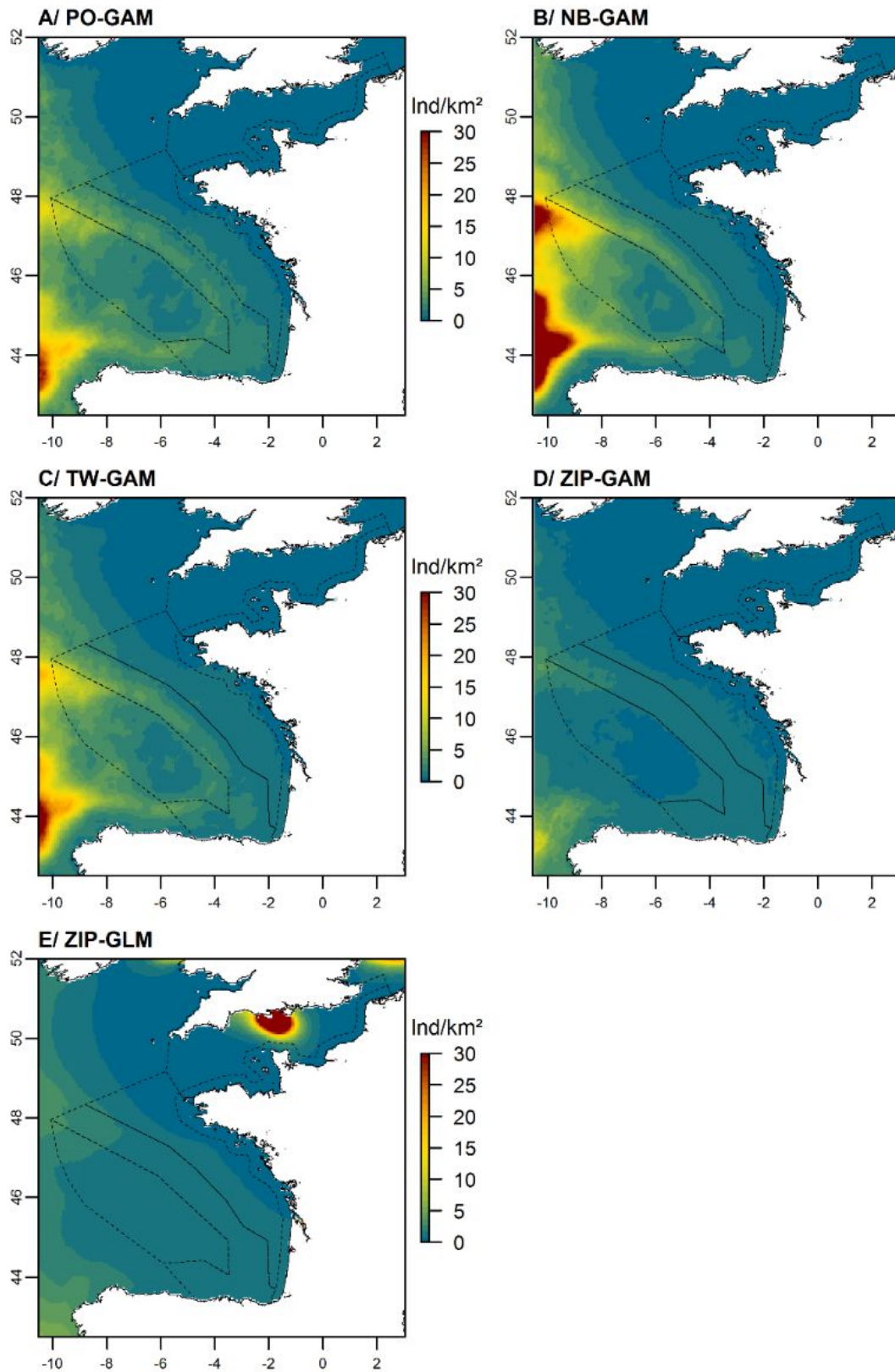
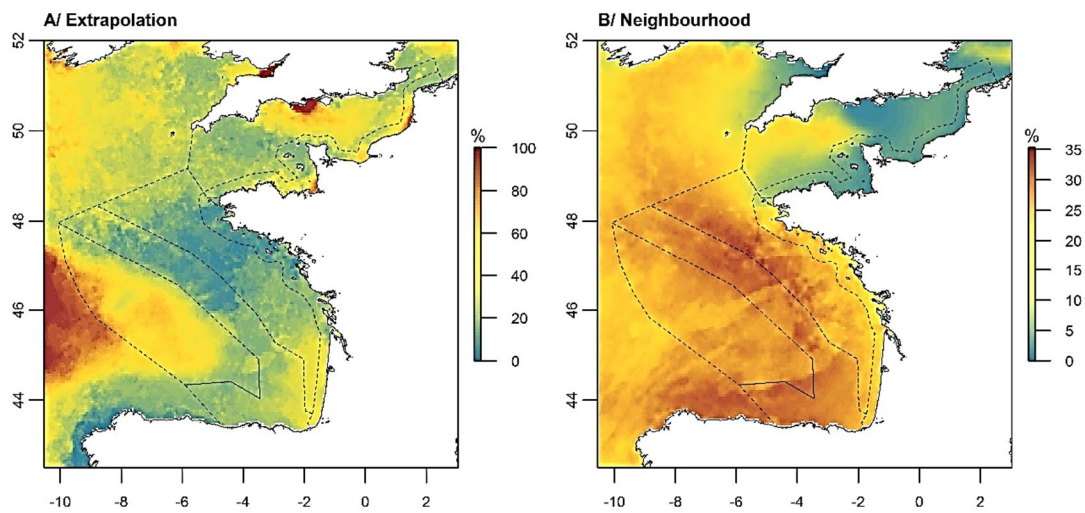


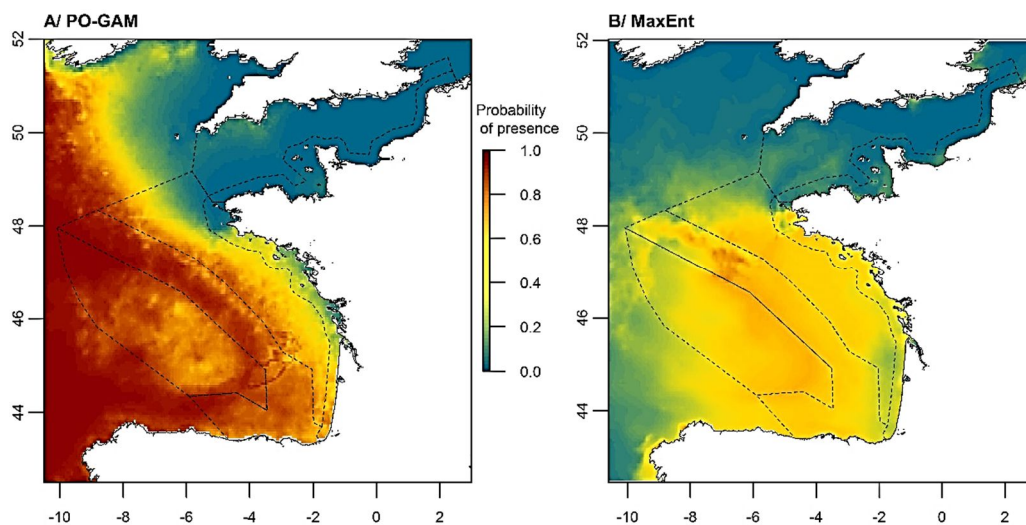
Fig. 3.4. Predicted distributions of small delphinids in individuals- $\text{km}^{-2}$  ( $\text{Ind}/\text{km}^2$ ) for each count-based model in the Bay of Biscay and the English Channel. Dotted lines represented the survey area. The scale of the PO-GAM was applied on all maps to facilitate the comparison, and Appendix 5 shows the maps with their own scale.

Extrapolation and neighbourhood maps (Fig. 3.5) allowed to assess the reliability of the predictions obtained with the fitted models. Overall, a high percentage of extrapolation and a low percentage of calibration data used to inform neighbouring cells (neighbourhood) indicated unreliable predictions. Hence, a model that predicted densities in the English Channel was inconsistent, which was particularly the case with ZIP-GLM and to a lesser extent, the case with ZIP-GAM. PO-GAM, NB-GAM and TW-GAM all predicted high densities outside the surveyed area, but according to Fig. 3.5, these predictions were fairly reliable because they were informed by approximately 20% of the data used to calibrate the model. However, NB-GAM made more extreme extrapolations than other models in the Bay of Biscay.



**Fig. 3.5. Extrapolation analysis using Gower's distance** (King and Zeng 2007). The extrapolation map (A) assesses whether a prediction was an extrapolation (100%) or an interpolation (0%) and the neighbourhood map (B) represents the percentage of calibration covariate data which informed each cell.

The MaxEnt model predicted higher probabilities of occurrence in the Bay of Biscay, particularly over the slope but also fairly evenly spread along the coasts of the Bay of Biscay and southwest England (Fig. 3.6). Compared to PO-GAM (see prediction map in probability of presence, Fig. 3.6), MaxEnt hardly geographically extrapolated beyond the sampled area and the predicted probabilities of presence were lower.



**Fig. 3.6. Distributions predicted by PO-GAM (A) versus MaxEnt (B) in the Bay of Biscay and the English Channel.** Dotted lines represent the survey area. The same scale was applied for the two model to facilitate the comparison.

### 3.3.2 Evaluation and comparison of the dolphin models

NB-GAM showed the lowest AIC and was followed by TW-GAM, whereas PO-GAM showed the highest AIC (Table 3.3). The explained deviances varied between 7.3% for ZIP-GLM and 39.1% for TW-GAM (Table 3.3). In addition, the lowest null and residual deviances were computed for NB-GAM, which indicated a better fit of the model compared to TW-GAM that, despite a high explained deviance, showed very high null and residual deviances (Table 3.3). ZIP models performed poorly compared to NB-GAM or TW-GAM but better than PO-GAM. Overall, NB-GAM showed a better predictive performance than other models.

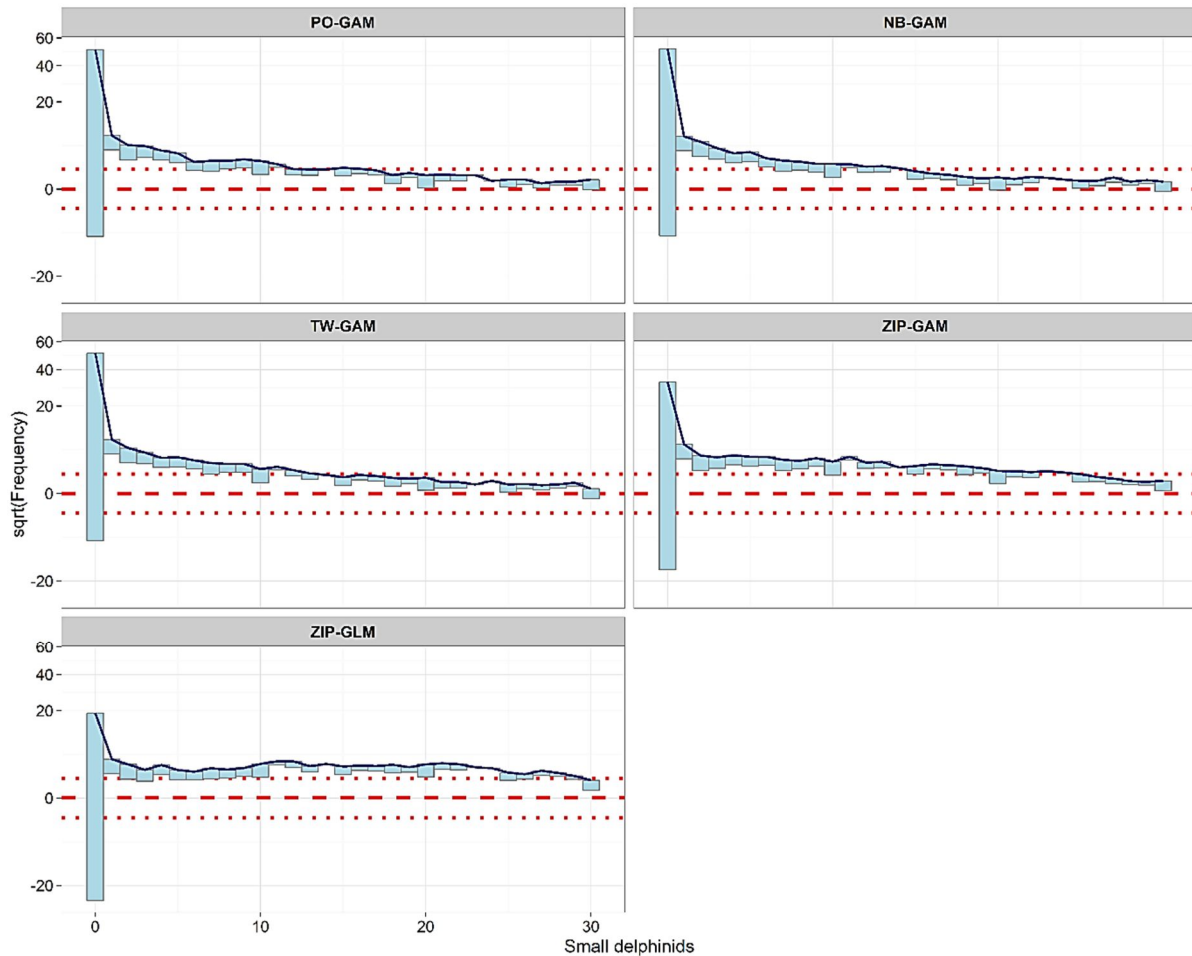
**Table 3.3. Indices used for the comparison of the count-based models.** AIC: Akaike Information Criterion.

	PO-GAM	NB-GAM	TW-GAM	ZIP-GAM	ZIP-GLM
AIC	70,082	4,284	5,869	25,438	28,001
Explained deviance	28.6 %	38.4 %	39.1 %	17.1 %	7.3 %
Null deviance	96,266	1,001	41,341	27,305	27,146
Residual deviance	68,742	616	25,168	22,635	25,164

In rootograms (Fig. 3.7), bars represent the observed frequencies and solid lines represent the fitted frequencies. To indicate a good fit of the models, bars have to be included into the confidence intervals represented by the red dot lines. The blue bars only have to intersect the confidence intervals to consider the fitted model as adequate. According to these rootograms, none of the fitted models was adequate for counts between 0 and 5 individuals. In all cases, the observed frequencies were not included in the confidence intervals. PO-GAM, NB-GAM and TW-GAM did not predict enough zeros and predicted too many sightings from 1 to 5 individuals. Likewise, ZIP-GAM and ZIP-GLM did not predict enough zeros and predicted too many sightings beyond 5 individuals. Albeit all fitted models tended to over-predict the frequencies between 1 and 5 individuals, the highest number of observed frequencies included in the confidence intervals was observed for NB-GAM thus making it the best fitted model.

## 3.4 STAGE 2: HOW MANY SIGHTINGS TO MODEL RARE SPECIES DISTRIBUTIONS?

From the first stage, I selected the best performing models, *i.e.* NB-GAM and TW-GAM, to conduct the sighting thinning experiment. I also kept ZIP-GAM because, despite its poor performance in the first stage (maybe because the dataset was not truly zero-inflated), I assumed a better performance with increasingly thinned datasets (true zero-inflation, since positive sightings are replaced by zeros). I also tested the MaxEnt model in the sighting thinning procedure to test the predictive performance of a presence-only model compared to count-based models when the data become rarer.



**Fig. 3.7. Rootograms obtained for each model.** The x-axis is the number of small delphinids and the y-axis represents the frequencies are in a square-root scale. Bars represent the observed frequencies, and solid lines represent the fitted frequencies. Red dot lines represent the confidence intervals, in which the blue bars have to be included to indicate a good fit of the models. The blue bars only have to intersect the confidence intervals to consider the fitted model as adequate.

### 3.4.1 Model selection and predictions of the auk baseline models

The explained deviances in the auk dataset reached 44.9% for NB-GAM, 40.9% for TW-GAM and 33.6% for ZIP-GAM. The variables selected by the three baseline models were depth, mean and gradient of SST and mean SSH (Fig. 3.8). Greater auk densities were associated with colder and shallower waters, stronger gradients of temperature and higher positive altimetry. The predicted distribution ranged from the coast to the edge of the continental shelf and predicted densities were particularly high in the eastern English Channel (Fig. 3.8). There was a good match between the sightings and the predictions of the model (Figs. 3.1 and 3.8) with high predicted densities associated with low coefficients of variation (Appendix B.5 of Annex B). The MaxEnt model, with an AUC of 0.842, generally predicted the same distribution as the other models with higher concentrations along the coast (Fig. 3.8).

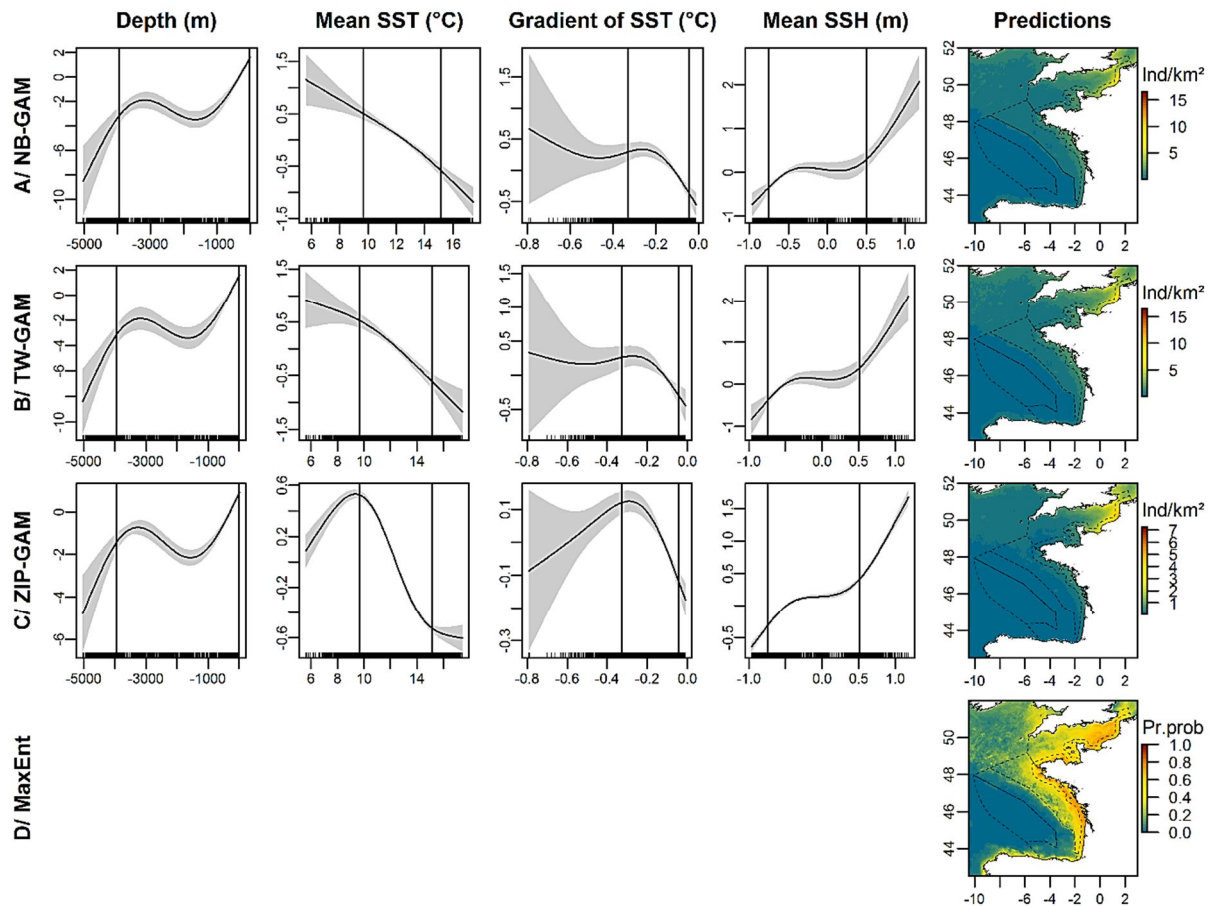


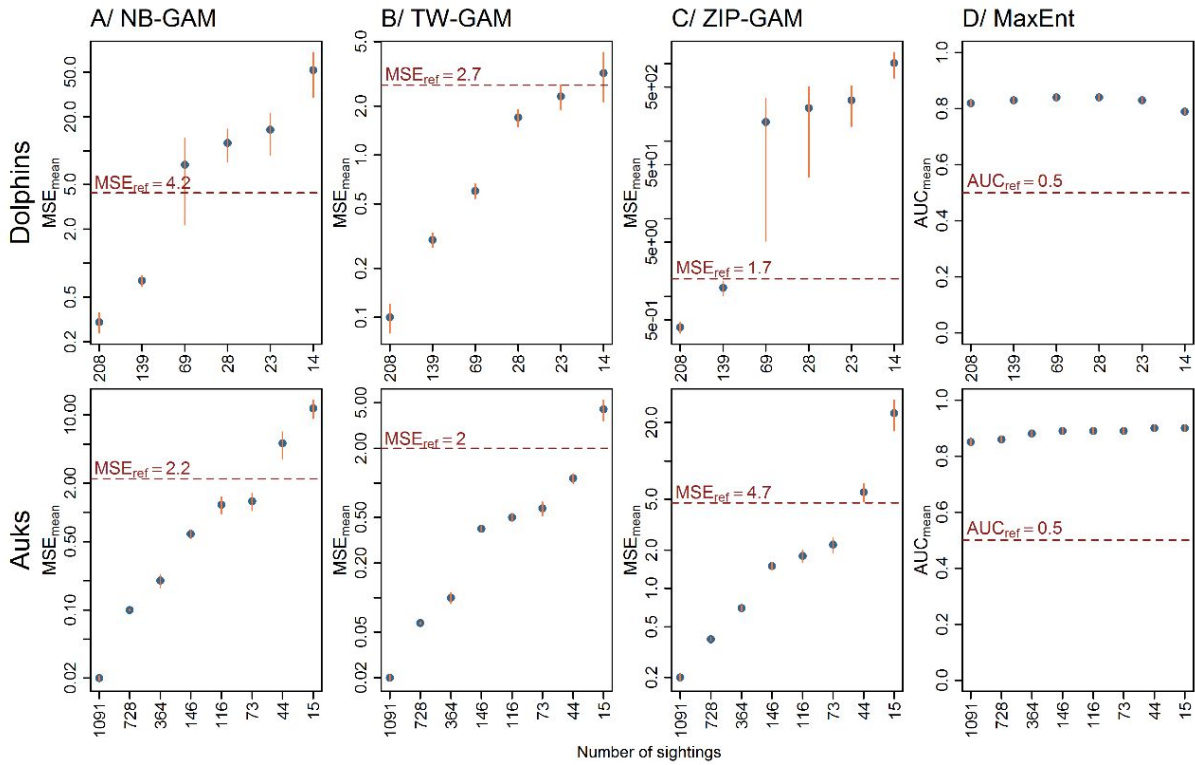
Fig. 3.8. Functional relationships for the selected covariates of the baseline models and the predicted distribution of auks in individuals.km<sup>-2</sup> (Ind/km<sup>2</sup>) for each count-based model and in presence probabilities (Pr.prob) for the MaxEnt model. The solid line in each plot is the estimated smooth function, and the shaded regions represent the approximate 95% confidence intervals. The y-axis indicates the number of individuals on a log scale, and a zero indicates no effect of the covariate. The best model fits are between the vertical lines indicating the 10<sup>th</sup> and 90<sup>th</sup> quantiles of the data. The black lines represent the bathymetric strata of the survey area.

### 3.4.2 Predictive performance of the experimental thinned models

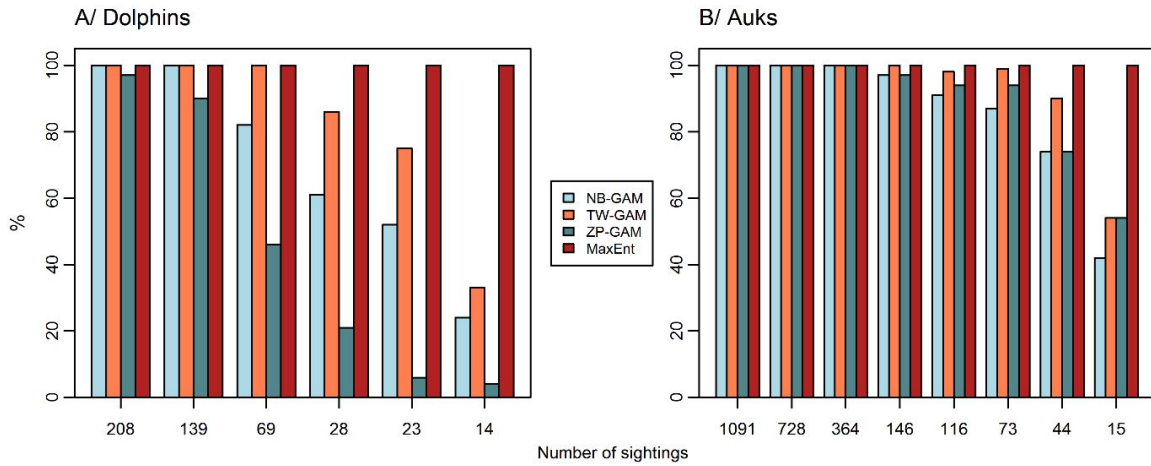
#### Small delphinids

As expected, a decrease in the number of sightings led to an increase in  $MSE_{mean}$  (Fig. 3.9). Predictions with 208 sightings (the lowest thinning rate) were closer to those of the baseline models than the predictions with only 14 sightings (the highest thinning rate). The comparison of  $MSE_{mean}$  with  $MSE_{ref}$  (representing the MSE between the baseline predictions and the null models), suggested that for less than 139 sightings,  $MSE_{mean}$  values for NB-GAMs and ZIP-GAMs were higher than  $MSE_{ref}$ . In contrast,  $MSE_{mean}$  values for TW-GAMs were lower than  $MSE_{ref}$ , except for the most extreme thinning rate that yielded as few as 14 sightings. Consequently, below 139 sightings, it was better to predict a homogeneous spatial distribution rather than to use the predictions provided by NB-GAMs and ZIP-GAMs. For TW-GAMs, this threshold was under 23 sightings. Furthermore, the number of experimental models in which the MSE was higher than the  $MSE_{ref}$  varied among model types (Fig. 3.10). With a decrease in the number of sightings, the proportion of experimental models in which predictions were better than a homogeneous spatial distribution decreased ( $MSE < MSE_{ref}$ ; Fig. B.6). For example, with 23 sightings, only 51% NB-GAMs and 6% ZIP-GAMs predicted better than a homogeneous spatial distribution compared to 75% TW-GAMs. For MaxEnt,  $AUC_{mean}$  values of the experimental models were

high ( $> 0.82$ ) and very similar and higher than the  $AUC_{ref}$ , which predicted a homogeneous distribution of the sites occupied by the species.



**Fig. 3.9. Evaluation of the predictive performance of the models using MSE and AUC.**  $MSE_{mean}$ : mean squared error averaged over 100 models;  $AUC_{mean}$ : area under the curve averaged over 100 models; Ref: reference index (*i.e.* a homogeneous spatial distribution). A log scale is applied on the y-axis. The vertical bars on each point represent the standard error calculated from 100 models.



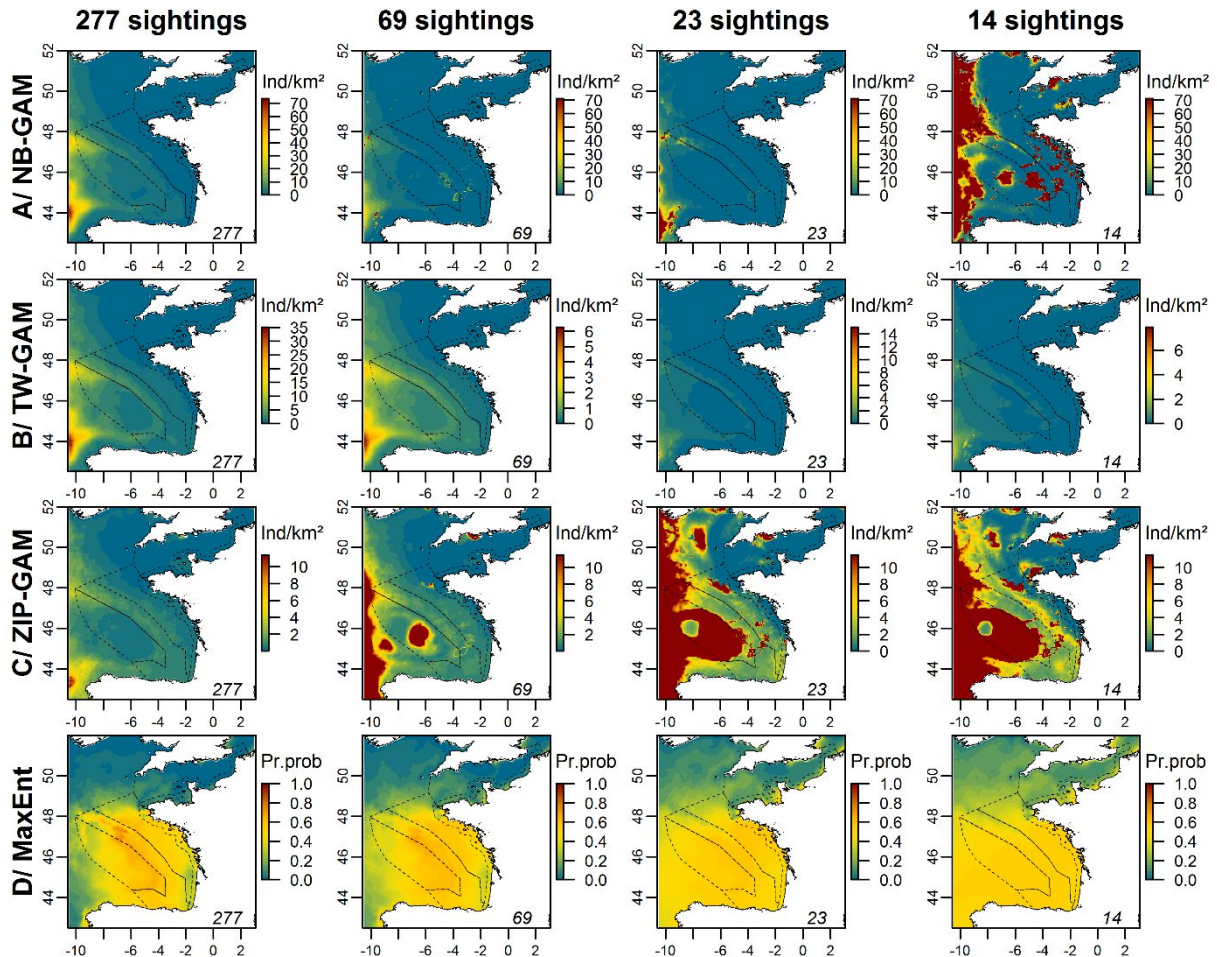
**Fig. 3.10. Proportion of experimental models better than a homogeneous spatial distribution.** Each bar represents the proportion of the experimental models out of the 100 fitted in which the MSE is lower than the  $MSE_{ref}$  for each number of sightings, *i.e.* the model that is better than a homogeneous spatial distribution. Each colour represents a different model type.

I noticed an important variation in the prediction maps among experimental models (Fig. 3.11; Appendices B.6 and B.7 of Annex B). Despite a decrease in the number of sightings, the distribution patterns of the baseline models were maintained down to 139 sightings for NB-GAMs and ZIP-GAMs. Beyond this threshold, the pattern disappeared or became unrealistic. Predictions from TW-GAM were similar to the distribution pattern of the baseline model with as few as 28 sightings. Beyond this



threshold, the pattern began to fade out. When compared to the baseline, the highest densities predicted by NB-GAM, TW-GAM and ZIP-GAM were associated with the highest uncertainties (Appendices B.8 and B.9 of Annex B).

Presence probability predicted by MaxEnt model became more uniform in area when the number of sightings decreased, with high probability areas located along the slope, and low probability areas located near the Aquitaine coast gradually fading out (Fig. 3.11).



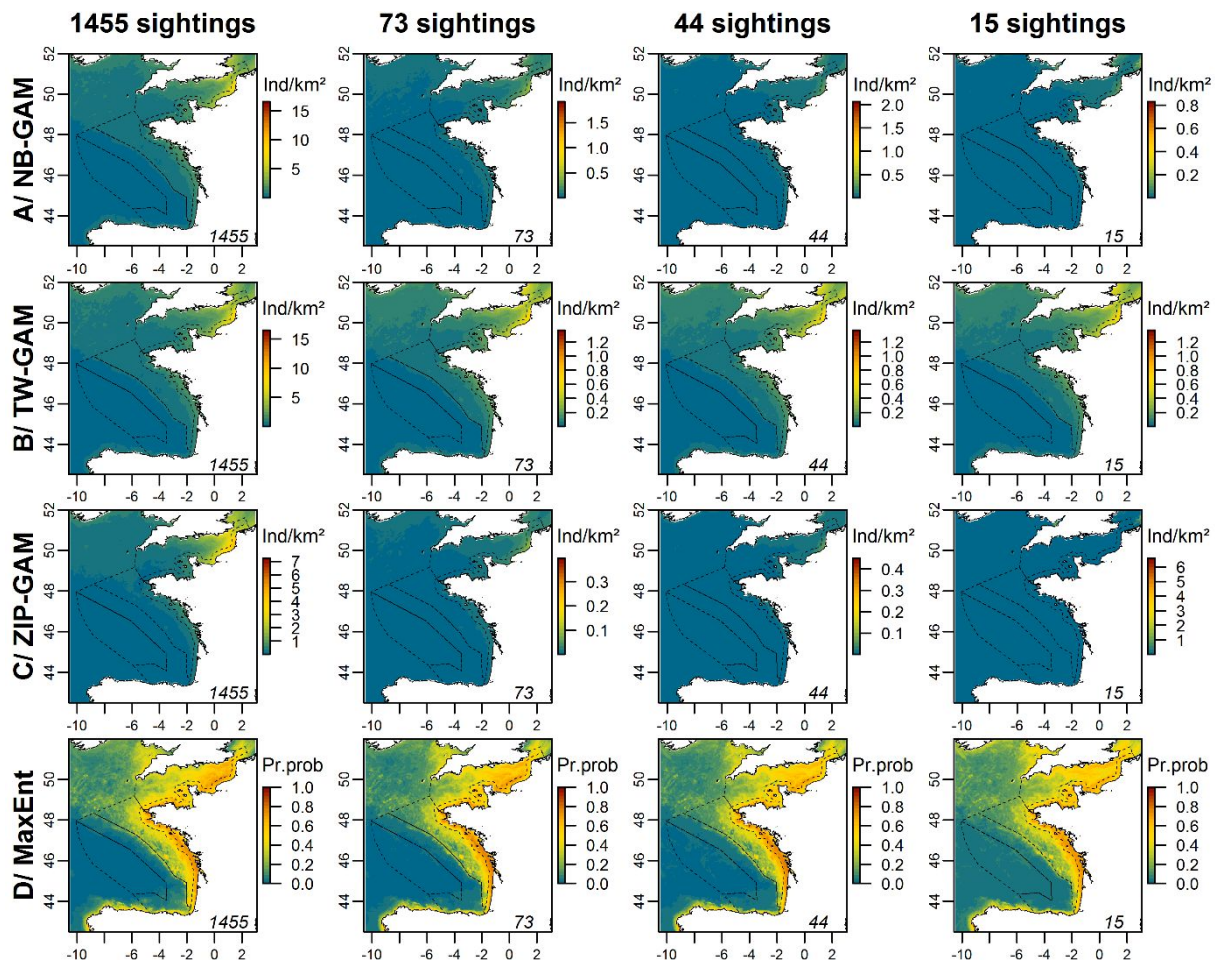
**Fig. 3.11.** Prediction maps of dolphins averaged over 100 models fitted to thinned datasets for each type of model in the Bay of Biscay and the English Channel. The rows represent the different types of generic models, and the columns represent the number of sightings used to fit the models. The numbers in the right corner of each map represent the number of sightings used to fit the model. The scale is in individuals.km<sup>-2</sup> (Ind/km<sup>2</sup>) for the NB-GAM, the TW-GAM and the ZIP-GAM and in the probability of presence (Pr.prob) for MaxEnt. This figure only shows the results for which a change was observed compared with the other predictions. All maps are presented in S5 and S6 Figs. The black lines represent the bathymetric strata of the survey area.

## Auks

MSE<sub>mean</sub> values increased with decreasing numbers of sightings (Fig. 3.9). As expected, predictions with 1,091 sightings (the lowest thinning level) were closer to those of the baseline model (1,455 sightings) than were the predictions with only 15 sightings (the highest thinning level). When the number of sightings was lower than 73, MSE<sub>mean</sub> values of NB-GAMs and ZIP-GAMs were higher than MSE<sub>ref</sub>, whereas MSE<sub>mean</sub> for TW-GAMs was higher than MSE<sub>ref</sub> with only 15 sightings. Consequently, with less than 73 sightings, the predictions provided by NB-GAMs and ZIP-GAMs were worse than a homogeneous spatial distribution. For TW-GAMs, this threshold was below 44 sightings. Similar to the results for dolphins, the number of models in which the MSE was higher than the MSE<sub>ref</sub> varied (Fig.

3.10). With 15 sightings, only 42% NB-GAMs compared to 54% TW-GAMs and ZIP-GAMs predicted better than a homogeneous spatial distribution. The  $AUC_{mean}$  values for the MaxEnt model were very high ( $>0.85$ ) and slightly increased with a decreasing number of sightings. Overall, the  $AUC_{mean}$  values were higher than  $AUC_{ref}$  (Fig. 3.9).

I noticed clear distinctions in averaged prediction maps between experimental models (Fig. 3.12; Appendices B.10 and B.11 of Annex B). For NB-GAMs, the prediction patterns were maintained down to 116 sightings, but under this threshold, patterns gradually disappeared. Despite a decrease in predicted densities, the distribution patterns predicted by the TW-GAMs remained the same down to 15 sightings. The distribution patterns predicted by ZIP-GAMs progressively disappeared below 364 sightings.



**Fig. 3.12.** Prediction maps of auks averaged over the 100 models fitted to thinned datasets for each type of model in the Bay of Biscay and the English Channel. The rows represent the different types of generic models, and the columns represent the number of sightings used to fit the models. The numbers in the right corner of each map represent the number of sightings used to fit the model. The scale is in individuals.km<sup>-2</sup> (Ind/km<sup>2</sup>) for the NB-GAM, the TW-GAM and the ZIP-GAM and in the probability of presence (Pr.prob) for the MaxEnt model. This figure only shows the results for which a change was observed compared with the other predictions. All maps are presented in Appendices B.10 and B.11 of Annex B. The black lines represent the bathymetric strata of the survey area.

Highest densities predicted by NB-GAMs, TW-GAMs and ZIP-GAMs fitted to thinned datasets were associated with lower uncertainties (Appendices B.12 and B.13 of Annex B). Furthermore, uncertainties of TW-GAMs were lower than those of NB-GAMs and ZIP-GAMs. The MaxEnt models showed some homogenisation of the distribution patterns with a decreasing number of sightings, but the general pattern was maintained irrespective of the number of sightings (Fig. 3.12).

## 3.5 GENERAL CONSIDERATIONS

In this chapter, I compared different types of habitat models, particularly count-based and presence-only models, to determine which would be the most suitable for scarce dataset of a rare species. In a first step, I found that GAM with a negative binomial distribution and GAM with a Tweedie distribution were the most appropriate model to fit the data. In contrast and fairly unexpectedly, the zero-inflated Poisson distributions showed less convincing results. I also found that MaxEnt provided fairly good results compared to PO-GAM. In a second step, I assessed the predictive performance of some of previous tested models using a reduced amount of available data. Findings suggested that the habitats for species that are rare or seldom seen are best described using a GAM with a Tweedie distribution (if effort data are available). GAMs with a negative binomial or zero-inflated Poisson distribution and MaxEnt models became inadequate for dataset under 130 sightings while TW-GAMs kept performing well down to a sample size of 30 sightings.

### 3.5.1 Biological systems

Dolphins and auks were used as biological models for two reasons. First, sightings of these taxa were large enough to allow proper statistical analyses and thinning to be conducted (277 dolphin sightings and 1,455 auk sightings). Second, dolphins and auks in the Bay of Biscay show well-defined and distinct patterns of distribution (Lambert et al. 2017a), which allows evaluating the predictive accuracy of the models.

Species groups pooled different species because of the difficulty to distinguish individuals at a species level from air. Pooling species into groups probably create categories with a broader habitat than the habitat of any of the constituting species, resulting in slightly larger sample size recommendations. However, the auk taxon is mainly dominated by the common guillemot and distribution patterns obtained in the study would mainly represent the common guillemot winter distribution. Indeed, auks wintering in the Bay of Biscay mostly originate from colonies located in the British Isles, where breeding populations of razorbills amount to 187,000 individuals, Atlantic puffin to 580,000 individuals and common guillemot to 1,416,000 individuals (Mitchell et al. 2004). Concerning dolphins, combining the two species resulted in a bimodal habitat. Indeed, shipboard surveys (CODA; partly SCANS-II and SCANS-III) have shown that if the two species are present in all offshore habitats, the common dolphin would predominate over the shelf and the shelf-break, whereas the striped dolphin would be more frequent in oceanic waters. Consequently, this species complex would reflect some kind of bimodal habitat characteristics that can be found *in natura*, as for instance in some delphinids like the bottlenose dolphin (*Tursiops truncatus*) with its pelagic and coastal ecotypes (Shirihai and Jarrett 2006).

Auks and dolphins differ widely in their habitat specificity, particularly regarding depth, an environmental variable of major importance to characterise marine habitats. Hence, the sighting thinning experiment conducted in both taxa simulated two different cases of rarity (*cf.* Chapter 1; Rabinowitz 1981). Thinning small delphinid sightings simulated a rare non-specialist species living in a broad habitat while thinning auk sightings generate a rare specialist species living in a narrower habitat. Modelling the habitat of the species characterised by locally high population sizes is not challenged by the number of sightings as the species is locally abundant, but is challenged by the location of the survey (if the survey was outside the core distribution of a species, sighting data would be scarce).

Consequently, only habitat modelling for rare species characterised by locally low population remain an issue. To provide a more complete answer regarding the sample sizes needed to characterise pelagic animal distributions, further analyses and meta-analyses with multiple and diversified datasets should be conducted to obtain robust recommendations. I am aware that determining the number of data needed to model the rare species habitats is an important challenge, for example to inform field efforts, but that a single study cannot consider all possible cases. An alternative research avenue would be to use virtual species instead of real species (Zurell et al. 2010), which would allow to control all the conditions of the procedure but would not reflect the complex reality of the ecosystems. A methodology can work with a virtual species but fail in a real case.

### 3.5.2 Reference and Baseline models

In this study, I chose a panel of eight environmental predictors, both static and dynamic, considered as proxies for primary production and consequently prey distribution (Austin 2002). In the first stage, to compare the habitat models of small delphinids, I used PO-GAM as a reference model because it characterises equi-dispersed data. Equi-dispersion is expected in the idealised situation where each detection event is independent of the others. The reference model showed a relatively high explained deviance (28.6%), interpolated more often than it extrapolated (Fig. 3.5) and was ecologically consistent with previous studies (Cañadas et al. 2009; MacLeod et al. 2009; Murphy et al. 2013; Lambert et al. 2017a), which strengthened model choice.

The choice of the challenger models was an important step in the comparison process. The aim was not to test all existing models but to answer the pragmatic question: “What type of model should I use if the dataset contains more than 90% zeros?” To answer this question, I built realistic models that included linear (GLM) or non-linear (GAM) relationships between the response variable and predictors and tested different structural choices for the data likelihood (Tweedie, negative binomial and ZIP). All these models handled differently the datasets with extra zeros (Appendix A.2 of Annex A). A ZIP model links two sub-models: a binomial model for the zero count that distinguishes between true and false absences, and a Poisson count model for non-zero observations. Conditional on an observation not being a true absence, equi-dispersion is assumed. True absences are in this case the only source of over-dispersion. Tweedie and negative binomial models directly include an assumption on the relationship between the mean and the variance, in that they address over-dispersion in a more phenomenological way because the micro-level process generating over-dispersion is not explicit (Dunn and Smyth 2005; Ridout et al. 1998; Zeileis et al. 2007, Wenger and Freeman 2014).

In the two stages, variable selection was only performed on the reference model (stage 1) and the baseline models (stage 2). As the performance of a model is largely controlled by its selected variables (Syphard and Franklin 2009), the models that used thinned-out sightings might be biased and are suboptimal (because some sighting data are ignored). Indeed, variables selected by a model fitted to few data could differ from models fitted on much larger datasets. I decided to hold the set of covariates constant over models to assess how the results were affected by the structural choice in the model only in the first stage and the number of sightings in the second stage. In the second stage, I did not attempt to find the best model fit but to test the robustness of model predictions to thinning; variable selection was, to a certain extent, secondary to our purposes. In an ideal situation, the habitat of the species is known a priori. In practice, this is rarely the case, but in realistic situations, a SDM is first developed and then used repeatedly until the need to update it becomes an imperative. Thus, the same SDM

specification may be used without undergoing rounds of variable selection each time a new datum is added to an existing dataset. In a similar fashion, while MSE give guarantee on the predictive performance on average (*i.e.* under repeated use of the same model with different data generated from the same process), more often than not a single dataset is available for a given area. Consequently, to approximate a real situation in which one needs to model rare species habitats from a single dataset, the predictive capacity of each experimental model has been assessed in order to determine the probability for a single experimental model to reproduce the baseline model predictions.

### 3.5.3 Thinning-out sighting data

The thinning rates applied in this study were arbitrarily determined to obtain, in the most extreme scenario, as few as 15-20 sightings, close to the numbers of deep-diver sightings commonly available in a large survey dataset, like SCANS or CODA for instance (Rogan et al. 2017). Overfitting can be an issue with small datasets, *i.e.* the selected model becomes too complex compared to the number of implemented sightings (Hawkins 2004; Subramanian and Simon 2013). Particularly, overfitting could have occurred in the models with the highest thinning rates. Nonetheless, NB-GAM, TW-GAM and ZIP-GAM performed differently with the same low number of sightings (14-15 sightings). NB-GAM and ZIP-GAM did not manage to predict consistent distribution patterns compared to the baseline models, whereas the TW-GAM did.

## 3.6 PREDICTING HABITATS OF RARE SPECIES

Although environmental variables used in the models were identical, each fitted model showed a different predictive performance based on its own characteristics. Overall, NB-GAM and TW-GAM were very similar in the improvement they provided over PO-GAM (Appendix A.3 of Annex A) and estimated similar non-linear relationships with environmental covariates. NB-GAM exhibited the best predictive performances with the smallest AIC and a moderate explained deviance. Habitat predictions obtained from PO-GAM, TW-GAM and NB-GAM were qualitatively similar, suggesting robustness with respect to extrapolation (Fig. 3.5) and consistency in the results. However, predicted densities were larger in magnitude with NB-GAM. The overall poor performance only found for GLM among all candidate generic models stressed the importance of non-linear relationships in habitat modelling of small delphinids in the Bay of Biscay, and potentially all mobile megafauna. Thanks to their flexibility, GAMs are appropriate for modelling the distribution of sparsely distributed megafauna either marine or terrestrial (Wood 2006a; Becker et al. 2010; Hegel et al. 2010). Thus, NB-GAM and TW-GAM were able to fit the data well despite huge numbers of zeros, as seen on the rootograms (Fig. 3.7). However, with the sighting thinning procedure, the NB-GAM provided less convincing results and unreliable predicted distribution patterns compared to the baseline model under approximately 130 sightings.

Due to the combination of a zero-inflated model and the non-parametric functions of a GAM (Barry and Welsh 2002), I expected a better performance of the ZIP-GAM. Fitted ZIP models showed lower explained deviances (17.1% for ZIP-GAM and 7.3% for ZIP-GLM), lower predictive performances (higher AIC) and less ecologically consistent predictions. Indeed, ZIP-GAM and ZIP-GLM predicted large densities of small delphinids in the English Channel (where no sightings were recorded) while previous studies showed that these species generally avoid this area (Cañadas et al. 2009; MacLeod et al. 2009; Murphy et al. 2013; Lambert et al. 2017a). In addition, the sighting thinning results were less convincing than

those obtained with a TW-GAM. Below approximately 130 sightings, the predicted distributions of the ZIP-GAM were unreliable compared to the predictions of the baseline model, whereas this threshold was as low as approximately 30 sightings for the TW-GAM. This is likely due to the current parametrisation of the ZIP family in the 'mgcv' package. In fact, the current parametrisation uses the linear predictors and linearly scales them on a logit scale to generate extra-zero observations (see the help pages in 'mgcv' v1.8-9; Wood 2013). This parametrisation implicitly assumes that areas with lower densities have a higher probability of non-detection. However, it does not allow for incorporating detection-specific covariates which may better explain non-detection patterns. Although 92% of the data were zeros, ZIP models showed worse results than NB-GAM or TW-GAM; hence over-dispersion was likely not mainly due to zero-inflation. Even though the best model selected here did not completely accommodate all zero observations, which suggests some zero-inflation (Fig. 3.7), but less prevalent than initially thought.

In stage 1, MaxEnt showed a fairly high predictive performance (AUC = 0.82) and distribution patterns relatively similar to those of PO-GAM. However, this model underestimated the probabilities of presence compared to PO-GAM and did not extrapolate beyond the study area. With the data-thinning experiment, all distributions predicted by the MaxEnt model were better than a homogeneous spatial distribution. Yet a gradual homogenisation of the probabilities of predicted dolphin presence over the whole area was shown, to such a point that with very few sightings (approximately 28) the model was no longer able to distinguish key areas of either high or low presence probabilities. In contrast, despite some degree of homogenisation of the predicted probabilities, the auk distribution patterns kept being correctly predicted even with as few as 15 sightings. Consequently, sighting thinning affected the performance of the models differently depending on the taxon. Thus, this presence-only model appeared relatively efficient to establish distribution patterns in a given survey area (Tsoar et al. 2007) and to identify areas of high probabilities of presence when only presence data were available (Zaniewski et al. 2002; Gormley et al. 2011). However, this conclusion regarding the performance of MaxEnt might be too optimistic and not completely representative of the actual performance of the model. Indeed since data were collected with a standard protocol that followed a regular sampling design across the Bay of Biscay, the obtained dataset conformed well to the assumptions associated with the appropriate use of presence-only models. Indeed, the main issue with presence-only models is that they cannot account for uneven effort and therefore assume that the sampling of the habitat was random in order to interpret MaxEnt predictions correctly (Yackulic et al. 2013). However, MaxEnt did not provide satisfactory results for the higher thinning rates. Therefore, like most models, beyond a critical amount of data, the performance of the MaxEnt model is reduced, and its use for rare species is questionable.

In addition, even if the TW-GAM provided good results with approximately 20-25 sightings, the results were based on the averages of 100 fitted models, which resulted in smoothing the predictions. In practice, habitat modellers have only one dataset to analyse. Therefore, I assessed the individual performance of each experimental model by computing the number of models in which the MSE was higher than the  $MSE_{ref}$ . It appeared that with 20 sightings, approximately 50% TW-GAMs predicted better than a homogeneous spatial distribution of the two species groups whereas with 40 sightings, 90% experimental models provided reliable results (Fig. 3.10). Consequently, to obtain robust predictions, I considered necessary that all fitted models would predict better than a homogeneous spatial distribution; a minimum of 50 sightings would represent a conservative empirical threshold. Moreover, by examining the explained deviances for each experimental Tweedie model (results not

shown), I found that explained deviances of the experimental models fitted to 28 and 69 sightings for dolphins were good (30-50%). For the smallest number of data (15 and 23 sightings), the explained deviances were very high (>50%) which suggested an over-fitting of the data. For this reason, a use of datasets with less than 50 sightings was not recommended. However, this would only be valid for TW-GAM because with NB-GAM and ZIP-GAM, the threshold at which all experimental models would provide good results (better than a homogeneous spatial distribution) was higher than 100 sightings (Fig. 3.10).

### 3.7 RECOMMENDATIONS FOR PRACTITIONERS

Finally, as Warton (2005) warned, “many zeros does not mean zero-inflation” of the data, and even 92% zeros does not necessarily mean zero-inflation. Even for scarce species, it would be recommendable to first test over-dispersed models such as GAMs with Tweedie or Negative Binomial distributions before testing zero-inflated models. Obviously, the predictive performance of the model has to be assessed for the tested model. A useful visual method to assess whether a model adequately addresses many zeros is the rootogram (Minami et al. 2007; Kleiber and Zeileis 2016). In addition, this study provided a first answer to the question commonly asked by habitat modellers: “What model should be used when studying rare species?” If modellers only have presence data, MaxEnt could be used with great caution and preferably for rare species with restricted distributions. With effort data, I would recommend using a GAM with a Tweedie distribution and a minimum of 50 sightings, which is a conservative empirical measure.

# Chapter 4

---

## DEEP-DIVER HABITAT PREFERENCES

---



© Laura Hedon

### CONTENTS

---

4.1 CONTEXT AND OBJECTIVES .....	58
4.2 METHODOLOGY .....	58
4.2.1 Data origin.....	58
4.2.2 Data processing .....	60
4.2.3 Habitat modelling .....	63
4.2.4 Environmental space coverage gap analysis.....	63
4.3 WHICH DISTRIBUTION FOR THE DEEP-DIVERS? .....	64
4.3.1 Effective strip width.....	64
4.3.2 Beaked whales.....	65
4.3.3 Sperm whales.....	66
4.3.4 Kogiids.....	67
4.4 A BASIN WIDE APPROACH TO MODEL THE DISTRIBUTION OF RARE MARINE SPECIES.....	69
4.4.1 Methodological considerations.....	69
4.4.2 Large-scale deep-diver habitats .....	70

**T**HIS chapter describes the data-assembling methodology developed into a basin wide approach to model the habitat preferences of beaked, sperm and kogiid whales. This chapter is based on an article that is currently in preparation and is about to be submitted (**Annex C**).



## 4.1 CONTEXT AND OBJECTIVES

In Chapter 3, I determined a minimum threshold of 50 sightings needed to provide reliable predictions of species distribution through habitat models. However, as introduced in the previous chapters, low sighting rates are usually reported for deep-divers (Waring et al. 2001; Barlow et al. 2006; Kiska et al. 2007). Each individual survey can rarely provide sufficient sightings (about 10-15 sightings per survey) to model species habitats, particularly at a large scale; this difficulty represents a major challenge when it comes to inform conservation strategies. To address this issue, I merged datasets from different visual surveys conducted in the North Atlantic Ocean and the Mediterranean Sea to increase the available number of sightings and model habitats used by deep-diving cetaceans and thus, understand the environmental processes that drive their basin-wide distribution.

Data-assembling is often necessary to successfully model habitat preferences of cetaceans (Roberts et al. 2016; Mannocci et al. 2017b, Rogan et al. 2017) but requires methodological considerations. Because of the various protocols, platforms and observation heights used in the different surveys, species detection and data quality vary among surveys. Indeed, each survey does not collect the same information, particularly regarding observation conditions; some surveys record only Beaufort seastate while other surveys also record additional parameters that influence species detection, such as sun glare, cloud coverage or wave height. Consequently, the homogenisation of these different data may require levelling down to the coarsest commonalities across datasets leading to some level of data degradation. Moreover, because surveys are carried out in different years and seasons, spatial and temporal heterogeneity in the data could be an issue.

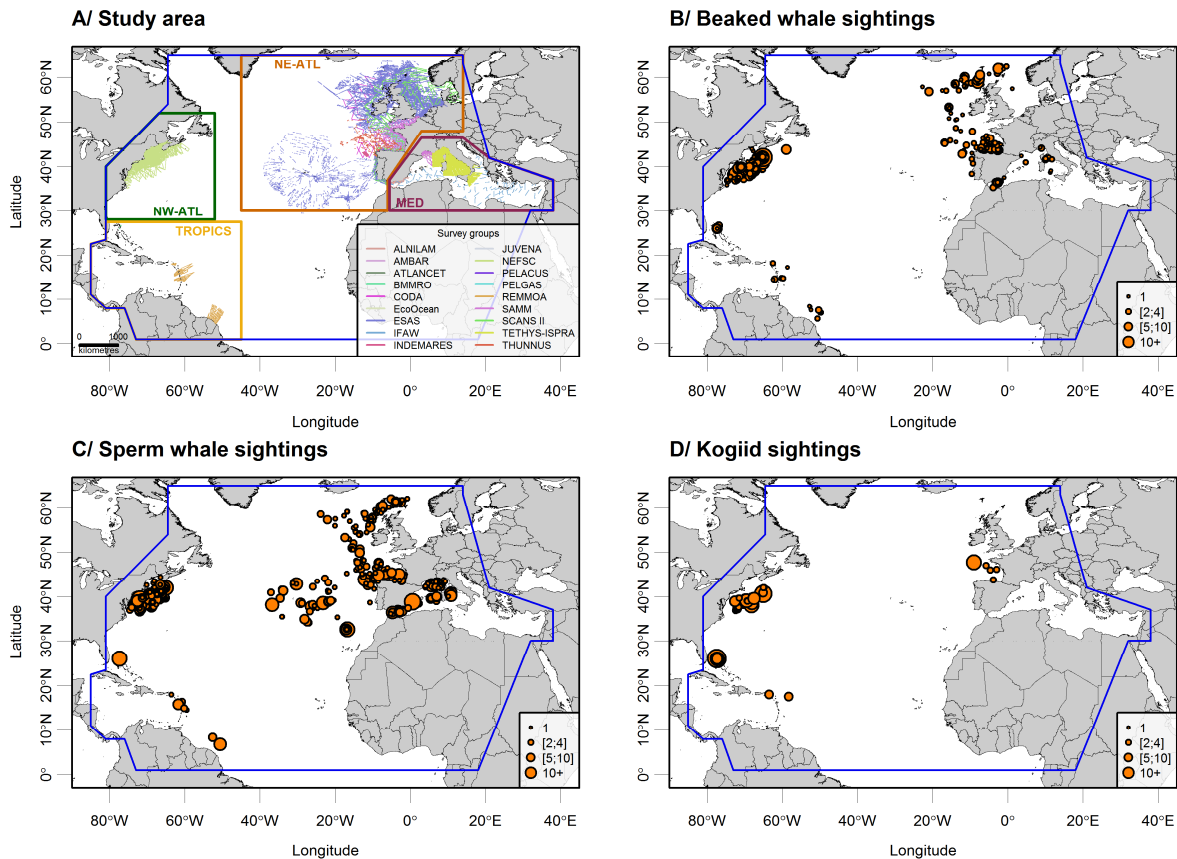
This study aims to model the habitats of deep-diving cetaceans at a large scale by assembling analogous datasets available from different surveys. In this work, I have aggregated cetacean visual survey datasets collected in the North Atlantic Ocean and the Mediterranean Sea. To take into account the various protocols, I implemented a meta-analysis of the detection process across platforms and observation conditions and modelled densities of the three groups of deep-diving cetaceans in a Generalised Additive Model framework. Finally, I performed an environmental space coverage gap analysis (Jennings 2000) to assess the reliability of the predictions outside the surveyed area. I thus produced the first basin-wide density maps for deep-diving cetaceans in the North Atlantic Ocean and the Mediterranean Sea.

## 4.2 METHODOLOGY

### 4.2.1 Data origin

As described in Chapter 2, the study area encompassed the North Atlantic and Mediterranean basins from the Guiana Plateau to Iceland, *i.e.* approximately from 1-65°N, excluding the Baltic, Red and Black Seas, the Gulf of Mexico and the Hudson Bay (Fig. 4.1A).

I aggregated visual shipboard and aerial surveys performed by 13 independent organisations in the North Atlantic Ocean and the Mediterranean Sea between 1998 and 2015 (details of the surveys in Appendix C.1 of Annex C). Cetacean sightings were recorded following line-transect methodologies that allow Effective Strip Width (ESW) to be estimated from the measurement of the perpendicular distances to the sightings (Buckland et al. 2015).



**Fig. 4.1.** The study area (A), and the beaked whale (B), sperm whale (C) and kogiid (D) sightings recorded during the surveys. The blue polygon delineates the study area. Surveys were carried out along transects (lines) following a line-transect methodology (details of the surveys in Appendix C.1 of Annex C). Sightings were classified by group sizes with each point representing one group of individuals and point size relating to the number of animals in a group.

A total of 630 sightings of beaked whales, 836 sightings of sperm whales and 106 sightings of kogiids, mainly distributed in the northeast and northwest Atlantic Ocean (north of the 35°N latitude), were assembled for the present study (Fig. 4.1B-D). Aggregated effort data represented about 1,240,300 km of on-effort transects (*i.e.* following a transect at specified speed/altitude with a specified level of visual effort) of which 58% was carried out by plane and the remaining effort by boats (Fig. 4.1A, Table 4.1). Only 9% of the effort was conducted under Beaufort seastate >4 and these data were removed from the analyses. Even if it is difficult to detect beaked whales and kogiids with a Beaufort seastate equal to 4, it was a trade-off between keeping a maximum number of data and limiting biases related to detection.

To account for differences between surveyed regions, four sub-regions were defined (Table 4.1): the northeast Atlantic Ocean (NE-ATL; from 40°W-10°E and 36°N-65°N), the northwest Atlantic Ocean (NW-ATL; from 80°W-55°W and 30°N-50°N), the tropics (from 80°W-45°W and 1°N-28°N) and the Mediterranean sea (MED; from 5°W-40°E and 30°N-46°N). Most sampling effort was performed in the northeast (37 %) and northwest (45 %) Atlantic Ocean. Mediterranean surveys represented only 16 % of the total sampling effort and surveys near the tropics represented only 2 %.

Encounter rates were calculated in each sub-region as:

$$(\text{number of encounters} / \text{total distance travelled}) * 100.$$

**Table 4.1. Effort performed by platform type or Beaufort seastate for all the surveys.** For the analyses, all segments with Beaufort seastate > 4 were excluded. NE-ATL: northeast Atlantic Ocean; NW-ATL: northwest Atlantic Ocean; MED: Mediterranean Sea.

Sectors	Total survey effort (km and %)	Total aerial effort (km and %)	Total shipboard effort (km and %)	Total effort by Beaufort seastate class (km and %)				
				[0-1]	]1-2]	]2-3]	]3-4]	]4-7]
NE-ATL	468,892	70,358	398,533	76,705	118,456	135,699	84,812	53,220
	37 %	15 %	85 %	16 %	25 %	30 %	18 %	11 %
NW-ATL	556,963	545,677	11,286	42,737	121,184	199,317	131,947	61,777
	45 %	98%	2 %	8 %	22 %	36 %	23 %	11 %
MED	195,440	86,930	108,510	92,126	69,882	26,649	5,984	799
	16 %	44 %	56 %	47 %	36 %	14 %	3 %	0.4 %
TROPICS	19,041	15,356	3,685	10,590	2,495	3,681	1,897	378
	2 %	81 %	19 %	56 %	13 %	19 %	10 %	2 %
<b>STUDY AREA</b>	<b>1,240,336</b>	<b>718,321</b>	<b>522,014</b>	<b>222,158</b>	<b>312,017</b>	<b>365,346</b>	<b>224,640</b>	<b>116,174</b>
		<b>58 %</b>	<b>42 %</b>	<b>18 %</b>	<b>25 %</b>	<b>30 %</b>	<b>18 %</b>	<b>9%</b>

#### 4.2.2 Data processing

##### Data-assembling

All survey datasets were standardised for units and formats (*e.g.* date, time and coordinates) and aggregated into a single common dataset. A specific coordinate projection encompassing the entire survey area was defined (Albers equal-area conic from <http://projectionwizard.org>). Effort data were linearized and discretised into 5 km segments using ArcGIS 10.3 (ESRI 2016) and the Marine Geospatial Ecology Tools software (Roberts et al. 2010). Because of the large disparity between aerial and shipboard surveys, aerial surveys had transect lengths of up to several tens of kilometres, while transects in shipboard surveys could be much shorter, the 5 km segment length was the value that best homogenised the various transect lengths of the different surveys. Finally, sightings were linked to the segments for each species group.

##### Environmental predictors

I used static and dynamic variables that are considered to influence the distributions of deep-divers (Table 4.2). All variables were resampled at a 0.25° resolution instead of a 5 km resolution that would match the 5 km effort segments, because of the very large size of the study area and the spatial resolution of the variables (Table 4.2). This implies that the same values of environmental variables are attributed to neighbouring segments.

Depth, slope and the surface of canyon and seamount habitats within each 0.25° cell are physiographic variables. Sea Surface Temperature (SST; mean, standard error and spatial gradients, calculated as the difference between the minimum and maximum SST values found in the eight pixels surrounding any given pixel of the grid), Sea Surface Height (SSH; mean and standard error) and Eddy Kinetic Energy (EKE; mean and standard error) are dynamic oceanographic variables related to the movements of water masses. Net Primary Production (NPP) is a biological variable used as a proxy of prey availability (Appendix C.2 of Annex C shows the maps of the averaged situation of the variables over the 18 years of surveys). Dynamic variables were computed at a monthly resolution, *i.e.* averaged

over the 29 days prior to each sampled day to avoid gaps in remote sensing oceanographic variables and to take into account the time-lag between an environmental condition and its effect on intermediate trophic levels (Jaquet 1996; Austin et al. 2006; Redfern et al. 2006; Cotté et al. 2009).

**Table 4.2. Candidate environmental predictors used for the habitat modelling.** All variables were resampled at a 0.25° resolution. A: Depth and slope were derived from GEBCO-08 30 arc-second database (<http://www.gebco.net/>); 30 arc-second is approximately equal to 0.008°. B: Surface per cell was calculated in ArcGIS 10.3 from the shapefile of canyons and seamounts provided by Harris et al. (2014). C: The mean, standard error and gradient of Sea Surface Temperature (SST) were calculated from the GHRSSST Level 4 CMC SST v.2.0 (Canada Meteorological Centre, <https://podaac.jpl.nasa.gov/dataset/CMC0.2deg-CMC-L4-GLOB-v2.0>). D: The Aviso ¼° DT-MADT geostrophic currents dataset was used to compute mean and standard deviation of Sea Surface Height (SSH) and Eddy Kinetic Energy (EKE; <https://www.aviso.altimetry.fr/en/data/products/sea-surface-height-products/global/madt-h-uv.html>). E: Net primary production (NPP) was derived from SeaWiFS and Aqua using the Vertically Generalised Production Model (VGPM; <http://orca.science.oregonstate.edu/1080.by.2160.8day.hdf.vgpm.m.chl.m.sst.php>).

Environmental variables and units	Original Resolution	Sources	Effects on pelagic ecosystems of potential interest to deep-divers
<b>Physiographic</b>			
Depth (m)	30 arc sec	A	Deep-divers feed on squids and fish in the deep water column
Slope (°)	30 sec arc	A	Associated with currents, high slope induce prey aggregation or enhanced primary production
Surface of canyons and seamounts in a 0.25° cell (km <sup>2</sup> )	30 sec arc	B	Deep-divers are often associated with canyons and seamounts structures; the variable indicates the proportion of this habitat in each cell
<b>Oceanographic</b>			
Mean of SST (°C)	0.2°, daily	C	Variability over time and horizontal gradients of SST reveal front locations, potentially associated with prey aggregations or enhanced primary production
Standard error of SST (°C)	0.2°, daily	C	
Mean gradient of SST (°C)	0.2°, daily	C	
Mean of SSH (m)	0.25°, daily	D	High SSH is associated with high mesoscale activity and enhanced prey aggregation or primary production
Standard deviation of SSH (m)	0.25°, daily	D	
Mean of EKE (m <sup>2</sup> .s <sup>-2</sup> )	0.25°, daily	D	High EKE relates to the development of eddies and sediment resuspension induce prey aggregation
Standard error of EKE (m <sup>2</sup> .s <sup>-2</sup> )	0.25°, daily	D	
Mean of NPP (mgC.m <sup>-2</sup> .day <sup>-1</sup> )	9 km, 8 days	E	Net primary production as a proxy of prey availability

### Effective Strip Width estimation

From sighting and effort data, I fitted a detection function to determine the ESW for each species group (Thomas et al. 2010; Buckland et al. 2015). The estimation of the ESWs was a key step in the data-assembling process to take into account heterogeneity in sighting conditions among segments in the models (Hedley and Buckland 2004). Even if I only considered line-transect survey data, protocols differed to some extent and datasets did not always provide the same information, in particular regarding the observation conditions. Some surveys recorded Beaufort seastate, cloud coverage, sun glare and subjective observation conditions while others only provided Beaufort seastate. Hence, Beaufort seastate was the only descriptor of observation conditions shared by all datasets. Consequently, the platform type, the observation heights and the Beaufort seastate were used as

covariates following the conventional distance sampling methodology (Marques and Buckland 2003; Buckland et al. 2015). In addition, I had not enough sightings to fit detection functions for each survey. Consequently, to take into account the various protocols, I performed a meta-analysis (Gurevitch et al. 2001; Higgins et al. 2009). Firstly, for each species group, truncation distance  $w$  was determined as the 95<sup>th</sup> percentile of the set of perpendicular distances; the 5% most distant sightings were discarded from the analysis. Then, I created classes to pool the different surveys; Classes of platform type (plane or boat), observation heights (e.g. 0-5 m; 5-10 m...) and Beaufort seastate (0-1; 1-2; 2-3 and 3-4). The meta-analysis was performed in R-3.3.1 (R Core Team 2016) in a Bayesian framework using JAGS version 4-6 and package 'rjags' (jags model available in Appendix C.3 of Annex C; Royle and Dorazio 2008; Plummer 2016). First, for each species group, perpendicular distances of all sightings were used to estimate a detection function with a hazard key.

For a sighting  $i$  made from survey  $s$  at height  $j$  in class of Beaufort seastate  $k$ , let  $d_{jks}^i$  denotes the perpendicular distance. The detection probability of sighting  $i$  is:

$$\begin{cases} p_{ijk}^s = g_s(d_{ijk}) = 1 - \exp\left(-\left(\frac{d_{ijk}}{\sigma_{jks}}\right)^{-\nu_s}\right) \\ \log(\sigma_{jks}) = \beta_{j0} + \beta_{j1} \times k + \alpha_s \end{cases}$$

where  $\beta_{j0}$  and  $\beta_{j1}$  are respectively random intercept and slope parameters for the effect of platform height; and  $\alpha_s$  and  $\nu_s$  are survey random effects. Bivariate random effects were specified with a Cholesky decomposition and using the priors for the Cholesky factors as Kinney and Dunson (2008). I used half Student-t distributions with 3 degrees of freedom and scale set to 1.5 as priors for dispersion parameters, and standard normal priors for all other parameters. Four chains were run with a burn in of 10,000 iterations, followed by another 10,000 iterations (with a thinning factor of 10). Parameter convergence was assessed with the Gelman-Rubin  $\hat{R}$  statistics. Posterior inferences are based on the pooled sample of 4,000 values (1,000 per chain).

The advantage of setting a hierarchical model to estimate detection functions is to borrow strength across the different datasets to increase the precision of estimates. For each combination of survey – platform type – observation height – Beaufort seastate, estimated detection functions are shrunk towards a common detection function (itself estimated from the data) according to the available data corresponding to this particular combination of survey – platform type – observation height – Beaufort seastate. If, for a given combination of parameters, there were few sighting data, the estimated detection function was very close to the common detection function, whereas if there were enough data, the estimated detection function could deviate from this common function. Upon model fitting and successful parameter estimation, the ESW for each combination of survey – platform type – observation height – Beaufort seastate was computed:

$$ESW_{jks} = \int_0^w g_s(x) dx = \int_0^w \left[ 1 - \exp\left(-\left(\frac{x}{e^{\beta_{j0} + \beta_{j1} \times k + \alpha_s}}\right)^{-\nu_s}\right) \right] dx$$

The posterior mean value of estimated ESW was then allocated to each segment with respect to species group, survey, platform type, seastate and observation height class.

### 4.2.3 Habitat modelling

From the results obtained in Chapter 3, to model deep-diver habitat preferences, I fitted Generalised Additive Models (GAMs; Hastie and Tibshirani 1986; Wood 2006b) with a Tweedie distribution to account for over-dispersion (Foster and Bravington 2013) with the ‘mgcv’ package (Wood 2013). I used the same variable selection procedure by removing combinations of variables with partial correlation coefficient higher than |0.7|, by testing all models with combinations of 4 variables (Mannocci et al. 2014a; Virgili et al. 2017a) and selecting the best model with the lowest mean prediction error determined by a leave-one-out cross validation process (minimum Generalised Cross-Validation score; Wood 2006a; Clark 2013).

Monthly predictions at 0.25° resolution were averaged over the entire time period (1998-2015) to produce maps of mean predicted densities which represent the average expected long-term distributional patterns of the beaked, sperm and kogiid whales. I did not attempt to correct predicted densities for availability bias thus predicted densities are relative densities. To lighten the reading, the relative densities will be hereafter labelled as densities. Finally, I provided uncertainty maps by computing the variance around the predictions as the sum of the variance around the mean prediction and the mean of the monthly variances. Then, the coefficient of variation was calculated as:

$$CV = 100 \times \sqrt{(\text{variance over the survey period}) / \text{mean over the survey period}}$$

### 4.2.4 Environmental space coverage gap analysis

To model the habitats of deep-divers in the North Atlantic and Mediterranean basins, I gathered data from a large region collected over a long period. The cumulative effort was not homogeneous and showed extensive geographical gaps. Therefore, I conducted gap analyses on environmental space coverage to identify areas where habitat models could produce reliable predictions outside the survey blocks, *i.e.* geographical extrapolation, whilst remaining within the ranges of surveyed conditions for the combinations of covariates selected by the models, *i.e.* areas of environmental interpolation (Jennings 2000).

To do this, I determined the extent of the environmental interpolation (*versus* extrapolation) obtained by combining the four variables selected by the models (one analysis was done for each species group). This was calculated by using the convex hull methodology defining effort data as the calibration dataset and climatological predictors (*i.e.* the average situation of each predictor over the 18 years period) for the entire study area as the prediction dataset (King and Zeng 2007; Authier et al. 2016). Here I used climatological predictors instead of monthly predictors to limit computation time (the analysis would have been done for each month and then averaged over 18 years). The convex hull of a set of points is the smallest convex envelope that contains these points, *i.e.* all effort data points described by the selected covariates. If prediction data fall inside the convex hull, they are interpolations while if they fall outside the convex hull, they are extrapolations; prediction made at any interpolation point within study area being considered as a more reliable (less model-dependent) than predictions made at extrapolation points (King and Zeng 2007; Authier et al. 2016).

Due to the large number of data (more than 280,000 points in the calibration dataset), convex hulls were estimated by random sub-sampling with the ‘WhatIf’ R-package (Stoll et al. 2014). I randomly

extracted a fraction of the calibration dataset (10,000 points) to estimate a convex hull and assess environmental extrapolation in the prediction dataset. A combination of climatological predictor values that falls inside the convex hull corresponds to interpolation. The combinations of climatological predictor values that were classified as interpolations were set aside but the other combinations were retained and further tested against another random sample of 10,000 points from the calibration data. This procedure was carried out until the full calibration dataset was examined.

The full procedure was conducted twice. Firstly, what I called ‘simple interpolation’, considered the full range of sampled variables to identify all points of the whole study area where the actual combinations of environmental variables had been sampled in survey blocks. Secondly, in the ‘precautionary interpolation’, I arbitrarily applied a 5% precautionary approach, *i.e.* 5% of the extreme values of the sampled variables were removed to include in the interpolation areas only the points whose associated combinations of covariates fell within the 95% core ranges sampled. This allowed the definition of two levels of confidence in the predictions.

Finally, I produced maps delineating the extent of the simple and the precautionary interpolation areas, and overlaid them to the density prediction maps to highlight areas with a greater reliability.

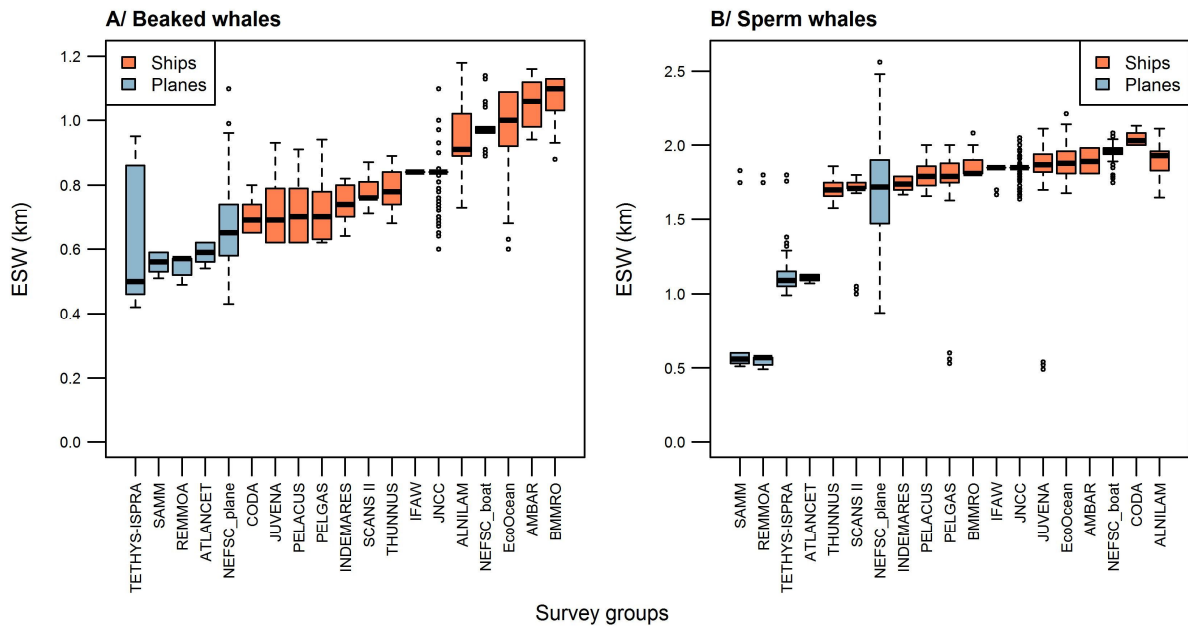
## 4.3 WHICH DISTRIBUTION FOR THE DEEP-DIVERS?

### 4.3.1 Effective strip width

The ESWs estimated with the meta-analysis varied with surveys and platform type; they were on average narrower in aerial than in shipborne surveys (Fig. 4.2). This is probably because aerial observers usually record animals below the aircraft while shipboard observers look further afield. ESWs were generally larger and more consistent, between surveys using the same platform type, in sperm whales than in beaked whales. There were not enough kogiid sightings to estimate an ESW for each survey particularly in shipborne surveys; consequently I pooled all aerial surveys and estimated an ESW of 1.1 km that I applied to all surveys (shipborne and aerial surveys). The outcomes of this analysis were consistent with expectations: a decrease in Beaufort seastate resulted in an increase in ESW estimations (Appendix C.4 of Annex C). Compared to ESWs obtained by using the more conventional Distance approach (Appendix C.4 of Annex C; Thomas et al. 2010; Buckland et al. 2015), ESW estimated in the present meta-analysis were shorter and their confidence intervals smaller.

Predictions of the three species groups can be considered as the summer habitats as most sightings were recorded from June-October (84% beaked whale, 76% sperm whales and 77% kogiids). Although effort was almost evenly distributed between the two seasons (53% in the summer at large – June to October – and 47% in the winter at large – November to May), there were not enough data to fit a model in winter possibly because of the poorer sighting conditions (mean Beaufort seastate was equal to 2.6 in the summer compared to 3.1 in the winter).

Overall, encounter rates were very low with 0.05 sightings·100 km<sup>-1</sup> for beaked whales, 0.07 sightings·100 km<sup>-1</sup> for sperm whales and <0.01 sightings·100 km<sup>-1</sup> for kogiids (Table 4.3). Highest encounter rates were recorded in the tropics for the three species groups, particularly for the kogiids. There was no sighting of kogiids in the Mediterranean Sea.



**Fig. 4.2. Beaked whale and sperm whale average ESWs estimated with the meta-analysis for each survey group and each platform type.** For each survey group, the boxplot represents the extent of estimated ESWs depending on Beaufort seastates and observation heights recorded within the group.

**Table 4.3. Encounter rates in sightings·100 km<sup>-1</sup> calculated for the entire study area and each sub-region.** NE-ATL: northeast Atlantic Ocean; NW-ATL: northwest Atlantic Ocean; MED: Mediterranean Sea.

	NE-ATL	NW-ATL	MED	TROPICS	STUDY AREA
<b>Beaked whales</b>	$4.2 \times 10^{-2}$	$5.8 \times 10^{-2}$	$3.5 \times 10^{-2}$	$2.2 \times 10^{-1}$	$5.1 \times 10^{-2}$
<b>Sperm whales</b>	$5.7 \times 10^{-2}$	$6.7 \times 10^{-2}$	$9.0 \times 10^{-2}$	$9.5 \times 10^{-2}$	$6.7 \times 10^{-2}$
<b>Kogiids</b>	$1.3 \times 10^{-3}$	$1.0 \times 10^{-2}$	0.0	$2.3 \times 10^{-1}$	$8.5 \times 10^{-3}$

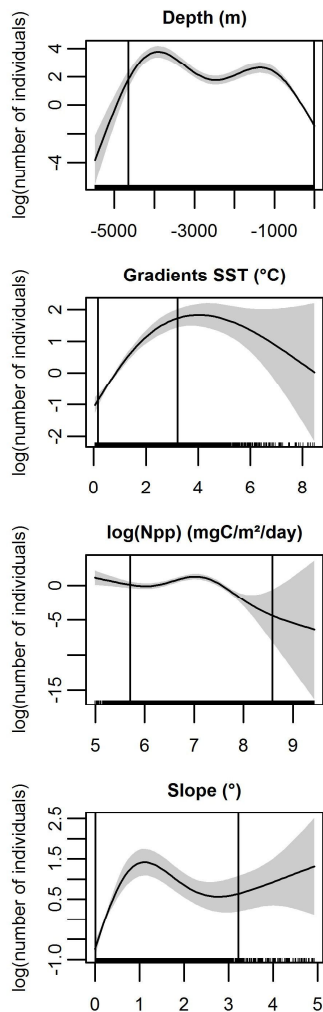
#### 4.3.2 Beaked whales

The beaked whale model accounted for 38.9% of the explained deviance (Fig. 4.3A). Depth, spatial gradients of SST, NPP and slope were the variables that most influenced the habitats of beaked whales. Highest densities were predicted for a large depth range with two modes (*ca.* 1,500 m and *ca.* 4,000 m), high slopes (*ca.* 1°), high gradients of SST (*ca.* 3°C) and medium productivity (*ca.* 1,100 mgC·m<sup>-2</sup>·day<sup>-1</sup>). This resulted in a concentration of individuals along steep slope areas associated with high depths, with highest densities predicted on the western side of the Atlantic Ocean (Fig. 4.3B). In the Mediterranean Sea, predicted densities were lower than in the Atlantic Ocean with highest densities predicted in the Alboran Sea, near the Gibraltar Strait, in the north of the Levantine basin, between Cyprus and Crete, and along the continental slopes. No individuals were predicted near Tunisia or in the northern Adriatic Sea (Fig. 4.3B). The gap analysis identified areas where the combination of the four variables selected by the best model had not been sampled, resulting in an absence of prediction in 6% of the sampled area, *i.e.* 94% of the sampled area was available for simple interpolation (Fig. 4.3B). However, the precautionary interpolation area obtained by retaining the 95% core distribution of the environmental variables represented only 53% of the study area (Fig. 4.3C), mostly because sampling effort in the open oceanic waters was insufficient to predict with confidence densities in the entire study area, particularly in the centre of the Atlantic Ocean. Coefficients of variation were higher in shallow waters associated

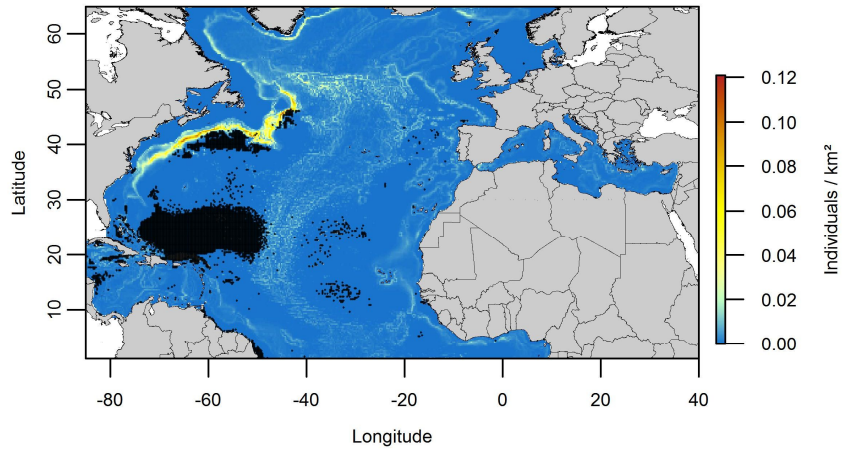


with high gradients of SST, where beaked whales have not been reported by any surveys (Appendix C.5A of Annex C).

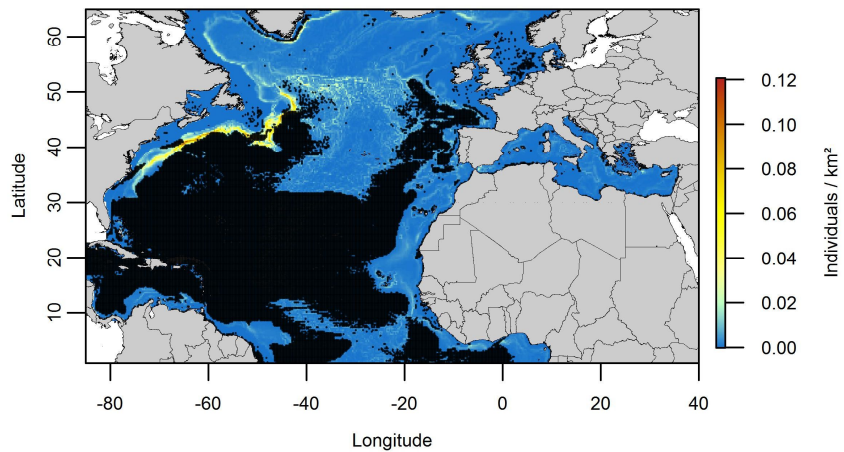
#### A/ Beaked whale model - $D^* = 38.9\%$



#### B/ Predictions - Interpolation 94%



#### C/ Predictions with a 5% precautionary approach - Interpolation 53%



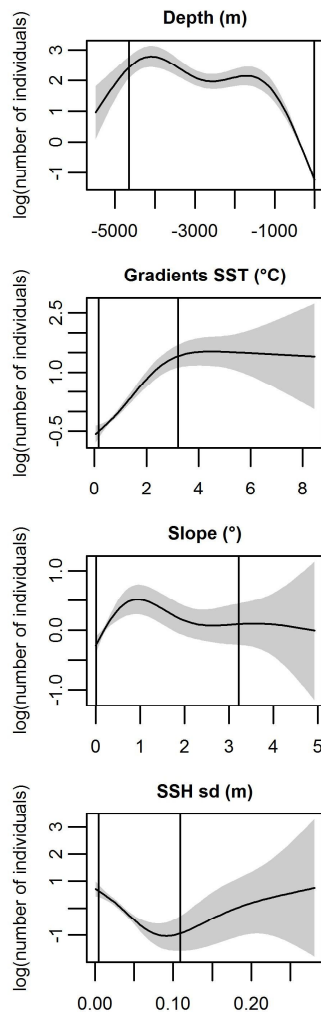
**Fig. 4.3.** Functional relationships for the selected variable (A) and the predicted relative densities of beaked whales in individuals.km<sup>-2</sup> (B and C). A: Solid lines are the estimated smooth functions, and the shaded regions represent the approximate 95% confidence intervals. The y-axes indicate the number of individuals on a log scale, where zero indicates no effect of the covariate. The vertical lines indicate the 2.5<sup>th</sup> and 97.5<sup>th</sup> quantiles of the data. Black areas on prediction maps (B: without precautionary approach and C: with a 5% precautionary approach) represent zones where I did not extrapolate the predictions. Percentages represent the proportion of the study area defined as interpolation with the gap analysis.

### 4.3.3 Sperm whales

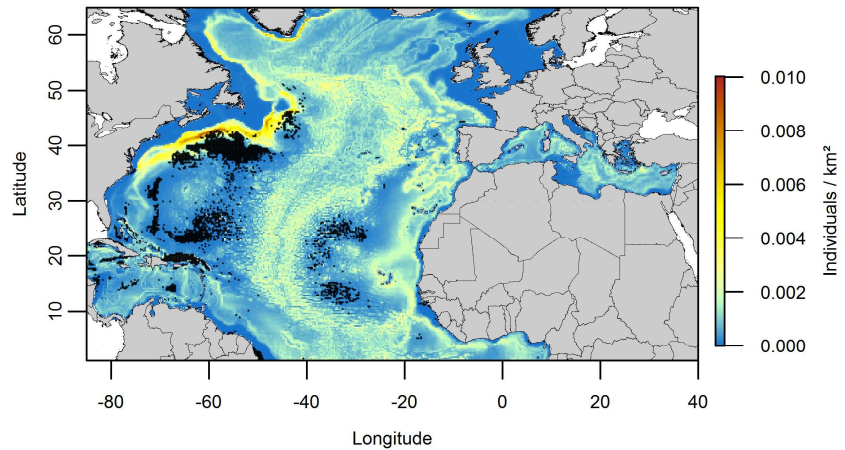
The explained deviance of the sperm whale model was 25.6% (Fig. 4.4A). As for beaked whales, depth, spatial gradients of SST and slope were the variables that most influenced the habitats of the species group, complemented by the standard deviation of SSH. Densities of sperm whales were predicted to increase in deep waters associated with steep slopes and high gradients of SST. The predicted habitats of sperm whales were more homogenous than for beaked whales, since the former appeared less restricted to slope areas (Fig. 4.4B). Highest densities were also predicted on the western side of the Atlantic basin, along the Gulf Stream. As for beaked whales, predicted densities of sperm whales were lower in the Mediterranean Sea than in the Atlantic Ocean. Highest densities were

predicted in the north of the Levantine basin, between Cyprus and Crete and fairly evenly predicted between the continental slopes and the oceanic waters, except near Tunisia and in the northern Adriatic Sea (Fig. 4.4B). Only 4% of the study area (Fig. 4.4B) corresponded to combinations of values of the selected covariates that had not been sampled during the surveys, but predictions within the core range of covariates only covered 44% of the study area. In fact, the highest predicted densities were partly outside this confidence zone (Fig. 4.4C). Coefficients of variation were highest in non- or poorly-sampled areas where uncertainty was therefore greatest (Appendix C.5B of Annex C).

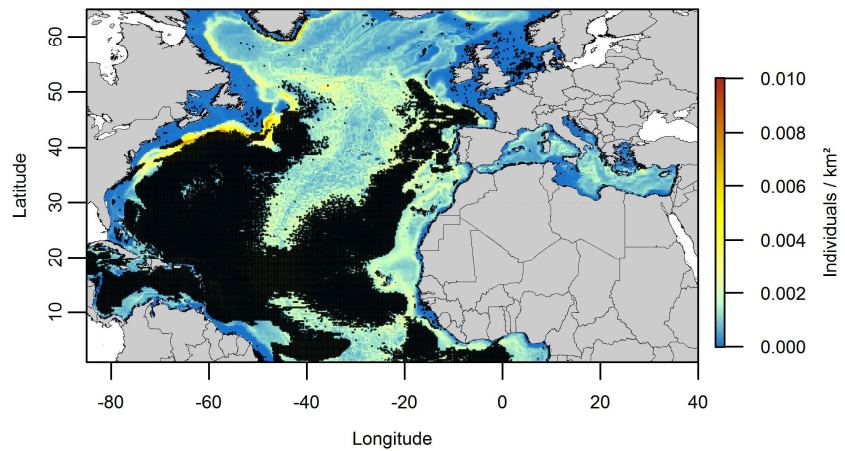
#### A/ Sperm whale model - $D^* = 25.6\%$



#### B/ Predictions - Interpolation 96%



#### C/ Predictions with a 5% precautionary approach - Interpolation 56%

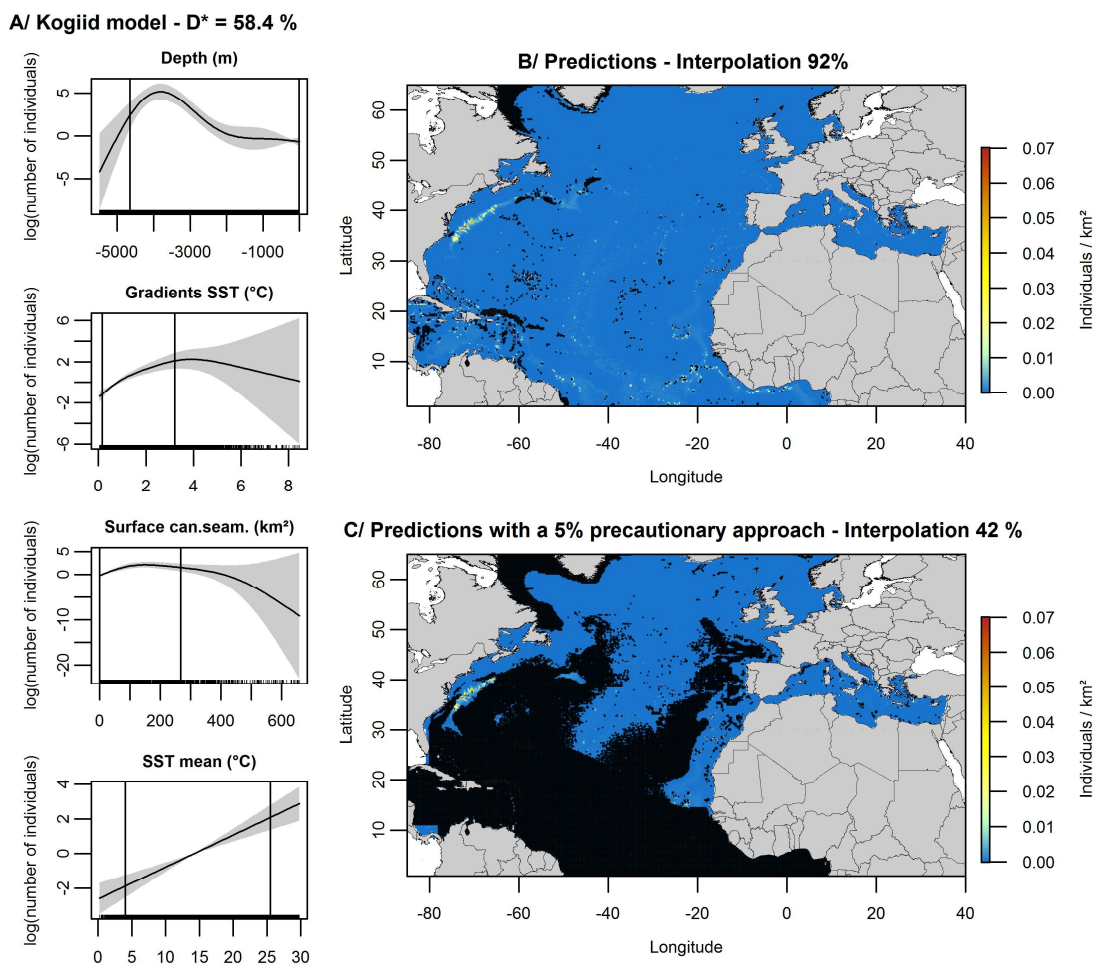


**Fig. 4.4.** Functional relationships for the selected variable (A) and the predicted relative densities of sperm whales in individuals.km<sup>-2</sup> (B and C). A: Solid lines are the estimated smooth functions, and the shaded regions represent the approximate 95% confidence intervals. The y-axes indicate the number of individuals on a log scale, where zero indicates no effect of the covariate. The vertical lines indicate the 2.5<sup>th</sup> and 97.5<sup>th</sup> quantiles of the data. Black areas on prediction maps (B: without precautionary approach and C: with a 5% precautionary approach) represent zones where I did not extrapolate the predictions. Percentages represent the proportion of the study area defined as interpolation with the gap analysis.

#### 4.3.4 Kogiids

The kogiid model accounted for 58.4% of the explained deviance (Fig. 4.5A). Depth, spatial gradients of SST, surface of canyon and seamount habitats per cell and mean SST were the variables that most influence the habitats of kogiids. Highest densities were predicted in warm and deep waters

associated with fronts and canyons or seamounts habitats. Consequently, individuals were not predicted in the northern part of the study area but mainly along the Gulf Stream where both fronts and canyons are abundant (Fig. 4.5B). Non-null densities were predicted in the Mediterranean Sea although no individuals were sighted (Fig. 4.5B). Because SST was among the selected covariates, 8% of the study area was classified as extrapolation zone (Fig. 4.5B). As there was little sampling effort in tropical and sub-polar regions, extreme temperature values were less sampled, resulting in a smaller prediction confidence zone for kogiids than for other species groups. The precautionary interpolation area, based on the 95% core distribution of the covariates' ranges, was reduced to 42% of the study area (Fig. 4.5C). Coefficients of variation were the highest in shallow waters and in the Mediterranean Sea, where kogiids have not been reported by any surveys (Appendix C.5C of Annex C).



**Fig. 4.5. Functional relationships for the selected variable (A) and the predicted relative densities of kogiids in individuals.km<sup>-2</sup> (B and C).** A: Solid lines are the estimated smooth functions, and the shaded regions represent the approximate 95% confidence intervals. The y-axes indicate the number of individuals on a log scale, where zero indicates no effect of the covariate. The vertical lines indicate the 2.5<sup>th</sup> and 97.5<sup>th</sup> quantiles of the data. Black areas on prediction maps (B: without precautionary approach and C: with a 5% precautionary approach) represent zones where I did not extrapolate the predictions. Percentages represent the proportion of the study area defined as interpolation with the gap analysis.

## 4.4 A BASIN WIDE APPROACH TO MODEL THE DISTRIBUTION OF RARE MARINE SPECIES

Deep-divers are species characterised by low sightings rates. Therefore, modelling their habitats is particularly challenging. This study merged different surveys to capitalise on more than 1,240,000 km of sighting effort deployed over the North Atlantic Ocean and the Mediterranean Sea in the past two decades. This data-assembling required that different protocols or platform types would be taken into account and therefore the different species-specific and survey-specific detectability. I then investigated the habitats of deep-divers using state-of-the-art statistical methods with a focus on how much confidence can be given to predictions. The habitats of deep-diving cetaceans were mainly influenced by static environmental variables such as depth or slope as well as spatial gradients of temperatures, revealing density hotspots in the western North Atlantic Ocean.

### 4.4.1 Methodological considerations

Over the past few years, data-assembling has been increasingly used for the study of top marine predators (Winiarski et al. 2014; Roberts et al. 2016; Mannocci et al. 2017b; Rogan et al. 2017). Due to the very low sighting rates of deep-diving cetaceans, taken separately each survey could not provide sufficient data to model the habitats of rare species, thus data-assembling was necessary. However, in contrast to Rogan et al. (2017), I did not assemble data collected with similar protocols but data collected with different protocols which implied to homogenise and somehow to degrade the data before developing a common spatial model. At a time when shared databases are becoming increasingly important (*e.g.* OBIS SEAMAP - <http://seamap.env.duke.edu/>, EMODnet - <http://www.emodnet.eu/>), sharing standardised observation protocols would be of utmost importance to facilitate data-assembling; this would allow to maintain higher data standards and help describing the large-scale habitats of the species.

Winiarski et al. (2014) warned how data collected in different surveys must be checked for compatibility, especially with respect to segment sizes. In this study, the great disparity between transects of aerial and shipboard surveys (from about two km to hundreds of kilometres) required to format the data in small segments of 5 km for modelling. The 5 km data format could induce a mismatch with covariate resolution, which was coarser due to the vast extent of the study area. However, this mismatch turned to be limited.

Regarding the environmental data used in the SDM, most of the oceanographic variables were related to processes affecting the euphotic zone (*i.e.* in the upper layer of the ocean). This is because most of environmental variables are based on satellite data and variables that describe deep water column are difficult to obtain. As deep-diving cetaceans spend most of their time in depth (Perrin et al. 2009), the use of surface variables might lead to a mis-interpretation of species habitats. Indeed, using proxies related to surface waters, I may not have identified the true causal relationships that explain the habitats but only indirect relationships (Austin 2002). However, explained deviances of the models were fairly high (from 25.6% to 58.4%) for cetaceans and the coefficients of variation were the highest on the continental shelf (Appendix C.5 of Annex C), where deep-divers are known to be mostly absent (Waring et al. 2001; Cañadas and Vázquez 2014; Arcangeli et al. 2015; Roberts et al. 2016). This indicated a good effectiveness of the models to make coherent predictions despite the use of indirect variables.

Nevertheless, the very high deviance of the kogiid model (58.4%) might indicate some level of model over-fitting due to the small number of data, even if predictions were consistent with the known ecology of the taxon (McAlpine 2009).

By assembling data collected in different regions (*e.g.* Mediterranean Sea and Atlantic Ocean or northwest and northeast Atlantic Ocean) I assumed similar relationships of deep-divers with their habitat across these multiple ecosystems. However these ecosystems are very different with an active frontal system associated with the Gulf Stream in the western Atlantic Ocean (Tomczak and Godfrey 2003) or an oligotrophic Mediterranean Sea (Bethoux *et al.* 1999). Consequently, deep-diver habitats may be different between regions with for example, a possible greater influence of thermal fronts on species habitat in the western Atlantic Ocean than in the eastern Atlantic Ocean or the Mediterranean Sea. Indeed, Roberts *et al.* (2016) evidenced in the western Atlantic Ocean an influence of the depth, the distance to fronts and to eddies on the habitats of the three species groups. On the other hand, both Cañadas and Vázquez (2014) and Rogan *et al.* (2017) found depth to be one of the most important predictors of deep-diver habitats in the Mediterranean Sea and in the northeast Atlantic Ocean respectively. This suggests a good consistency in the habitats of these species groups, which are highly associated with topographic features. Consequently, a data-assembling at such a large scale seem consistent. However, bimodal response to depth for beaked whales and sperm whale with modes of densities predicted at 1,500 m and 4,000 m might reveal different habitats, the species groups probably use different habitats to forage. The 1,500 m mode was essentially made up of sightings from the Mediterranean Sea while the 4,000 m mode was essentially made up of sightings from the northwest and northeast Atlantic Ocean (Fig. 4.1; Appendix C.2 of Annex C). A model for each ecoregion or the inclusion of an interaction with ecoregion in the model could help determining whether the variables selected by the different models would be identical.

#### 4.4.2 Large-scale deep-diver habitats

Physiographic variables were highly predictive of deep-diver habitats. In the three models, depth was one of the most influential variable. The surface of canyon and seamount habitats per cell was a significant variable only for the kogiids. This is consistent with the influence of topographic features noticed in smaller regions (Fergusson *et al.* 2006; MacLeod *et al.* 2011; Whitehead 2013; Wong and Whitehead 2014). Oceanographic variables were also important. For each species group, spatial gradients of SST significantly contributed to the models. Deep-divers seemed to concentrate in areas of strong gradients such as thermal fronts in which prey aggregate (Brandt 1993; Bost *et al.* 2009; Woodson and Litvin 2015). Hence, the Gulf Stream, which is the most active frontal zone in the study area compared to the eastern boundary currents that are broader and much slower, may explain the high densities of deep-divers on the western side of the North Atlantic Ocean (Griffin 1999; Waring *et al.* 2001; Hamazaki 2002; Roberts *et al.* 2016).

In this study, I geographically extrapolated the deep-diver habitats to the entire North Atlantic Ocean and Mediterranean Sea while remaining within sampled environmental conditions (*i.e.* within environmental interpolation). At a local scale, predictions were consistent with known distributions. As Cañadas and Vázquez (2014), I identified a beaked whale density hotspot in deep waters of the Alboran Sea but predicted densities were lower and more extended towards the Gibraltar Strait. In the predictions, the Tyrrhenian and Ligurian Seas also appeared as suitable habitats for beaked whales, consistent with the previous results (Arcangeli *et al.* 2015; Lanfredi *et al.*, 2016). In addition, recorded

strandings of Cuvier's beaked whale along the coasts of the Ligurian and Ionian Seas and the eastern coasts of the Mediterranean Sea (Podestà et al. 2006) revealed the presence of the taxon close to these coasts, as suggested by the predictions. For sperm whales, predictions agreed with Praca and Gannier's (2008) results with potential habitats predicted on the continental slope off France and off islands of the western Mediterranean Sea. Sperm whale codas recorded in the Ligurian, Tyrrhenian and Ionian Seas (Pavan et al. 2000) revealed the presence of the species in these areas, as suggested by the predictions. In the Bay of Biscay, highest densities of beaked whales and sperm whales were predicted along the slope, consistent with encounter rates estimated from platforms of opportunity (Kiszka et al. 2007) and abundances predicted from shipboard and aerial surveys (Rogan et al. 2017). In the western Atlantic Ocean, models predicted highest densities of beaked whales and sperm whales along the continental slope consistently with Roberts et al. (2016) but predicted densities were lower (about 50% lower). Regarding the kogiids, there is little published literature allowing predictions to be compared. In the northwest Atlantic Ocean, kogiids were predicted in warm deep waters, which was consistent with their known ecology (McAlpine 2009) and the patterns of distribution predicted by Mannocci et al. (2017b), except that no individual was predicted off the coast of Florida. However this was an area of environmental extrapolation in the precautionary approach.

In the present study from 92 to 96% of the study area, with no precautionary approach, and from 42 to 56% of the study area, with a 5% precautionary approach, were considered to provide confident predictions because they corresponded to pixels with environmental condition encompassed by 95% of sampled pixels. Large gaps in environmental space coverage were revealed, especially in deeper waters of the central north Atlantic gyre and in tropical waters. It can be noted that areas of interest for deep-divers were predicted at the margin of the precautionary interpolation zone in particular because deeper waters and steeper slopes were within the upper 2.5% quantiles of aggregated survey coverage for these two physiographic covariates. This suggested that sampling effort was not sufficient in deeper and steeper areas and more intensive sampling effort performed in these areas could help to better describe habitats used by deep-divers.

Meanwhile, the predicted habitats provided in this study could be included in a marine spatial planning. This consists in analysing and allocating the spatial and temporal distribution of human activities in marine areas, here for example anthropogenic sound, to achieve ecological objectives, such as species conservation (Douve 2008). Thus, with this methodology, the conservation of deep-diving cetaceans could be improved.

Finally, modelling rare species habitats is particularly challenging because habitat models require large datasets yet rare species typically yield low numbers of sightings. As a result, assembling datasets appeared to be an appropriate strategy to model the large scale habitats of deep-divers, the beaked whales, sperm whales and kogiids, across the North Atlantic and the Mediterranean basins.

Thanks to a data-assembling methodology, I predicted the large scale distribution of deep-divers. At a local scale, predicted habitats were consistent with previous studies. Predictions at a larger scale highlighted a gradient of predicted densities (with highest densities predicted on the western side of the study area) which would not have been evident at a local scale and showed pronounced influence of active frontal zones, such as the Gulf Stream, on the habitats of these species groups. Even though gaps remain at such a large scale, I was able to predict the habitats of these species groups throughout the Atlantic basin and thus identify potential habitats, even in non-sampled areas. In addition, due to

the large extent of the study area, a prediction relevance assessment was needed. Through an environmental space coverage gap analysis, I identified areas in tropical and deep oceanic waters where sampling effort was insufficient and need to be intensified to increase prediction reliability. Finally, by developing a data-assembling procedure that could be applied to any species and to any local or extended study area, I helped to improve the knowledge of deep-diver distribution. In addition, I noticed a possible change of the habitat drivers between the different regions that could be interesting to investigate.

# Chapter 5

---

## DATA-ASSEMBLING: A MATTER OF ECOSYSTEMS SIMILARITY

---



© Laura Hedon

### CONTENTS

---

5.1 CONTEXT AND OBJECTIVES .....	74
5.2 METHODOLOGY .....	75
5.2.1 Data origin.....	75
5.2.2 Model fitting and predictions.....	77
5.2.3 Model assessment.....	78
5.3 CAN PREDICTIONS BE EXTRAPOLATED IN DIFFERENT ECOSYSTEMS? .....	78
5.3.1 Are the models transferable between regions?.....	78
5.3.2 Does data-assembling allow to make more reliable predictions? .....	79
5.4 CAN DIFFERENT ECOSYSTEMS BE ASSEMBLED? .....	83
5.5 WHAT SCALE TO CONSIDER FOR DATA-ASSEMBLING? .....	84

**T**HIS chapter aims to assess, through a model transferability approach, the extrapolated predictions provided in the previous chapter and whether merging data from multiple ecosystems allowed to improve these predictions. This chapter is planned to be published as a separate stand-alone paper.



## 5.1 CONTEXT AND OBJECTIVES

Habitat preferences of deep-diving cetaceans, and rare species in general, are difficult to model due to the small number of available sightings. Habitat models require large datasets and data-assembling appeared as a key strategy to predict species distribution (Chapter 4; Roberts et al. 2016; Mannocci et al. 2017b; Rogan et al. 2017). To model the large scale habitats of deep-diving cetaceans, datasets from different sources were assembled by merging data collected in the Mediterranean Sea and the North Atlantic Ocean. Despite a fairly broad coverage of the study area by the numerous surveys that were pooled in the analysis, many geographical and habitat gaps persist, making extrapolation of the predictions outside the sampled areas challenging. In this context, an environmental space coverage gap analysis identified areas of environmental interpolation where predictions provided by the models would be reliable because the combinations of values taken by the variables selected by the models remained within the ranges of surveyed conditions (Jennings 2000).

However, even within interpolation areas, the same oceanographic processes can differently drive species habitats from one ecosystem to another. A very good example, outside the study area, is the effect of eddies on primary production concentration in the Mozambique Channel and the South Pacific Ocean. In the Mozambique Channel, anticyclonic and cyclonic eddies induce high chlorophyll biomass concentration (Schouten et al. 2003, De Ruijter et al. 2004; Longhurst 2007) while in the South Pacific Ocean, within the anticyclonic gyre, eddies do not induce a rise of nutrient-rich deep waters because of a sink of hypersaline surface waters induced by a high evaporation (Rougerie and Rancher 1994). Thus, across several ecosystems, a process, such as eddies, may affect species habitats in a different way.

In the study area, multiple ecosystems are present. The North Atlantic Ocean and the Mediterranean Sea are two physically and biologically distinct regions but even within the North Atlantic Ocean, the northwest and northeast sub-basins also differ from each other as the extremely active frontal system associated with the Gulf Stream in the west has no equivalent in the eastern Atlantic Ocean (*cf.* Chapter 2). Longhurst (2007) has divided the Global Ocean into four biomes (equivalent to terrestrial biomes) which represent the main types of oceanic phytoplankton and are defined according to the seasonality of the mixing layer, nutricline and euphotic depths. These biomes are divided into biogeochemical provinces with specific environmental and oceanographic conditions (*e.g.* bathymetry, regional circulation, stratification, river discharges), delineated by boundaries such as areas of convergence, divergence or frontal oceanic zones and characterised by a specific fauna and flora (Fig. 5.1). Particularly, he considered the Mediterranean Sea as a separate entity and segregated the North Atlantic Ocean into multiple provinces with coastal and oceanic provinces. By merging data collected in these different provinces, I assumed similar relationships between deep-divers and their habitat but drivers may in fact differ between regions (Mediterranean Sea, northeast or northwest Atlantic Ocean).

Some studies partly answered this question by assessing model transferability between regions (Randin et al. 2006; Vanreusel et al. 2007; Mannocci et al. 2015; 2017; Redfern et al. 2017). Indeed, if a model is transferable from one region to another, this indicates a consistency in species habitat drivers. However, the ability of the models to extrapolate beyond the surveyed areas appeared variable. In similar ecoregions, models were highly transferable (Vanreusel et al. 2007) while in dissimilar regions, models were not transferable (Randin et al. 2006; Redfern et al. 2017). Considering that, Mannocci et al. (2017b) and Redfern et al. (2017) suggested that collecting data from multiple ecosystems would be a way to more reliably extrapolate predictions.

Consequently, the aim of this study was to extrapolate the predictions from one region to another, particularly in regions that belonged to the interpolation zones of the general model, but also to assess whether the merging of data from multiple ecosystems allowed to improve the species habitat predictions. I first built a model for each region (with data collected only in the corresponding region) and evaluated its ability to predict the species habitats in other regions. I compared these model predictions to the predictions of the model fitted in the other region. Then, I built a model using all the data collected in the Atlantic Ocean and a model using all the data from the study area (Atlantic Ocean and Mediterranean Sea) to test the effect of input data on the predictions.

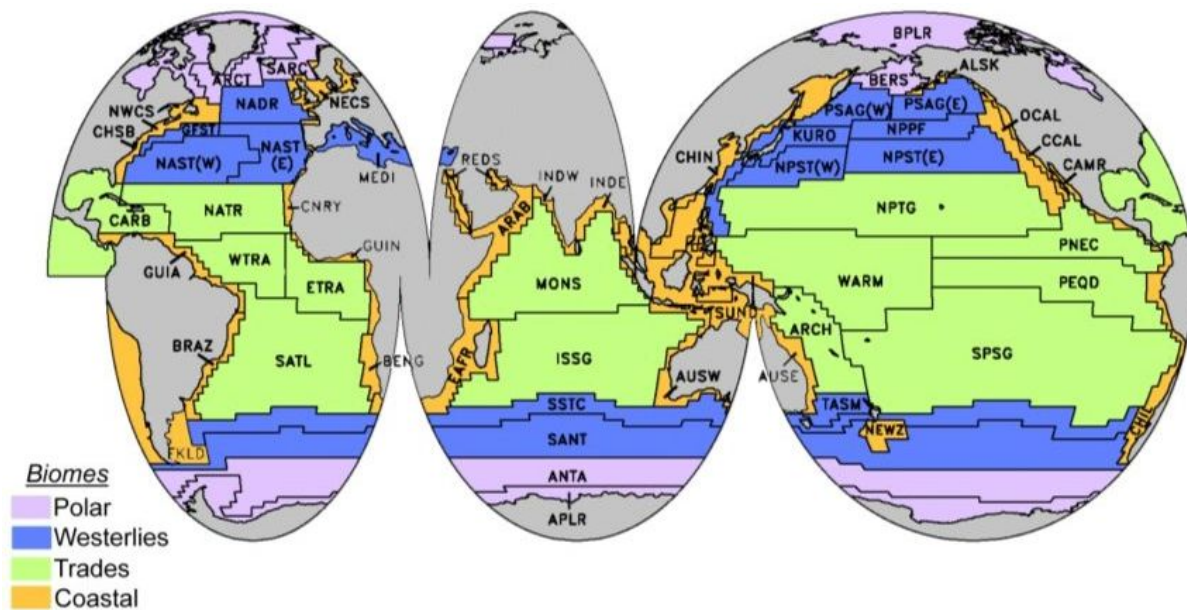


Fig. 5.1. The biogeochemical provinces from Longhurst (2007).

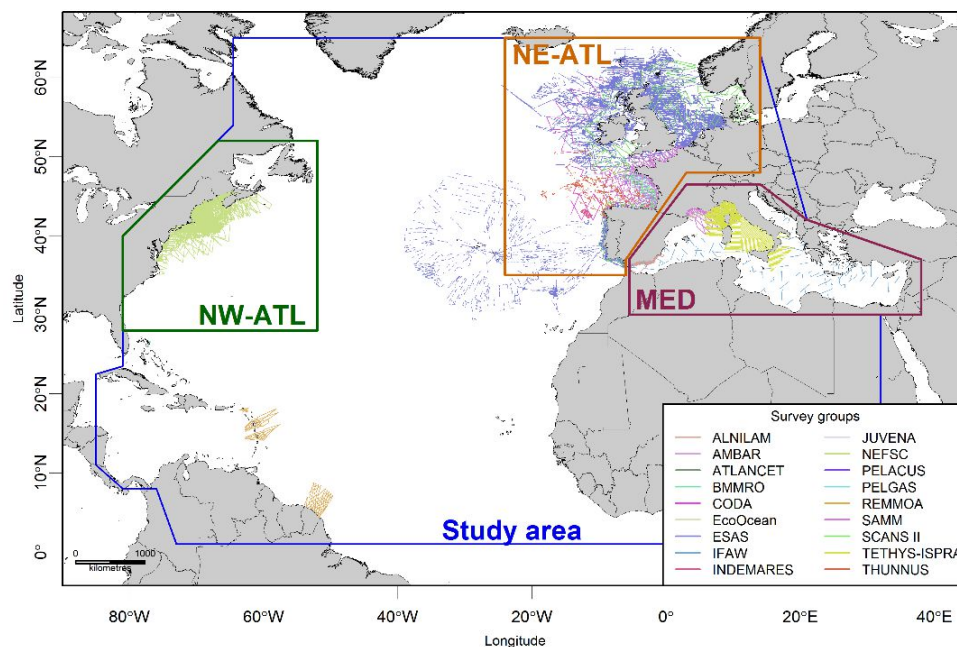
## 5.2 METHODOLOGY

### 5.2.1 Data origin

Effort data were assembled from surveys carried out in the Mediterranean Sea and the Atlantic Ocean while environmental variables were extracted for the entire time period (1998-2015) and survey area (*cf.* Fig. 4.1 and Table 4.2). To limit computing times, I reduced the number of variables by using only those selected by the five best models in the general study (Virgili et al. in prep.; Chapter 4; Table 5.1). From the whole study area, I delineated three regions, the north-west Atlantic Ocean (hereafter labelled as 'NW-ATL'), the north-east Atlantic Ocean ('NE-ATL') and the Mediterranean Sea ('MED'; Fig. 5.2). These regions corresponded to distinct ecosystems and to the most heavily sampled areas. From the whole datasets, called 'GENERAL dataset', I extracted 4 subsets. A dataset which contained only data from the NW-ATL (hereafter labelled as 'NW-ATL dataset'), a dataset, which contained only data from the NE-ATL ('NE-ATL dataset'), a dataset which contained only data from the MED ('MED dataset') and a dataset which pooled data from the NW-ATL and the NE-ATL ('N-ATL dataset'). In this study, I only used the beaked and sperm whale datasets because there was not enough kogiid data to fit a model in each region.

**Table 5.1. Candidate environmental predictors used for the habitat modelling.** All variables were rescale at a 0.25° resolution. Sources: A: Depth and slope were computed from GEBCO-08 30 arc-second database (<http://www.gebco.net/>); 30 arc-second is approximately equal to 0.008°. B: The mean and gradient of Sea Surface Temperature (SST) were calculated from the GHR SST Level 4 CMC SST v.2.0 (Canada Meteorological Centre, <https://podaac.jpl.nasa.gov/dataset/CMCO.2deg-CMC-L4-GLOB-v2.0>). C: The Aviso ¼° DT-MADT geostrophic currents dataset was used to compute mean and standard deviation of Sea Surface Height (SSH) and Eddy Kinetic Energy (EKE; <https://www.aviso.altimetry.fr/en/data/products/sea-surface-height-products/global/madt-huv.html>). D: Net primary production (Npp) was derived from SeaWiFS and Aqua using the Vertically Generalised Production Model (VGPM; <http://orca.science.oregonstate.edu/1080.by.2160.8day.hdf.vgpm.m.chl.m.sst.php>).

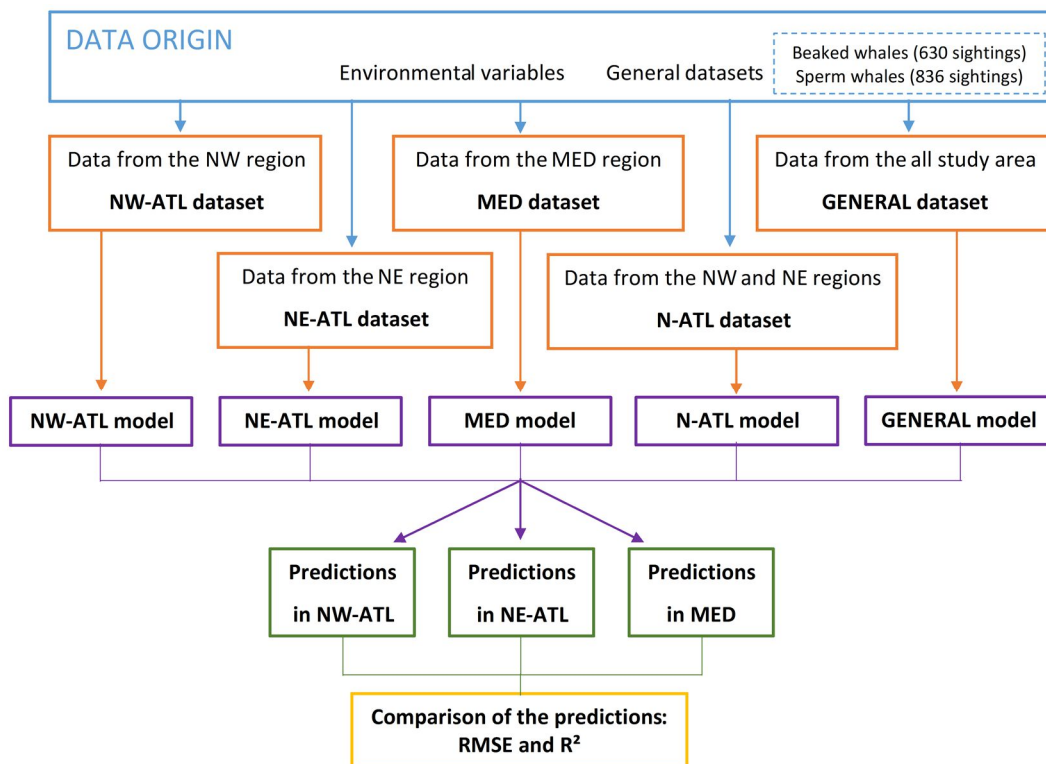
Environmental predictors, units and abbreviations	Resolution	Sources	Effects on pelagic ecosystems of potential interest to top predators
<b>Physiographic</b>			
Depth (m)	30 arc sec	A	Shallow waters are often associated with high primary production
Slope (°)	30 sec arc	A	Associated with currents, high slope induce prey aggregation or enhanced primary production
<b>Oceanographic</b>			
Mean of SST (°C) – ‘SSTm’	0.2°, daily	B	Variability over time and horizontal gradients of SST reveal front locations, potentially associated with prey aggregations or enhanced primary production
Mean gradient of SST (°C) – ‘SSTgrad’	0.2°, daily	B	
Mean of SSH (m) – ‘SSHm’	0.25°, daily	C	High SSH is associated with high mesoscale activity and enhanced prey aggregation or primary production
Mean of EKE (m <sup>2</sup> .s <sup>-2</sup> ) – ‘EKEm’	0.25°, daily	C	High EKE relates to the development of eddies and sediment resuspension induce prey aggregation
Mean of Npp (mgC.m <sup>-2</sup> .day <sup>-1</sup> ) – ‘Npp’	9 km, 8 days	D	Net primary production as a proxy of prey availability



**Fig. 5.2. Study area and regions delineated for the study.** The blue polygon delineates the study area, the green polygon delineates the north-west region (NW-ATL), the purple polygon delineates the Mediterranean region (MED) and the orange polygon delineates the north-east region (NE-ATL). The N-ATL region pooled the NW-ATL and the NE-ATL. Lines represent the survey tracks.

### 5.2.2 Model fitting and predictions

Generalised Additive Models (GAMs) with a Tweedie distribution were fitted (Hastie and Tibshirani 1986; Wood 2006a) by using the 'mgcv' R-package (R-3.3.1 version; Wood 2013) for each dataset (NE-ATL, NW-ATL, MED, N-ATL and GENERAL datasets; Fig. 5.3). The parameter of the Tweedie distribution was directly estimated by the 'mgcv' function. I used a variable selection procedure by removing combinations of variables with Spearman partial correlation coefficient higher than  $|0.7|$ , by testing all models with combinations of four variables (Mannocei et al. 2014a; Virgili et al. 2017a) and selecting the best model with the lowest generalised cross validation score (Wood 2006b; Clark 2013). A maximum of four covariates per model was used to avoid excessive complexity of models and difficulty in their interpretation (Mannocei et al. 2014a; Virgili et al. 2017a). To facilitate comparisons of models and predictions, count data were transformed into presence-absence data, *i.e.* any sighting, regardless of associated group size, was considered as a single observation ('1') while the absence of sighting was set to '0'. Indeed, various methods, such as Area under the Receiving Curve (AUC; Elith et al. 2011), kappa statistic (Monserud and Leemans 1992) or True Skill Statistics (TSS; Allouche et al. 2006), are commonly and easily used to assess the model performance but can only be used on binary data so that I had to transform count data into presence-absence data. This does not change the predicted distribution patterns but provides maps in probabilities of presence instead of densities.



**Fig. 5.3. Flowchart of the method used in the study.** NE-ATL: northeast Atlantic region; NW-ATL: northwest Atlantic region; MED: Mediterranean region; N-ATL: north Atlantic region and pooled northwest and northeast Atlantic regions; GENERAL: study area and pooled the three regions; RMSE: Root Mean Squared Error;  $R^2$ : coefficient of determination.

Each model fitted to the five datasets (NE-ATL, NW-ATL, MED, N-ATL and GENERAL models) were used to predict the species habitats in the three regions, NE-ATL, NW-ATL and MED (Fig. 5.3). Following Virgili et al. (*in prep.*), monthly predictions at  $0.25^\circ$  resolution were averaged over the entire time period (1998-2015) to produce maps of mean predicted probabilities of presence and represent expected

general patterns of the beaked and sperm whales in the NE-ATL, NW-ATL and MED. I did not attempt to correct predicted probabilities of presence for availability bias thus predicted probabilities of presence are relative probabilities.

### 5.2.3 Model assessment

In a first step, the performance of each model was assessed by comparing explained deviances (Wood 2006b), AUC (as for MaxEnt in Chapter 3, Elith et al. 2011) and TSS (Allouche et al. 2006). TSS is only used for binary data and is independent of model prevalence (*i.e.* the proportion of observed sites in which the species was recorded as present). TSS takes into account the model sensitivity (*i.e.* proportion of presences accurately predicted) and specificity (*i.e.* proportion of absences accurately predicted) and is defined as  $sensitivity + specificity - 1$ . It ranges between -1 and +1; a value of |1| indicates a perfect discrimination of sites with and without the species and values below 0 indicate a performance no better than random distribution (Allouche et al. 2006).

The next step consisted in comparing the predictions provided by the models in each region. For each of the three regions, I defined a 'reference prediction' for comparison with the other predictions (hereafter labelled as 'experimental predictions'). This reference was defined as the prediction of a given region provided by the model fitted to the data of the same particular region. For example, the prediction in the NE-ATL provided by the NE-ATL model which used the data of the NE-ATL. Experimental predictions were obtained in a region from the models that used the data from the other regions. For example, the prediction in the NE-ATL obtained from the NW-ATL model that used data from the NW-ATL. Then I used the Root Mean Squared Error (RMSE; Barnston 1992) and the coefficient of determination ( $R^2$ ; Nagelkerke 1991) to compare experimental predictions to this reference prediction. The RMSE measures the difference between the predicted values of the experimental predictions and the predicted values of the reference prediction. RMSE specifies if a prediction is well-, over- or underestimated compared to the reference prediction; a low value indicates a good match between the predictions. The  $R^2$  measures the adequacy between the predictions by determining how well reference prediction is replicated by the experimental predictions. If predicted distribution patterns are identical,  $R^2 = 1$ , if they are completely different,  $R^2 = 0$ . These two criteria allowed to assess if the experimental predictions were consistent with the reference prediction.

## 5.3 CAN PREDICTIONS BE EXTRAPOLATED IN DIFFERENT ECOSYSTEMS?

### 5.3.1 Are the models transferable between regions?

In a first step, I assessed the performance of the models to fit the different datasets (NE-ATL, NW-ATL, MED, N-ATL and GENERAL datasets). For the two species groups, beaked and sperm whales, the five models performed well, with explained deviances ranging from 22% to 49%, AUC ranging from 0.83 to 0.98 and TSS ranging from 0.54 to 0.90 (Appendices D.1 and D.2 of Annex D). The only exception was the MED model for sperm whales which showed fairly poor performance with an explained deviance of 6%, an AUC of 0.69 and a TSS of 0.29. Overall, NW-ATL models showed a better performance than other models (highest explained deviances, AUCs and TSS) and N-ATL models showed a better performance than the GENERAL models. Selected variables and relationships between the probability of presence and the variables varied between the different models, suggesting different relationships with the

environment depending on the region. Particularly, relationships in the MED for beaked whales were highly different from the others with a unimodal relationship with depth, a linear relationship with slope, an inverse relationship with SSH and an influence of eddies (EKE) that were undetected in other regions (Appendices D.1 and D.2 of Annex D). In other regions, even if the selected variables were not identical (absence of depth in the NE-ATL model), relationships were similar (same pattern). Overall, depth and slope were predominantly selected, confirming the importance of physiographic variables in deep-diver habitat use.

Concerning the probabilities of presence predicted by the five models in the three regions, the NW-ATL models provided distribution patterns similar to the reference prediction in the NE-ATL with highest probabilities of presence predicted along the continental slopes for beaked whales and fairly homogeneously from the continental slopes to oceanic waters for sperm whales. However, NW-ATL models over-estimated the probabilities of presence for the two species groups (Fig. 5.4.A-B). For beaked whales, NE-ATL model provided a distribution pattern similar to the reference prediction in the NW-ATL, with highest probabilities of presence predicted along the continental slope, and a fairly good capacity of prediction ( $R^2 = 0.51$ ; Fig. 5.5.1/A-B). For sperm whales, the predictions were less convincing for the NE-ATL model with a fairly high RMSE (0.01) and a low  $R^2$  (0.28). However, for the two species groups, predictions were extrapolated in a larger area than for the reference prediction, indicating a larger environmental coverage in the NE-ATL than in the NW-ATL. For the two species groups, in the NE-ATL and NW-ATL, the  $R^2$  of the MED model was very low. Models were unable to reproduce the reference predictions and under-estimated the presence probabilities (Figs. 5.4.C and 5.5.C). Similarly, predictions provided by the NE-ATL and NW-ATL models in the MED were poor (slightly less so for the NE-ATL sperm whale model) with an over-estimation of predicted presence probabilities (Fig. 5.6.A-C).

### 5.3.2 Does data-assembling allow to make more reliable predictions?

N-ATL and GENERAL models predicted distribution patterns were similar to the prediction of the reference models in the NE-ATL (Fig. 5.4.D-E). Although predicted values were different, highest probabilities of presence were predicted along the continental slopes for beaked whales and fairly homogeneously from the continental slopes to oceanic waters for sperm whales. For beaked whales, predictions of the N-ATL model were the most similar to the reference prediction (smallest RMSE and highest  $R^2$ ) while for sperm whales the difference between predictions of the N-ATL and GENERAL models was not as significant, despite a more similar distribution pattern for the N-ATL model. For the two species groups, predictions of the N-ATL and GENERAL models were better than the predictions of the NW-ATL and MED models.

In the NW-ATL,  $R^2$  was higher for N-ATL models (Fig. 5.5). N-ATL models predicted better the distribution patterns than the other models but for sperm whales predicted probabilities of presence were over-estimated (higher RMSE than for the GENERAL model). Overall, N-ATL and GENERAL models predicted better than the NE-ATL and MED models and allowed to extrapolate over the entire area. Consequently, data-assembling at a large scale allowed to cover a large range of variable values.

As NE-ATL and NW-ATL models, N-ATL and GENERAL models performed poorly in the MED with fairly high RMSE and low  $R^2$  (Fig. 5.6.D-E). Predicted probabilities of presence were over-estimated with the N-ATL and GENERAL models for the two species groups and distribution patterns differed from the reference predictions, with probabilities of presence predicted in almost the entire area while distribution patterns of the reference predictions were patchy.

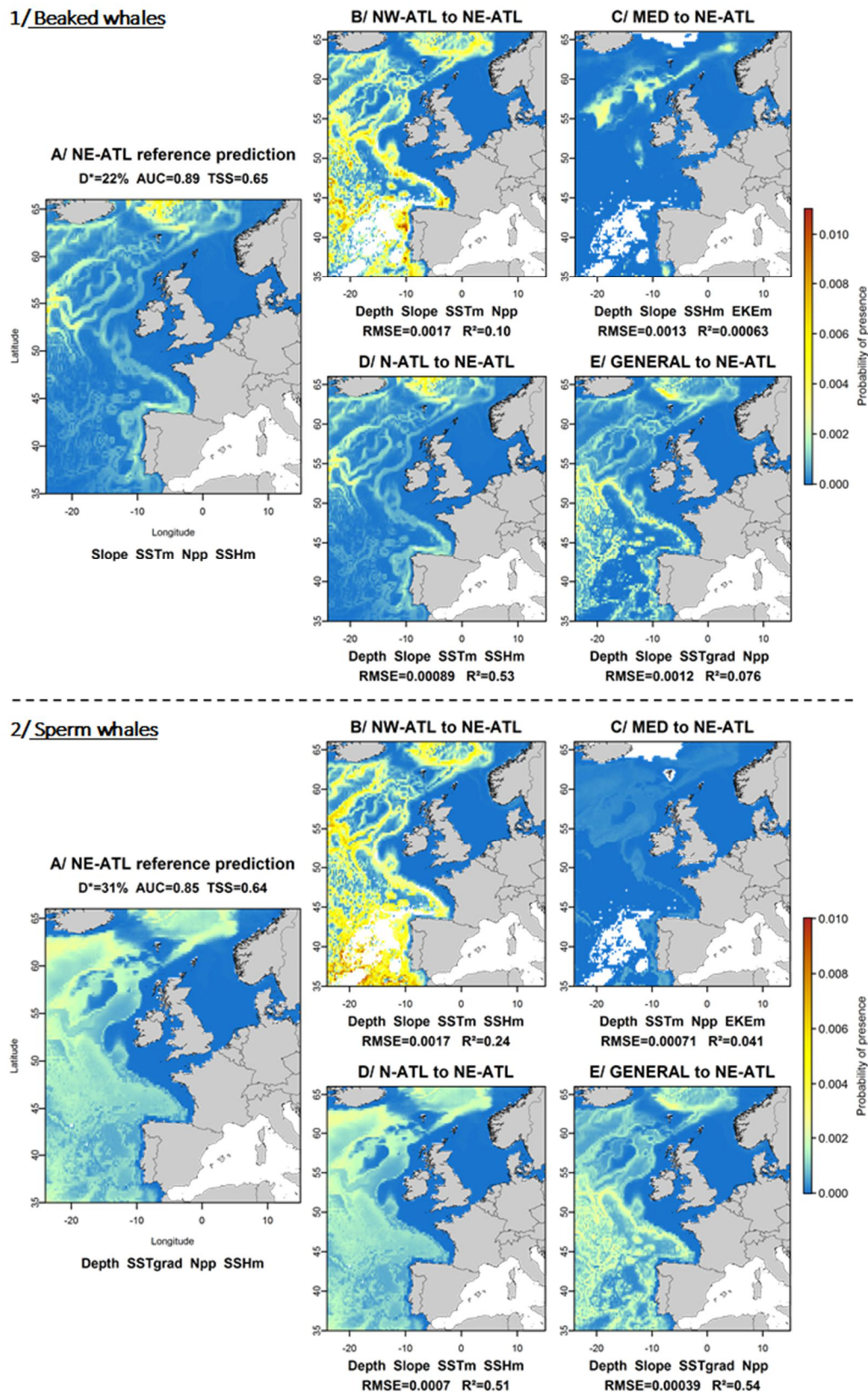


Fig. 5.4. Predictions provided by each model in the NE-ATL for beaked whales (1) and sperm whales (2). A/ is the reference prediction, B-D/ are the experimental predictions. Titles of the plots indicate the model used to predict species habitats in the NE-ATL. D\*: explained deviance; AUC: area under the curve; TSS: True skill statistic. Variables selected by each model are listed below the plots (*cf.* table 5.1). RMSE: Root Mean Squared Error; R<sup>2</sup>: coefficient of determination. They are calculated between the values of the experimental predictions and the reference prediction.

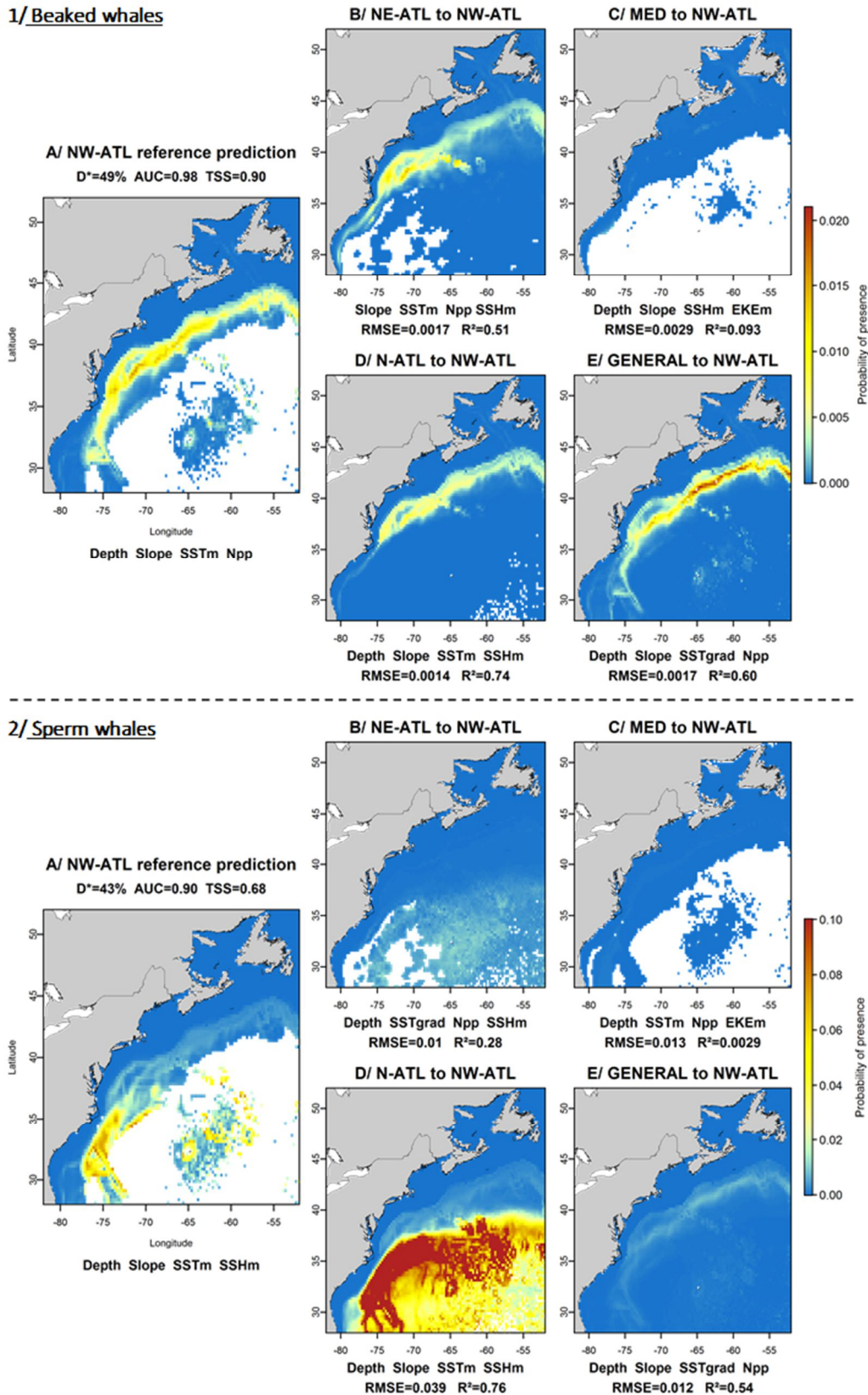


Fig. 5.5. Predictions provided by each model in the NW-ATL for beaked whales (1) and sperm whales (2). A/ is the reference prediction, B-D/ are the experimental predictions. Titles of the plots indicate the model used to predict species habitats in the NW-ATL. D\*: explained deviance; AUC: area under the curve; TSS: True skill statistic. Variables selected by each model are listed below the plots (*cf.* table 5.1). RMSE: Root Mean Squared Error; R<sup>2</sup>: coefficient of determination. They are calculated between the values of the experimental predictions and the reference prediction.



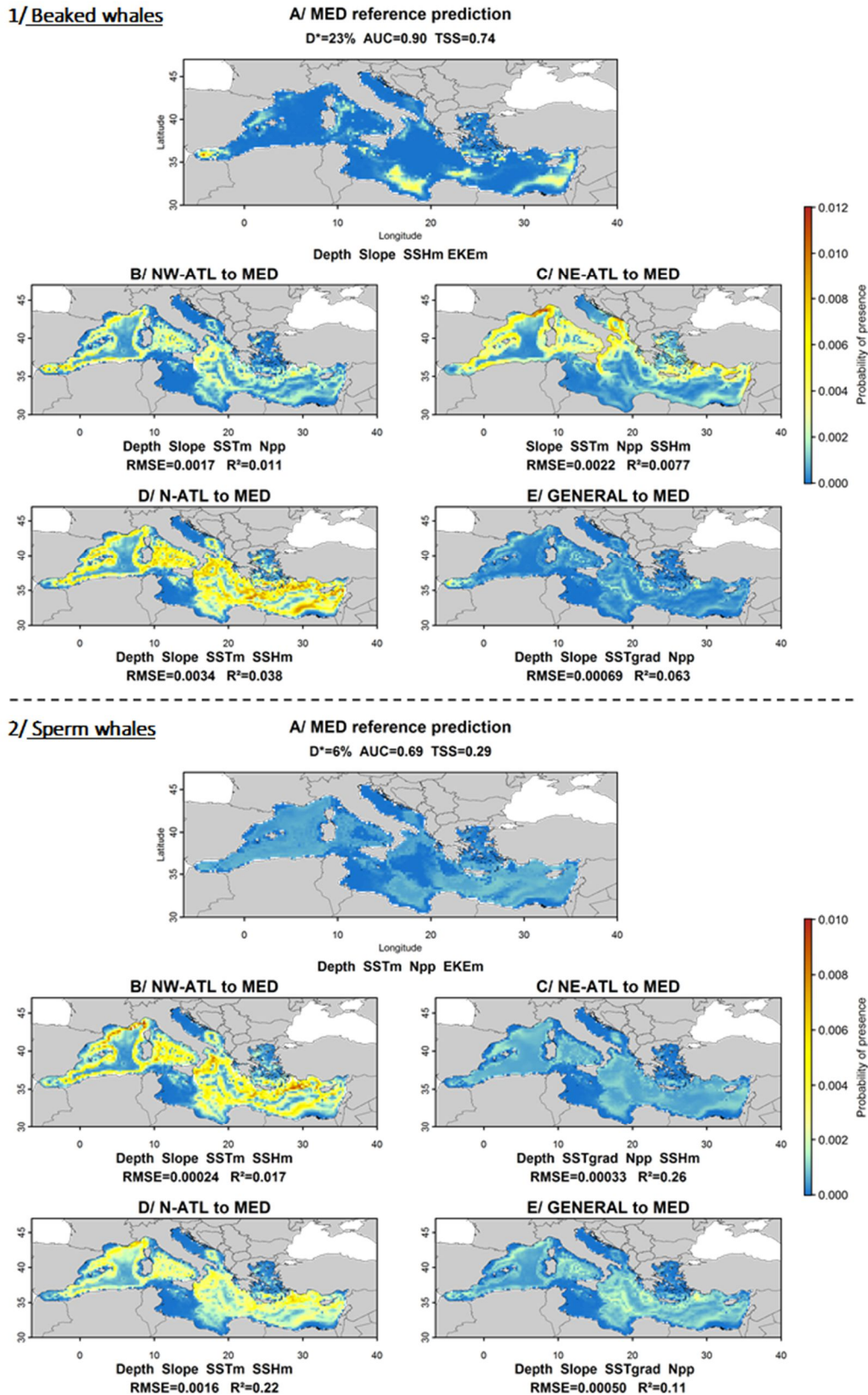


Fig. 5.6. Predictions provided by each model in the MED for beaked whales (1) and sperm whales (2). A/ is the reference prediction, B-D/ are the experimental predictions. Titles of the plots indicate the model used to predict species habitats in the MED. D\*: explained deviance; AUC: area under the curve; TSS: True skill statistic. Variables selected by each model are listed below the plots (*cf.* table 5.1). RMSE: Root Mean Squared Error; R<sup>2</sup>: coefficient of determination. They are calculated between the values of the experimental predictions and the reference prediction.

These results suggested that the Mediterranean Sea should be treated as a separate entity from the North Atlantic Ocean, whereas both side of the North Atlantic would be better considered jointly. By adding the MED data, predictions were less convincing than N-ATL models in the NW-ATL and NE-ATL and NW-ATL, NE-ATL and N-ATL models were not able to reproduce the prediction reference patterns in the MED.

## 5.4 CAN DIFFERENT ECOSYSTEMS BE ASSEMBLED?

Deep-diver habitats were predicted at a large scale by combining data collected in multiple ecoregions. Even if precautions were taken, by only predicting species habitats in interpolation zones where the combination of variables selected by the models were sampled, relationships between the species and their habitats may vary between the various ecosystems sampled. Consequently, through a model transferability approach, I wanted to assess if drivers of deep-diver habitats changed between the three most extensively sampled regions of the study area. In addition, I wanted to examine if merging data collected in multiple ecosystems would improve model predictions, as suggested by Redfern et al. (2017).

Heikkinen et al. (2012) and Duque-Lazo et al. (2016) showed a fairly good transferability of GAMs compared to other species distribution models such as boosted regression trees or classification and regression trees. Here, I showed that this transferability strongly depended on the regions between which the models were transferred. For the two species groups, NE-ATL models provided good predictions in the NW-ATL, NW-ATL models provided less convincing predictions in the NE-ATL and all models were poor at predicting species distribution patterns in the Mediterranean region. Models in the North Atlantic Ocean were thus not able to reproduce habitat specificities of the Mediterranean Sea with the data sampled in the North Atlantic Ocean while they were able to reproduce the habitats in both sides of the North Atlantic Ocean. This may be due to the particular characteristics of the Mediterranean Sea with its higher temperatures and salinity and lower productivity than in the North Atlantic Ocean (Bethoux et al. 1999; Longhurst 2007; Pujo-Pay et al. 2011) or due to differences in ecological processes between the regions. The seasonal cycle of primary production and consumption in the Mediterranean Sea is actually more similar to the cycle of the subtropical Atlantic Ocean (GFST and NAST provinces; Fig. 5.1) than to the cycle of the northeast Atlantic Ocean (NADR province; Fig. 5.1; Longhurst 2007). Additionally, topography is more similar between the two sub-basins of the Atlantic Ocean than between the Mediterranean basin and the Atlantic basins (Longhurst 2007). Consequently, topography would be the main factor that allows model transferability between the two Atlantic sub-basins.

In addition, variables selected in the five models were different, consequently I can suppose that drivers of deep-divers habitats differed between regions. As Roberts et al. (2016) in the NW-ATL, Rogan et al. (2017) in the NE-ATL and Cañadas and Vázquez (2014), depth or slope were selected in each model, suggesting a constant affinity of deep-divers for topographic features in the Atlantic sub-basins. In contrast relationships with these variables differed in the MED. In addition, the activity of eddies, that are predominant along the Mediterranean coasts (Pinaridi and Masetti 2000; Sarda et al. 2004; Tanhua et al. 2013), seemed to drive deep-diver habitats in the Mediterranean Sea, since eddy kinetic energy was only selected in the MED models for the two species groups. All these results suggested that models are transferable between similar ecoregions (such as NE-ATL and NW-ATL), as shown by Vanreusel et

al. (2007) but not transferable between regions that differ too much in their ecological processes (Randin et al. 2006; Torres et al. 2015; Redfern et al. 2017).

## 5.5 WHAT SCALE TO CONSIDER FOR DATA-ASSEMBLING?

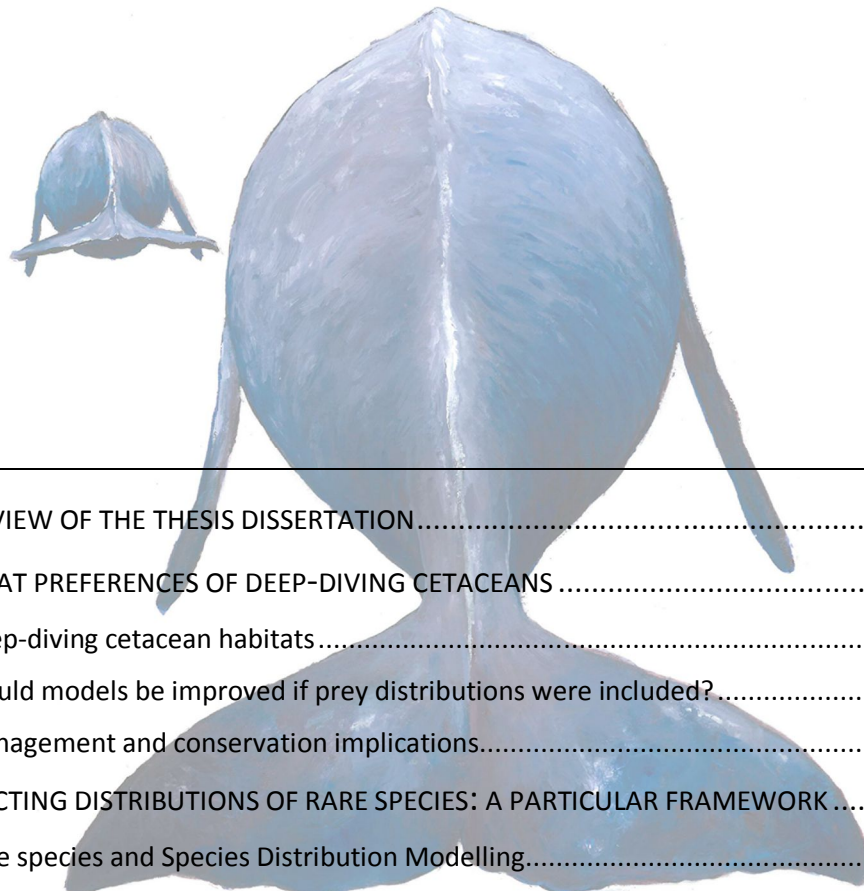
In this study, the effectiveness of data-assembling to model species habitats was confirmed. At the ecoregion scale, models fitted to merged data showed a very good performance (high explained deviances, AUCs and TSS). They provided good predictions of species habitats, consistent with previous studies (Roberts et al. 2016; Mannocci et al. 2017b; Rogan et al. 2017), with highest probabilities of presence predicted on the continental slopes for beaked whales and from the continental slopes to oceanic waters for sperm whales. This confirmed that data-assembling was a very good approach to model species habitats, in particular for rare species.

At the basin scale (North Atlantic Ocean and Mediterranean Sea), I noticed unexpected results. N-ATL models showed a better performance than GENERAL models to predict species habitats in the NW-ATL and NE-ATL. In contrast, both were poor at predicting species habitats in the MED. This suggested that the Mediterranean Sea was ecologically too different from the Atlantic Ocean to be included in the basin scale model (Bethoux et al. 1999; Longhurst 2007; Pujo-Pay et al. 2011). The separation of the Mediterranean Sea from the Atlantic Ocean proposed by Longhurst (2007) was reflected here. In contrast, the different provinces within the Atlantic Ocean were not clearly revealed, maybe because Atlantic regions of the study (NW-ATL and NE-ATL) encompassed various Longhurst's provinces (for example, the NADR and NAST(E) provinces in the NE-ATL; Fig. 5.1) and thus multiple ecosystems shared by the different provinces of the North Atlantic Ocean belonging to the same westerlies biome. This would suggest that predictions extrapolated in ecologically different regions such as the Caribbean Sea or upwelling zones along the African coasts, which does not belong to the same biome as northwest and northeast Atlantic sub-basins, where there was no survey effort, have to be considered with caution; additional surveys could help validating these predictions.

Finally, data-assembling is a matter of ecosystems similarity. At the ecoregion scale, where ecosystem processes are spatially consistent, data-assembling is strongly advised to model species distribution, even of common species, because the more data we have, the better the predictions we obtain. At a larger scale, drivers of species habitats may differ between ecosystems and we must be aware of the data-assembling limitations. It is wiser to assemble data from regions with similar ecological processes and to consider independently disparate regions such as coastal seas or very dynamic local areas. In any case, data-assembling allow to extrapolate species distribution patterns in larger areas, whilst remaining in interpolation areas, which could eventually help habitat modellers creating models at a global scale and improve species conservation.

# Chapter 6

## GENERAL DISCUSSION



© Laura Hedon

### CONTENTS

6.1 OVERVIEW OF THE THESIS DISSERTATION.....	86
6.2 HABITAT PREFERENCES OF DEEP-DIVING CETACEANS.....	89
6.2.1 Deep-diving cetacean habitats.....	89
6.2.2 Would models be improved if prey distributions were included?.....	90
6.2.3 Management and conservation implications.....	93
6.3 PREDICTING DISTRIBUTIONS OF RARE SPECIES: A PARTICULAR FRAMEWORK.....	94
6.3.1 Rare species and Species Distribution Modelling.....	94
6.3.2 Different assembling strategies to predict species distribution.....	95
6.3.3 A matter of caution.....	97
6.4 PERSPECTIVES.....	98
6.4.1 A finer scale resolution and processes in depth.....	98
6.4.2 More surveys and fewer gaps.....	99
6.4.3 What if we put all available data together?.....	99

**T**HE strategy adopted in this thesis and the effects of using prey distribution variables in the models are discussed in this chapter. I deepen the challenges of statistical modelling applied to rare species and the management implications of the present work as well as various perspectives.

## 6.1 OVERVIEW OF THE THESIS DISSERTATION

As stated in the introduction, rare species are by definition characterised by low encounter rates (Cunningham and Lindenmayer 2005). They can be threatened by natural and anthropogenic pressures and play a key role in both terrestrial and marine ecosystems (Lyons et al. 2005). Even though these species are generally protected, they often remain poorly known. Indeed, little is known about their distribution, abundance or migration patterns. To improve knowledge, habitat modelling is a major tool for predicting their distribution (Franklin 2010). However, low encounter rates of rare species make it difficult to use these models which generally need large datasets to properly fit, *i.e.* have an acceptable explained deviance while keeping overfitting and uncertainties as low as possible (Welsh et al. 1996; Barry and Welsh 2002; Cunningham and Lindenmayer 2005).

In this context, the aim of this thesis was to find a habitat model methodology that would be suitable for rare species in order to predict their habitats. The methodology adopted in this thesis proposes solutions that allowed the problem of low sample size to be circumvented. By first comparing various habitat models and then by using them with increasingly small datasets, I found that GAMs with a Tweedie distribution best fitted datasets with few sightings and that count-based models were generally better at predicting top predator distribution than presence-only models. Although GAMs with a Tweedie distribution proved to be suitable for datasets with few sightings, a minimum of 50 sightings was proposed to be required to obtain reliable predicted densities with this model. Reaching such a sample size threshold can be a challenge for many studies on rare species and therefore assembling multiple data sources can be the only possible way forward to reach this threshold.

Comparing models is fairly common in habitat modelling (*e.g.* Brotons et al. 2004; Leathwick et al. 2006; Tsoar et al. 2007; Gormley et al. 2011; Martínez-Rincón et al. 2012; Opperl et al. 2012) but it is uncommon when it comes to scarcely detected species (Welsh et al. 1996; Ridout et al. 1998). While the present model comparison procedure was not exhaustive and other models could be tested (see below), the most commonly used models for top marine predator habitats have been considered. Similarly, the sighting thinning experiment developed in Chapter 3 is an interesting and innovative approach, which allowed to determine the model that would best suit rare species data. However, this analysis would probably require further investigation. Indeed, the main limitation of this work is the limited number of species on which the methodology was tested. To make robust recommendations on the number of sightings needed to make reliable predictions, the methodology should be tested on a larger number of examples. However, due to the limited duration of the project, only two cases were tested. Nevertheless, this allowed me to adapt the methodology in the following chapters. In addition, I am aware that a model fitted to a complete dataset would be different than the one fitted to a thinned dataset but my objective was here to assess the ability of the models to predict distribution patterns and not their ability to explain the ecological processes (Elith and Leathwick 2009; Shmueli 2010). Of course ecological understanding is necessary to analyse causal drivers of species distributions (Mac Nally 2002), this is why I carried out a variable selection procedure in the baseline models (with the complete datasets). However, the assessment of model prediction performance was essential in this study because another objective of the thesis was to predict the potential habitats of deep-divers at the scale of the North Atlantic and Mediterranean basins. In this context, the GAM with a Tweedie distribution proved to be the most efficient.

In the present case, available data for deep-diving cetaceans, sperm, beaked and kogiid whales, were too scarce to properly model habitat solely from data collected within French Atlantic and Mediterranean waters. Therefore, it was decided to assemble different surveys from a larger study area to increase this number of sighting data. Indeed, over the past few years, data-assembling has been increasingly used for the study of top marine predators and allowed to efficiently model distributions at the basin scale (Winiarski et al. 2014; Roberts et al. 2016; Mannocci et al. 2017b; Rogan et al. 2017). After a fairly complex data homogenisation process designed so as to cope with disparate line-transect survey protocols, the data-assembling procedure allowed to gather a large number of sightings, which would be impossible to obtain from any single survey for species such as beaked whales, sperm whales and kogiids. The procedure was performed at a very large scale (North Atlantic Ocean and Mediterranean Sea), which implied assembling very different surveys and required to develop a meta-analysis to take into account the various survey protocols. The meta-analysis is commonly used to link the results of different studies (Gurevitch et al. 2001; Higgins et al. 2009) but here it has been applied to relate data from different surveys to estimate the Effective Strip Widths (ESW) of each survey. Usually, ESW is estimated using the conventional distance sampling methodology (Marques and Buckland 2003; Buckland et al. 2015), which was the basis of the meta-analysis procedure, but here, because of the low number of data, I could not fit a detection function for each survey. The meta-analysis was useful here because, for a particular combination of parameters, the estimated detection functions were shrunk towards a common detection function (itself estimated from the data) according to the available data corresponding to this combination of parameters. Consequently, even if each survey provided few sightings, an ESW could be estimated. This innovative methodology is needed if we want to pool datasets with different data acquisition protocols. The procedure is fairly complex and probably worth discussing, but it is an approach to which habitat modellers should probably converge.

Assembled data were then used to model deep-diver habitats by using a GAM with a Tweedie distribution and to provide predicted density maps across the Atlantic and Mediterranean basins. For the three species groups, highest densities were predicted in waters from c. 1500-4000 m and close to thermal fronts. Predictions identified areas of concentration along the continental slopes, in particular in the western North Atlantic Ocean along the Gulf Stream. Here, assessing prediction uncertainty was necessary because datasets included a lot of absences and the study area was very large. Generally, uncertainty is assessed by means of coefficients of variation (Mannocci et al. 2014a; Roberts et al. 2016; Lambert et al. 2017a), which measure the dispersion of the results around the mean. In addition to using this approach, and considering that only 5% of the study area (1°N-65°N; 82°W-40°E) were covered by the aggregated surveys, my objective was also to delineate interpolation zones where the combination of variables selected by the model was actually sampled, thus defining areas where predictions would be the most reliable. Consequently, I delineated, through an environmental space coverage gap analysis, two confidence thresholds in the predictions which defined areas where the predictions were environmental interpolations and thus reliable despite the local absence of sampling effort. The simple interpolation considered the full range of sampled variables to identify all points of the whole study area where the actual combinations of environmental variables had been sampled in survey blocks while in the precautionary interpolation, 5% of the extreme values of the sampled variables were removed to include in the interpolation areas only the points whose associated combinations of covariates fell within the 95% core ranges sampled. This revealed that deeper areas of the North Atlantic gyre were mostly areas of environmental extrapolation where no prediction should be made. To my knowledge, this procedure has been little applied so far in marine habitat modelling but deserves special attention,

especially if we converge towards data-assembling methods because even if we assemble a lot of data, extrapolation and confidence in predictions remain a major challenge.

Although data-assembling proved to be helpful to model deep-diver habitats, some limitations remain. The most obvious limit is sampling bias. Indeed, because most surveys were concentrated on the continental shelves and slopes, sampling remained low in oceanic areas and especially in the middle of the North Atlantic gyre, which revealed gaps in these regions through the gap analysis. In addition, due to the low number of animals sighted in each survey and because most survey effort was conducted during summer, I did not have enough data to fit a model per month, per year or even per season. I was able to provide general species distribution patterns over the 18 years of surveys but unfortunately I could not explore seasonal or annual changes in species distributions.

Moreover, I have only assembled datasets from visual surveys for which data collection protocols were comparable, if not identical. I made this choice because I was limited in time but it could be considered to assemble datasets from different sources (*e.g.* telemetry, acoustic, see below) to supplement knowledge of these species distributions. Additionally, it would be interesting to assess availability bias for these species groups (*i.e.* the proportion of animals missed by observers during their dive cycle; Marsh and Sinclair 1989; Barlow 2015). Indeed, in this study I only provided relative densities instead of absolute densities because it is difficult to evaluate the availability bias for deep-divers and thus to correct estimated densities by a factor which mostly depends on the immersion time of the animal. Following Laran et al. (2017b), I could apply a correction factor by using the proportion of time spent at surface by the animal (namely 20% for sperm whales, 10% for kogiids and 9% for beaked whales). These densities should also be corrected by a perception bias (*i.e.* observers' failure to detect animals present at the surface; Marsh and Sinclair 1989; Barlow 2015) that is specific to each survey but was not available here. Another limitation that could be identified in this study, and which is common in habitat modelling (Austin 2002; Guisan and Thuiller 2005), is the omission of possible interactions between variables in the models. Indeed such interactions could exist and their incorporation in models could improve the model fitting (Guisan et al. 1999; Thuiller et al. 2003). However, these interactions are difficult to interpret, and above all, would add a considerable number of parameters in the models (Rushton et al. 2004). To limit the number of parameters, it would be necessary to have an *a priori* idea of the possible interactions between these variables, which was not the case and that is why interactions have not been considered here.

Finally, as I pooled data collected in different ecosystems, even within interpolation areas, habitat drivers may vary from one ecosystem to another. Even if sampling effort is massive, differences between ecosystems can lead to very different relationships. Therefore, working at a very large scale may mask specific relationships between the taxon of interest and its environment that would operate at more local scales. Consequently, I wanted to assess if, even within interpolation areas, habitat drivers were similar between the different ecosystems (or regions). Through a model transferability approach, which consisted in fitting a model to the data from one region (*e.g.* the northeast Atlantic Ocean) and predicting probabilities of presence in another region (*e.g.* the northwest Atlantic Ocean), it appeared that models of the two North Atlantic Ocean sub-basins were not transferable in the Mediterranean Sea (and vice versa) while they were transferable from one sub-basin to the other. In addition, models that pooled data from the eastern and western North Atlantic sub-basins were better than models that pooled data from these two sub-basins and the Mediterranean Sea. This suggested that the Mediterranean Sea was ecologically different from the North Atlantic Ocean and that data-assembling would be more recommended in similar regions than in disparate regions where habitat drivers may

differ. Therefore, although data-assembling is strongly advised to model species distribution, I would recommend to perform it within similar ecosystems. Overall, if variables selected by models fitted to different regions are similar, these regions can be included in the same analysis, whereas if selected variables are radically different, data from these different regions should not be merged. This study should be continued, as it would be interesting to test whether the models would be transferable from one ocean to another in similar biomes or eco-regions (*e.g.* Atlantic and Pacific oceans) or if these drivers vary within a particular region. Particularly, it would be possible that habitat drivers differ between coastal and oceanic habitats. However, data used in the present study were not good candidates to assess this effect because deep-divers are mostly absent from shelf habitats.

## 6.2 HABITAT PREFERENCES OF DEEP-DIVING CETACEANS

### 6.2.1 Deep-diving cetacean habitats

The main objective of this thesis was to find a suitable methodology to model rare species habitats and to apply this methodology to deep-diving cetaceans. Thanks to the data-assembling procedure, I was able to provide the first relative density maps of beaked whales, sperm whales and kogiids in the North Atlantic Ocean and the Mediterranean Sea but also to highlight the importance of static (*i.e.* physiographic) and dynamic (*i.e.* oceanographic) variables in deep-diver habitat use. For the three species groups, depth and slope were significant variables that drove habitat selection and many studies have shown the importance of topographic features, such as slope or canyons, for deep-divers (Fergusson et al. 2006; MacLeod et al. 2011; Whitehead 2013; Wong and Whitehead 2014; Roberts et al. 2016; Rogan et al. 2017). However, few of them revealed the importance of dynamic structures, especially of the thermal fronts. Indeed, here gradients of temperatures significantly contributed to the models indicating that deep-divers would concentrate in areas of strong thermal fronts.

By studying the large scale distribution of the species, I highlighted an unexpected density gradient between the Atlantic Ocean and the Mediterranean Sea but also between the two North Atlantic sub-basins. Indeed, the maximum densities predicted in the Atlantic Ocean were 5 times for sperm whales, 7 times for kogiids and 12 times for beaked whales, higher than those predicted in the Mediterranean Sea. Similarly, the maximum densities predicted in the northwest Atlantic Ocean were 5 times for sperm whales, 7 times for kogiids and nearly 10 times for beaked whales, higher than those predicted in the northeast Atlantic Ocean. This difference could be due to the active frontal zone generated by the Gulf Stream in the northwest Atlantic Ocean in which deep-divers would concentrate (Griffin 1999; Waring et al. 2001; Hamazaki 2002).

In addition, at the basin scale (northwest and northeast Atlantic Ocean and Mediterranean Sea), I showed that for beaked whales and sperm whales, the most significant variables in the models varied from the general model (fitted to all data) but also between sub-basins, which was not particularly identified in studies carried out at the basin scale (Cañadas and Vázquez 2014; Roberts et al. 2016; Rogan et al. 2017). In these three studies, depth was a significant predictor but was the only parameter shared by the three studies. In contrast in Chapter 5, the model selection procedure was based on the same variables in the three regions and the selected variables were different from one region to another, which revealed a distinction in habitat drivers between the three regions. Indeed, while slope or depth variables have been selected in all three regions, temperature gradient was generally replaced by sea surface temperature and eddy activity as important drivers of deep-diver habitats in the



Mediterranean Sea compared to the other regions. Consequently, even if some habitat drivers are shared between regions, it should be kept in mind that habitat selection would ultimately depend on how local environmental conditions would determine target species prey distributions.

### 6.2.2 Would models be improved if prey distributions were included?

As mentioned in the introduction, environmental variables, such as depth, slope or sea surface temperature are supposed to be good indicators of the distribution of lower trophic levels and thus good proxies of the distribution of top predators (Ferguson et al. 2006; Redfern et al. 2006; Mannocci et al. 2014a). However, there is a time-lag between a change in an environmental condition and its effects on upper trophic levels (Jaquet 1996; Austin et al. 2006; Redfern et al. 2006; Cotté et al. 2009). Also, the relationships between these distal predictors and the actual quality of the habitat for a predator can vary with the underlying ecological processes (see above). The use of more proximal variables, such as prey distribution, could reduce these lags because marine top predators are supposed to be mostly sensitive to prey abundance (Österblom et al. 2008). Nevertheless, field data on prey distributions are not available at the scale of the Atlantic and Mediterranean basins and will not be so in a foreseeable future. To cope with this gap, a numerical model, the Spatial Ecosystem And POPulation DYNamics Model (SEAPODYM), provides simulation of distributions of zooplankton and six functional groups of micronekton at the global scale. It has been initially used to model tuna populations (Lehodey et al. 2008) but its usage was recently extended to predict turtle and cetacean habitats (Abecassis et al. 2013; Lambert et al. 2014). Consequently, in a work in progress (Annex E), I aimed to explore if models fitted by using data of prey distributions predicted better the deep-diver distribution than models using more conventional environmental data. I also aimed to explore if the combination of environmental and prey distribution data would further improve model results. To do that, for each taxon of deep-diving cetaceans (beaked whales, sperm whales and kogiids), I compared the performance of three models that used either environmental variables, or SEAPODYM variables or a combination of environmental and SEAPODYM variables (Figs. 6.1 and 6.2).

By comparing models built with SEAPODYM variables ('SEAPODYM model') and with environmental and SEAPODYM variables ('MIXED model') to the models built with only environmental variables ('ENVIRONMENTAL model'), it appeared that for beaked whales and sperm whales, SEAPODYM models were slightly better than ENVIRONMENTAL models and MIXED models were better than the two others for the three species groups. However the difference between model performances was not always significant and the best performance of MIXED models was not demonstrated for all species groups.

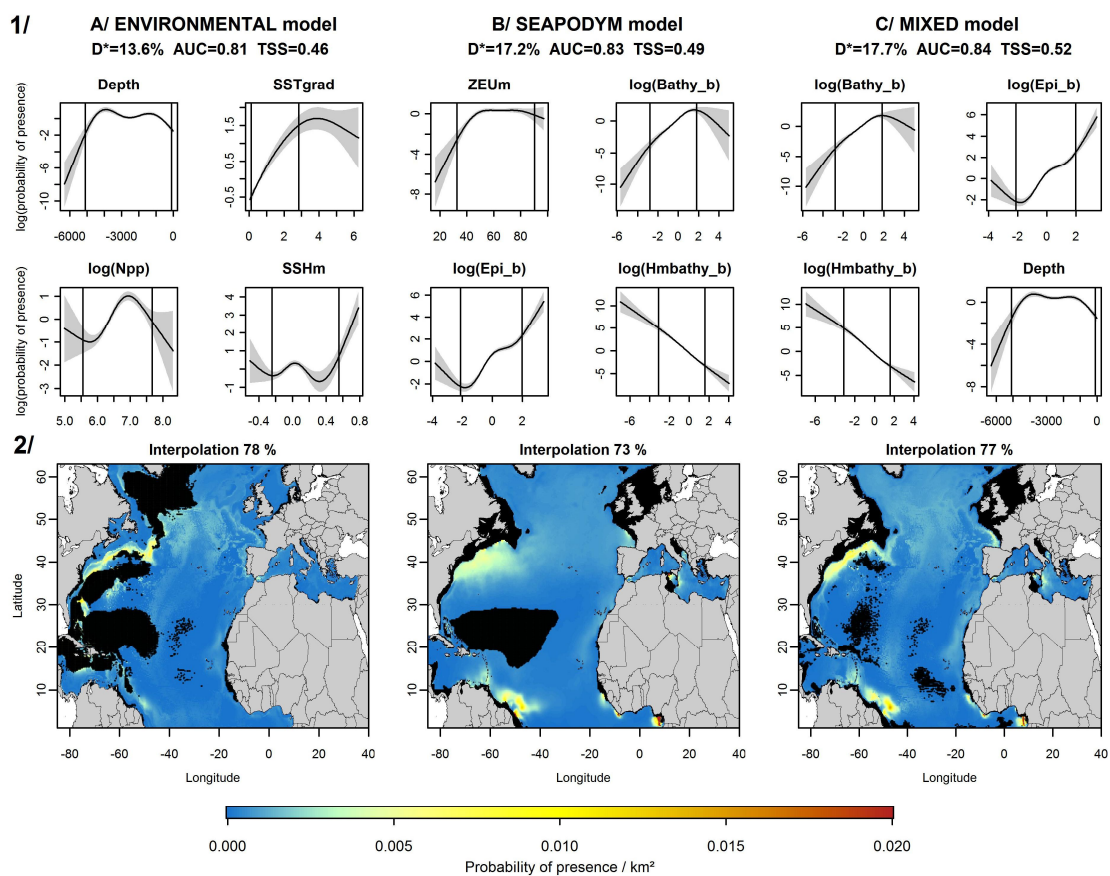
In addition, as for Lambert et al. (2014), prediction maps of species probabilities of presence were slightly smoother with the SEAPODYM variables compared to the environmental variables. This may be due either to the large resolution of the variables (0.25°) or to data averaging. Indeed, as in Chapter 4, SEAPODYM variables were monthly averaged (*i.e.* averaged over the 29 days prior to each sampled day), which would probably smooth predictions. A finer temporal resolution (*e.g.* weekly) may reduce this effect. However, by adding environmental variables in the models, such as depth, predictions were more structured. Thus, MIXED model results were encouraging.

The advantages of the SEAPODYM variables are that they are more proximal predictors of predator distributions than environmental variables and they can give an idea of the prey categories targeted by the species groups (Lehodey et al. 2008; 2010; 2014). In this specific case, the three species groups were closely related to high biomasses of bathypelagic and epipelagic organisms and low biomasses of highly

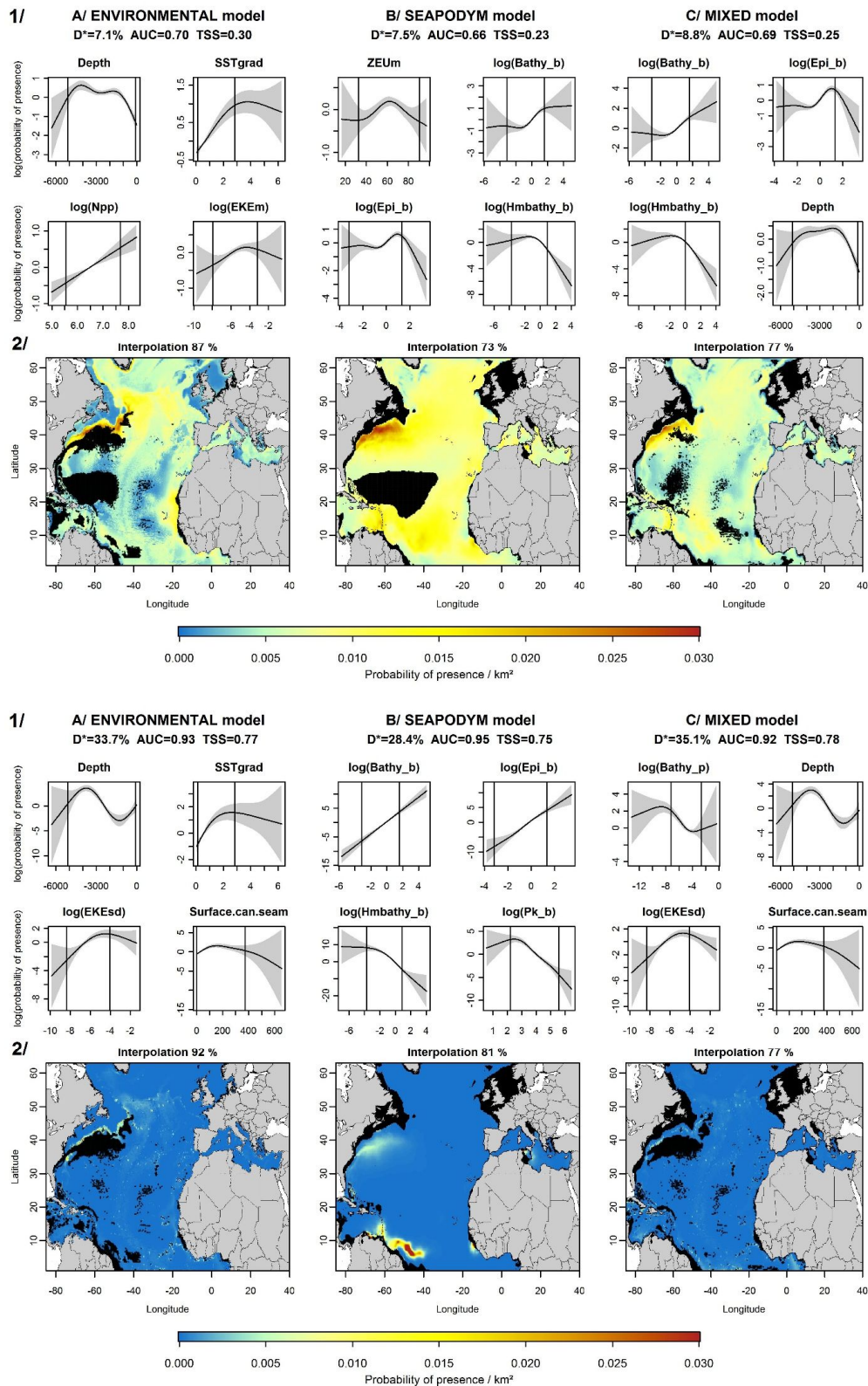
migrant bathypelagic species, which was consistent with the known diet of the species groups (Spitz et al. 2011).

The main issue with SEAPODYM variables was the absence of data over the continental shelf. This induced an absence of predictions on the continental shelf but also a loss of data, particularly for sperm whales which are regularly sighted at the edge of the continental shelf. This is why these ENVIRONMENTAL models were less performant than models described in Chapter 4 (lowest explained deviances). This effect might be a problem for more coastal species (*e.g.* small delphinids). Thus, the development of the SEAPODYM model over the continental shelf would be extremely helpful, although probably difficult because of interactions between the three pelagic layers of SEAPODYM with the epi-benthic and demersal species assemblages.

Finally, these results are encouraging and the combination of environmental and prey distribution variables could be a key to improve the predictions of species distributions through habitat models.



**Fig. 6.1.** Functional relationships for the selected variables (1/) and the predicted probabilities of presence of beaked whales (2/) for the ENVIRONMENTAL model (A), the SEAPODYM model (B) and the MIXED model (C). 1/: Solid lines are the estimated smooth functions, and the shaded regions represent the approximate 95% confidence intervals. The y-axes indicate the probabilities of presence on a log scale, where zero indicates no effect of the covariate. The vertical lines indicate the 2.5<sup>th</sup> and 97.5<sup>th</sup> quantiles of the data. D\*: explained deviance; AUC: Area Under the receiving Curve; TSS: True Skill Statistics. 2/: Black areas on prediction maps represent zones where we did not extrapolate the predictions. Percentages represent the proportion of the study area defined as interpolation with the gap analysis.



**Fig. 6.2.** Functional relationships for the selected variables and the predicted probabilities of presence of sperm whales (above) and kogiids (below) for the ENVIRONMENTAL model (A), the SEAPODYM model (B) and the MIXED model (C). 1/: Solid lines are the estimated smooth functions, and the shaded regions represent the approximate 95% confidence intervals. The y-axes indicate the probabilities of presence on a log scale, where zero indicates no effect of the covariate. The vertical lines indicate the 2.5<sup>th</sup> and 97.5<sup>th</sup> quantiles of the data. D\*: explained deviance; AUC: Area Under the receiving Curve; TSS: True Skill Statistics. 2/: Black areas on prediction maps represent zones where we did not extrapolate the predictions. Percentages represent the proportion of the study area defined as interpolation with the gap analysis.

### 6.2.3 Management and conservation implications

The management and conservation of species and ecosystems are increasingly based on habitat models. The ability of these latter to predict species occurrence in non-sampled or poorly documented areas is very useful (Fleishman et al. 2001, Lumaret and Jay-Robert 2002) because the implementation of dedicated survey is sometimes impracticable in these areas due to budgetary and logistical challenges of dedicated surveys. For example, in the framework of this study, it is almost impossible to carry out surveys dedicated to cetacean observation in the middle of the North Atlantic Ocean, however, by collecting data on both sides of the Atlantic Ocean, densities were predicted in this area.

Habitat models are also used to extrapolate the impact of global change on species distribution (Pearson and Dawson 2003, Cheung et al. 2009). For example, by projecting into 2050 the distribution of several fish species, Cheung et al. (2009) showed local extinctions due to climate change. According to the models fitted in this study, the deep-diver distribution was closely related to the sea surface temperature, thus an ocean global warming could lead to a change of this distribution, especially for kogiids which could display a northward shift in distribution.

Habitat models are also applied in conservation planning, which consists in designing protected area networks to protect biodiversity *in situ* (Cañadas et al. 2005, Louzao et al. 2006). For example, Cañadas et al. (2005) used generalised linear models to predict cetacean probabilities of occurrence and to propose Special Areas of Conservation and Specially Protected Area in the Mediterranean Sea.

However, these applications are disputed. One reason is that habitat models reflect the realised niche rather than fundamental species niche and generally do not explicitly account for biotic interactions (modelling cetacean SDM by using data or predictions describing prey field, as done above with the SEAPODYM predictions, is one exception). Therefore, projections into the future could not reveal the true expected distribution (Guisan and Thuiller 2005). Another reason is related to the very high uncertainty associated with the use of predicted species distribution data (Guisan and Thuiller 2005) which may lead to completely different protected area networks depending on the used model (Wilson et al. 2005). In addition, the feasibility and the usefulness of pelagic protected areas is also questioned, particularly for mobile species, because they would require the implementation of very broad protection zones that would be incompatible with anthropogenic activities but also with national jurisdictions and surveillance or management capabilities (Game et al. 2009).

In this thesis, I showed that deep-diving cetaceans are closely associated to stable topographic features, thus it could be possible to delineate marine protected areas according to Cañadas et al.'s (2005) procedure, which defined a Special Area of Conservation as an area that covered at least 60% of the principal habitats used by the species. However, these species are also responsive to dynamic structures at a large scale, such as thermal fronts, and it is not currently possible to obtain seasonal patterns or annual trends in species distribution. As a result, very large protected areas should be delineated, which would not be realistic both in terms of regulation in the high sea, and in terms of practicalities of its implementation. Nevertheless, in a Marine Spatial Planning approach (Douvere 2008), it would be worthwhile to overlay density maps with anthropogenic pressure maps (Halpern et al. 2008) to define areas where pressure should be minimised. In addition, it would be interesting to develop a dynamic ocean management (*i.e.* a real-time management), defined by Maxwell et al. (2015) as 'management that changes rapidly in space and time in response to the shifting nature of the ocean and its users based on the integration of new biological, oceanographic, social or economic data in near

real-time'. For example, to reduce turtle bycatch in the Central North Pacific, NOAA scientists used satellite tracking to determine temperature preferences of loggerheaded sea turtle (*Caretta caretta*) and regularly inform fishermen of the fishing areas they should avoid depending on the movements of temperature fronts (Maxwell et al. 2015). This methodology is probably currently not applicable to deep-divers because it requires real-time habitat predictions but the development of models using prey distribution data might help using such techniques in the future.

## 6.3 PREDICTING DISTRIBUTIONS OF RARE SPECIES: A PARTICULAR FRAMEWORK

### 6.3.1 Rare species and Species Distribution Modelling

As previously stated, the aim of Chapter 3 was to find a habitat model adapted to the study of rare species. Based on the literature, I initially selected some relevant models for testing, such as zero-inflated models (Welsh et al. 1996; Cunningham and Lindenmayer 2005), GAMs (Redfern et al. 2006; Mannocci et al. 2014a) or MaxEnt models (Elith et al. 2011). Contrary to Welsh et al. (1996) and Cunningham and Lindenmayer (2005), zero-inflated models showed less convincing results than GAMs with a Tweedie distribution, which appeared to be the best model for modelling rare species habitats, which could explain the growing use of the Tweedie distribution in ecology (Shono et al. 2008; Żydelis et al. 2011; Mannocci et al. 2015; Redfern et al. 2017)

However, it should be admitted that only a small number of models have been tested among all available models. Particularly, according to Latimer et al. (2006) the use of hierarchical models through a Bayesian framework (*i.e.* models in which data can enter at different stages and describing conceptual and unobservable processes) could provide more powerful inference about species distributions. Indeed, they proposed to build hierarchical models that included problems of irregular sampling intensity or spatial dependence. However, these techniques are rarely used in species distribution modelling and did not show more convincing results than non-hierarchical models (Elith and Leathwick 2009). In addition, I think it would be difficult to use them in the case of rare species because we have very little information that could be implemented in the models. In addition, in this work, I only used correlative (or niche-based) models that relate presence and absence data to sampled environmental variables at those sites but it could be also planned to use mechanistic (or process-based) models. These models assess the bio-physiological aspects of a species to generate the conditions in which the species can ideally persist, based on *in-situ* observations (Morin and Thuiller 2009). In Morin and Thuiller's (2009) study, contrary to niche-based models, mechanistic models proved to take into account the phenotypic plasticity and local adaption of tree species to predict species colonisation. However, these models require a large quantity of data that is not available for rare species and therefore could hardly be applied to deep-divers. Concerning presence-only models, I could have tested other models such as ecological niche factor analysis (Hirzel et al. 2002) or genetic algorithm for rule-set production (Stockwell and Peters 1999) but I wanted to focus my research on count-based models that showed better results than the presence-only models (Brotans et al. 2004). However, it appeared necessary to test a presence-only model because these models are commonly used in species distribution modelling (Zaniewski et al. 2002), in particular for rare species where opportunistic presence-only data can be seen as a useful complement of the scarce presence-absence data collected during dedicated sighting surveys. Even if MaxEnt models showed less convincing results than GAMs, I think they can be used for conservation

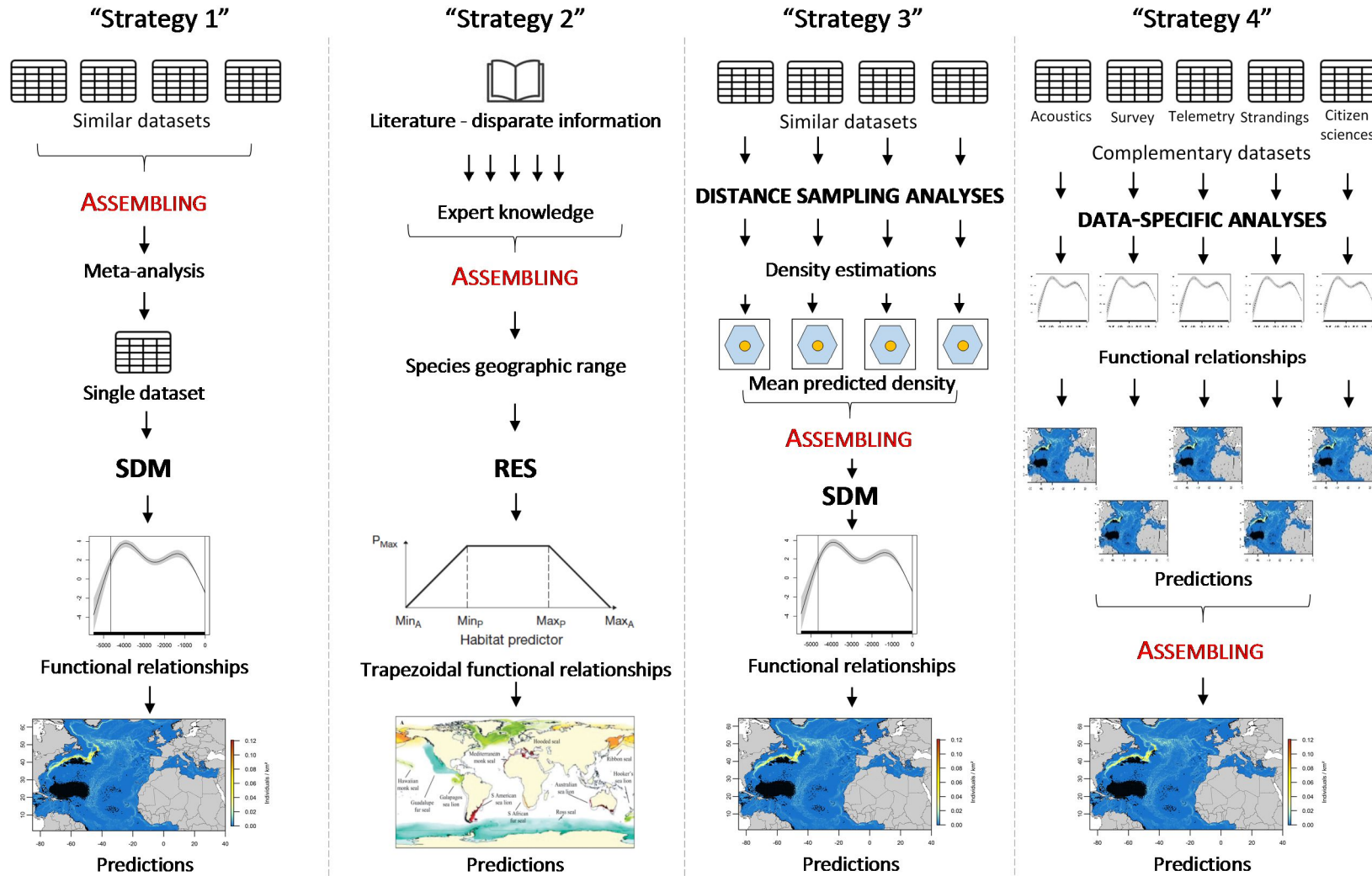
purposes but when effort data are available, presence-absence and count-based models should be prioritised.

By using GAMs with a Tweedie distribution, I did not claim that it was the optimal solution to model rare species habitats. Progress still needs to be made and the most obvious solution would be to intensify sampling effort. However, this model has proved to be relevant in this study, notably because it would limit the effect of over-dispersion and the overfitting of the data, inherent in the study of rare species (Hawkins 2004; Subramanian and Simon 2013). Generally, the explained deviances were fairly high (>25%) and the variables used seemed relevant, but the addition of other variables could improve the models. Species distribution is probably not only related to prey distributions but also to intra- or interspecific interactions (Guisan and Thuiller 2005). For sperm whales, females and their offspring seem to create groups while adult males would migrate episodically to meet females and ensure breeding (Engelhaulpt et al. 2009), which may lead to different distribution patterns that would not be explained only by the biological production of the oceans. Similarly, interspecific competition or interaction between beaked whales and sperm whales that feed on fairly similar prey and occupy fairly similar habitats could be hypothesised. Thus, there may be still unknown links which could improve the models. Consequently, a better knowledge thanks to telemetry data or citizen sciences data could help improving species distribution models for rare species. In the meantime, data-assembling seems to be an effective solution to overcome the problem.

### 6.3.2 Different assembling strategies to predict species distribution

Many studies that specifically deal with the distribution of deep-diving cetaceans are generally restricted to small study areas, *e.g.* Kiszka et al. (2007) in the Bay of Biscay, Claridge and Durban (2009) in the Great Bahama Canyon (Bahamas), Whitehead (2013) in the Gully submarine canyon (near the edge of the eastern continental shelf of North America) or Wong and Whitehead (2014) in the Sargasso Sea. These studies are of course essential for the conservation of deep-diving cetaceans but the knowledge of the large scale distribution of these species groups could improve their conservation, especially by identifying density hotspots (Boyd et al. 2008). However, in order to model large scale species distributions, it is necessary to assemble data (at different levels) because each individual survey is restricted to a comparatively small area and the smaller the sampled area, the smaller the interpolation area in which habitats can be predicted. In this context, based on the literature and discussions with colleagues, I identified four strategies of analysis that would allow to model large-scale species distributions (Fig. 6.3). The strategies proposed here, and throughout the thesis, are mainly applied in the marine environment but could of course be applied in the terrestrial environment.

Strategy 1 corresponds to the strategy developed in this thesis (Chapter 4). This strategy consists in directly assembling similar (if not identical) datasets, for example datasets collected in visual surveys. This requires to homogenise the various datasets and to perform a meta-analysis to take into account the different protocols. This results in a single dataset used as input to a species distribution model (SDM) which provides functional relationships and finally allow to predict the large-scale species distributions. This strategy has been increasingly used over the past few years on top marine predator datasets (Winiarski et al. 2014; Roberts et al. 2016; Mannocci et al. 2017b; Rogan et al. 2017).



**Fig. 6.3. Different assembling strategies to predict species distribution.** SDM: Species distribution model; RES: Relative environmental Suitability;  $Min_A$  and  $Max_A$ : absolute minimum and maximum predictor ranges;  $Min_P$  and  $Max_P$ : 'preferred' range, in terms of habitat usage of a given species; Mean predicted density (Strategy 3): density of individuals averaged for each dataset.

Strategy 2 corresponds to a strategy developed by Kaschner et al. (2006). This strategy is based on expert knowledge, extracted from the literature, which are assembled in order to determine the species geographic ranges, *i.e.* the maximum area between the known limits of a species occurrence. Through a Relative Environmental Suitability (RES) model, trapezoidal functional relationships, which described the predictor ranges used by species, are defined. These relationships are then used to generate a relative environmental suitability index (between 0 and 1) that scores how well each variable matched with the known species habitat use. Finally, this index is used to predict relative environmental suitability maps for each species.

As in Strategy 1, analyses of Strategy 3 are based on similar datasets (*e.g.* visual survey datasets). For each dataset, densities are estimated through a Distance sampling analysis over the survey area delineated by the dataset (Thomas et al. 2010) and then averaged to obtain a single average density value for each dataset to be linked to the geographical centre of the corresponding survey. Next, these data are assembled in a new dataset and used as input data for a new SDM. New functional relationships are obtained and densities are predicted at a large scale.

The principle of Strategy 4 consists in assembling prediction maps provided by completely different but complementary datasets. Each dataset is collected by a specific method (*e.g.* acoustics, telemetry, surveys, strandings, citizen sciences) on which are applied data-specific analyses (*e.g.* SDM, drift modelling) in order to obtain functional relationships and prediction maps. These maps are then assembled (*e.g.* averaged) to produce a large-scale prediction map. This strategy was for example used by Louzao et al. (2009) who combined Cory's shearwater (*Calecnoctris diomedea*) tracking data and shipboard survey data in the Mediterranean Sea or Thiers et al. (2014) who combined frigatebirds tracking data and aerial and shipboard survey data in the Mozambique Channel.

Each of these assembling strategies has advantages and limitations. In the four cases, it would be possible to predict species distributions from a local to a global scale. I hypothesise that result uncertainty would be higher in Strategies 2, 3 and 4 because uncertainties associated with each analysis have to be cumulated whereas in Strategy 1 uncertainty solely lies in a single model but a comparative approach would be very interesting to explore more acutely the potentials and limitations of each strategy.

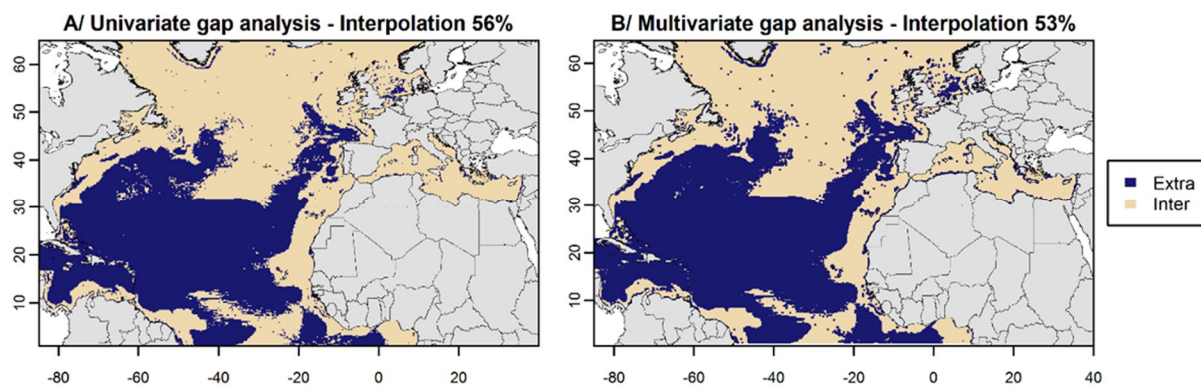
### 6.3.3 A matter of caution

As stated in the introduction, the scale to be considered in a study highly depends on the ecological question to which the study aims to answer (Mannocci et al. 2017a). Fine temporal and spatial scales identify local and ephemeral mechanisms that influence species distribution, while larger temporal and spatial scales identify distribution patterns associated with stable structures (Wu and Li 2006).

In this study I mainly focused on large-scale processes in order to determine the large-scale deep-diver distribution patterns. However, this involved taking precautions when extrapolating geographically. Indeed, due to specific species-environment relationships, it is difficult to predict these relationships in different ecosystems and could lead to highly uncertain predictions in non-sampled areas (Elith and Leathwick 2009; Elith et al. 2010). Consequently, to provide reliable predictions, it is necessary to remain within environmental interpolation, *i.e.* within the ranges of surveyed conditions. However, this highly depends on sampling effort, the larger the sampling scale, the more extended the sampled variable ranges.



The extent of environmental interpolation is usually delineated by overlaying the ranges of each variable selected in the habitat model ('univariate gap analysis'; Elith et al. 2010; Mannocci et al. 2015). In these studies, predictions were constrained within the envelope of the sampled variable values to limit extrapolation. However, this approach considers sampled variable ranges one by one but not the combination of these variables whereas a non-sampled combination of variables should be viewed as a true extrapolation even if each of the variable considered in isolation would fit in the corresponding sampled range (Zurell et al. 2012). Consequently, to provide the most reliable predictions, it is important to analyse the coverage of data collections in a multivariate approach ('multivariate gap analysis'; Elith et al. 2010; Dormann et al. 2012; Mesgaran et al. 2014), which was done in this work. To assess the effect of univariate *versus* multivariate gap analysis, the two methods were compared (Fig. 6.4). As expected, gaps were larger with the multivariate approach than with the univariate approach but in fairly low proportions. Therefore, the multivariate approach was more precautionary than the univariate one.



**Fig. 6.4. Univariate versus multivariate environmental space gap analysis.** The two analysis were performed with the same variables (depth, slope, gradients of sea surface temperature and net primary production). The univariate gap analysis consisted in simply overlay the ranges of sampled values of each variable while the multivariate gap analysis determined if the combination of the four variables was sampled. Percentages represent the proportion of the study area defined as interpolation with the gap analysis.

To sum up, the study of rare species is challenged by the low number of data which requires the use of suitable models. In this context, data-assembling appeared as a promising strategy which allows, even for the rarest species, to combine enough data to model species distribution (Roberts et al. 2016; Mannocci et al. 2017b; Rogan et al. 2017). However, caution must be taken to avoid extrapolating predictions by delineating areas of reliable predictions (Authier et al. 2016) or even assembling and predicting only in ecologically similar areas to ensure that species-environment relationships are conserved (Elith and Leathwick 2009; Elith et al. 2010).

## 6.4 PERSPECTIVES

### 6.4.1 A finer scale resolution and processes in depth

To improve deep-diver conservation, it was important to model their large-scale distributions. However, for the sake of operational practicality for end users, it would be essential to understand the relationships of deep-divers with their environment at a finer resolution. To do this, it would be interesting to examine a number of hotspots identified in Chapter 4 and 5 (*e.g.* canyon areas in the southern Bay of Biscay, in northeast Tyrrhenian Sea, near the Sable Island Bank or the Grand Bank in the

western Atlantic) and to model at high resolution habitat utilisation by deep-divers at these hotspots, notably by using new covariates selected for their capacity to express processes that could determine local prey availability. Considering that depth, slope and the presence of canyons have been identified as significant explanatory variables, it would be proposed that the interaction of these physiographic features with deep-sea circulation might play a key role.

In this context, it would be interesting to examine the physical characteristics of the canyons within the hotspot areas such as depth, slope, aspect and roughness to better describe the structure of the canyons by using the bathymetric data provided by Becker et al. (2009). Besides, it is known that flows within and around the canyons (*e.g.* upwelling, downwelling...) increase nutrient supply to the surface layer and thus increase the abundance of the preys in the ecosystem. Consequently, deep currents from ocean general circulation models and their fine-scale interactions with the local bathymetry might be a more direct proxy of deep-divers' prey availability than the more widely used surface oceanographic predictors are. This approach would allow to understand more precisely why deep-divers are concentrated within specific underwater structures and thus allow the development of more accurate mitigation or management strategies.

#### 6.4.2 More surveys and fewer gaps

It would be necessary to reduce the gaps identified in the gap analysis, since even if the assembled effort was substantial, it represented only a small part of the study area (about 5%). Particularly, by using a gap analysis, I defined two confidence thresholds in the predictions which highlighted that oceanic areas were poorly sampled, especially in the western Atlantic sub-basin. Consequently, by increasing sampling effort in these areas, information would be more comprehensive, analyses could be complemented and uncertainties associated with predictions in oceanic areas could be reduced.

However, another important avenue for progress would be to focus on the implementation of really standardised observation protocols. Currently, each organisation uses its own observation protocols that is based on shared general principles but differ in many operational details that prevent a straightforward dataset assemblage. This also leads to methodological choices such as dataset degradation to reach a common standard or dataset discard because protocols would be too divergent. Over the past few years, there has been a development of platform for data exchange on which datasets can be shared (*e.g.* OBIS SEAMAP - <http://seamap.env.duke.edu/>; Halpin et al. 2006; 2009), allowing me to identify datasets that would be interesting for my study. However, I think that observation protocols should also be shared in order to help standardising data collection and facilitate the development of data-assembling approaches.

#### 6.4.3 What if we put all available data together?

By reconsidering the four strategies proposed in 6.3.2 (Fig. 6.1), I think that Strategy 4 could prove to be very relevant if it was applied as a whole. So far, only certain methods have been assembled, mainly survey data and telemetry or tracking data (Louzao et al. 2009; Thiers et al. 2014), but each data collection method provides additional information such as diving profiles, causes and locations of animal death, number of individuals in the groups, etc. Therefore, it would be very interesting to be able to assemble all these data in order to provide the most reliable prediction maps. In addition, as shown in Strategy 1, instead of assembling the results of Strategy 4, it would be optimal to assemble the data of

Strategy 4. However, the current problem probably lies in the modelling strategy to be used. A way might be to use a multi-model approach by defining a common metric, such as abundance of individuals or presence and convert each dataset obtained from the different sources into this metric. For example, transforming an intensity of an acoustic signal or the time spent in a cell by a tagged animal into the common metric and then build a meta-model (Talluto et al. 2016). This approach is recent and represents an interesting perspective for future PhD students and researchers.



## REFERENCES

- Abecassis, M., Senina, I., Lehodey, P., Gaspar, P., Parker, D., Balazs, G., Polovina, J. (2013). A model of loggerhead sea turtle (*Caretta caretta*) habitat and movement in the oceanic North Pacific. *PLoS One* 8(9): e73274.
- Aïssi, M., Ouammi, A., Fiori, C., Alessi, J. (2014). Modelling predicted sperm whale habitat in the central Mediterranean Sea: requirement for protection beyond the Pelagos Sanctuary boundaries. *Aquatic Conservation: Marine and Freshwater Ecosystems* 24(S1): 50-58.
- Akaike, H., (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19(6): 716-723.
- Allouche, O., Tsoar, A., Kadmon, R. (2006). Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of applied ecology* 43(6): 1223-1232.
- Anderson, R.C., Corbett, E.A., Anderson, M.R., Corbett, G.A., Kelley, T.M. (2001). High white-tailed deer density has negative impact on tallgrass prairie forbs. *Journal of the Torrey Botanical Society* 128: 381-392.
- Anderson, R.P. (2012). Harnessing the world's biodiversity data: promise and peril in ecological niche modeling of species distributions. *Annals of the New York Academy of Sciences* 1260: 66-80.
- Apps, C.D., McLellan, B.N., Kinley, T.A., Flaa, J.P. (2001). Scale-dependent habitat selection by mountain caribou, Columbia Mountains, British Columbia. *The Journal of wildlife management*: 65-77.
- Arcangeli, A., Campana, I., Marini, L., MacLeod, C.D. (2015). Long-term presence and habitat use of Cuvier's beaked whale (*Ziphius cavirostris*) in the Central Tyrrhenian Sea. *Marine Ecology* 37: 269-282.
- Arcuti, S., Calculi, C., Pollice, A., D'Onghia, G. (2013). Spatio-temporal modelling of zero-inflated deep-sea shrimp data by Tweedie generalized additive. *Statistica* 73: 87-101.
- Arrêté du 1er juillet 2011 fixant la liste des mammifères marins protégés sur le territoire national et les modalités de leur protection NOR: DEVL1110724A - version consolidée au 31/08/2017.
- Atwood, T.C., Fry, T.L., Leland, B.R. (2011). Partitioning of anthropogenic watering sites by desert carnivores. *The Journal of Wildlife Management* 75(7): 1609-1615.
- Austin, M.P., (2002). Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling* 157: 101-118.
- Austin, D., Bowen, W.D., McMillan, J.I., Iverson, S.J. (2006). Linking Movement, Diving, and Habitat to Foraging Success in a Large Marine Predator. *Ecology* 87: 3095-3108.
- Authier, M., Saraux, C., Péron, C. (2016). Variable Selection and Accurate Predictions in Habitat Modelling: a Shrinkage Approach. *Ecography* 39: 001-012.
- Balcomb, K.C., and Claridge, D.E. (2001). A mass stranding of cetaceans caused by naval sonar in the Bahamas. *Bahamas Journal of Science* 8: 2-12.
- Ballance, L.T., Pitman, R.L., Fiedler, P.C. (2006). Oceanographic influences on seabirds and cetaceans of the eastern tropical Pacific: a review. *Progress in Oceanography* 69(2): 360-390.
- Bailey, H., and Thompson, P.M. (2009). Using marine mammal habitat modelling to identify priority conservation zones within a marine protected area. *Marine Ecology Progress Series* 378: 279-287.
- Baird, R.W., Nelson, D., Lien, J., Nagorsen, D.W. (1996). Status of the pygmy sperm whale, *Kogia breviceps*, in Canada. *Canadian Field-Naturalist* 110: 525-532.

- Barlow, J., Forney, K.A., Hill, K.A., Brownell, R.L., Caretta, R.L., Demaster, D.P., Julian, D.P., Lowry, M.S., Ragen, M.S., Reeves, R.R. (1997). U.S. Pacific marine mammal stock assessments: 1996. NOAA Technical Memorandum NMFS-SWFSC 248: 223 pp.
- Barlow, J. and Cameron, G.A. (2003). Field experiments show that acoustic pingers reduce marine mammal by-catch in the California drift gill net fishery. *Marine Mammal Science* 19(2): 265-283.
- Barlow, J., Ferguson, M., Perrin, W., Gerrodette, T., Joyce, G., Macleod, C., Mullin, K., Palka, D., Waring, G. (2006). Abundance and densities of beaked and bottlenose whales (family *Ziphiidae*). *Journal of Cetacean Research and Management* 7: 263-270.
- Barlow, J. (2015). Inferring trackline detection probabilities,  $g(0)$ , for cetaceans from apparent densities in different survey conditions. *Marine Mammal Science* 31(3): 923-943.
- Barnston, A.G. (1992). Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score. *Weather and Forecasting* 7(4): 699-709.
- Barry S.C., and Welsh A.H. (2002). Generalized additive modelling and zero inflated count data. *Ecological Modelling* 157(2-3): 179-188.
- Basille, M., Calenge, C., Marboutin, E., Andersen, R., Gaillard, J.M. (2008). Assessing habitat selection using multivariate statistics: Some refinements of the ecological-niche factor analysis. *Ecological Modelling* 211(1): 233-240.
- Beatson, E. (2007). The diet of pygmy sperm whales, *Kogia breviceps*, stranded in New Zealand: implications for conservation. *Reviews in Fish Biology and Fisheries* 17(2-3): 295-303.
- Becker, J.J., Sandwell, D.T., Smith, W.H.F., Braud, J., Binder, B., Depner, J. et al. (2009). Global bathymetry and elevation data at 30 arc seconds resolution: SRTM30\_PLUS. *Marine Geodesy* 32(4): 355-371.
- Becker, E.A., Forney, K.A., Ferguson, M.C., Foley, D.G., Smith, R.C., Barlow, J., Redfern, J.V. (2010). Comparing California Current cetacean-habitat models developed using in situ and remotely sensed sea surface temperature data. *Marine Ecology Progress Series* 413: 163-183.
- Bethoux, J.P., Gentili, B., Morin, P., Nicolas, E., Pierre, C., Ruiz-Pino, D. (1999). The Mediterranean Sea: A miniature ocean for climatic and environmental studies and a key for the climatic functioning of the North Atlantic. *Progress in Oceanography* 44: 131-146.
- Block, W.M., and Brennan, L.A. (1993). The habitat concept in ornithology. *Current ornithology* 11: 35-91.
- Boersma, P.D. (1998). Population trends of the Galápagos penguin: impacts of El Niño and La Niña. *Condor*: 245-253.
- Bosc, E., Bricaud, A., Antoine, D. (2004). Seasonal and interannual variability in algal biomass and primary production in the Mediterranean Sea, as derived from 4 years of SeaWiFS observations. *Global Biogeochemical Cycles* 18(1).
- Bost, C.A., Cotté, C., Bailleul, F., Cherel, Y., Charrassin, J.B., Guinet, C., Ainley, D.G., Weimerskirch, H. (2009). The importance of oceanographic fronts to marine birds and mammals of the southern oceans. *Journal of Marine Systems* 78: 363-376.
- Bowles, A.E., Smultea, M., Wursig, B., Demaster, D.P., Palka, D. (1994). Relative abundance and behavior of marine mammals exposed to transmissions from the Heard Island Feasibility Test. *Journal of the Acoustical Society of America* 96: 2469-2484.
- Boyd, C., Brooks, T.M., Butchart, S.H., Edgar, G.J., Da Fonseca, G.A., Hawkins, F. et al. (2008). Spatial scale and the conservation of threatened species. *Conservation Letters* 1(1): 37-43.

- Brandt, S.B. (1993). The effect of thermal fronts on fish growth: a bioenergetics evaluation of food and temperature. *Estuaries* 16: 142-159.
- Broadus, J.M., Ericson, D.B., Fleming, R.H., LaMourie, M.J., Barnes, C.A., Namias, J. (2009). Atlantic Ocean. *Encyclopedia Britannica*.
- Brotos, L., Thuiller, W., Araújo, M.B., Hirzel, A.H. (2004). Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography* 4: 437-448.
- Brownell, R.L., Yamada, T., Mead, J.G., Helden, A.L. (2004). Mass stranding of Cuvier's beaked whales in Japan: U.S. Naval acoustic link? *Journal of Cetacean Research and Management* 7: 1-10.
- Buckland, S.T., Rexstad, E.A., Marques, T.A., Oedekoven, C.S. (2015). *Distance sampling: methods and applications*. Springer.
- Caldwell, D.K., and Caldwell, M.C. (1989). Pygmy sperm whale *Kogia breviceps* (de Blainville, 1838): dwarf sperm whale *Kogia simus* Owen, 1866. *Handbook of marine mammals* 4: 235-260.
- Campbell, J.W., and Aarup, T. (1992). New production in the North Atlantic derived from seasonal patterns of surface chlorophyll. *Deep Sea Research Part A: Oceanographic Research Papers* 39: 1669-1694.
- Canada Meteorological Center. (2012). GHRST Level 4 CMC0.2deg Global Foundation Sea Surface Temperature Analysis (GDS version 2). Ver. 2.0. PO.DAAC, CA, USA.
- Cañadas, A., Sagarminaga, R., Garcia-Tiscar, S. (2002). Cetacean distribution related with depth and slope in the Mediterranean waters off southern Spain. *Deep Sea Research Part I: Oceanographic Research Papers* 49(11): 2053-2073.
- Cañadas, A., Sagarminaga, R., De Stephanis, R., Urquiola, E., Hammond, P. S. (2005). Habitat preference modelling as a conservation tool: proposals for marine protected areas for cetaceans in southern Spanish waters. *Aquatic Conservation: Marine and Freshwater Ecosystems* 15(5): 495-521.
- Cañadas, A., Donovan, G.P., Desportes, G., Borchers, D.L. (2009). A short review of the distribution of short beaked common dolphins (*Delphinus delphis*) in the central and eastern North Atlantic with an abundance estimate for part of this area. *NAMMCO Scientific Publications* 7: 201-220.
- Cañadas, A., and Vázquez, J.A. (2014). Conserving Cuvier's beaked whales in the Alboran Sea (SW Mediterranean): Identification of high density areas to be avoided by intense man-made sound. *Biological Conservation* 178: 155-162.
- Carrillo, M., and Ritter, F. (2010). Increasing numbers of ship strikes in the Canary Islands: proposals for immediate action to reduce risk of vessel-whale collisions. *Journal of Cetacean Research and Management* 11(2): 131-138.
- Certain, G., and Bretagnolle, V. (2008). Monitoring seabirds population in marine ecosystem: The use of strip-transect aerial surveys. *Remote Sensing of Environment* 112: 3314-3322.
- Cheung, W.W., Lam, V.W., Sarmiento, J.L., Kearney, K., Watson, R., Pauly, D. (2009). Projecting global marine biodiversity impacts under climate change scenarios. *Fish and fisheries* 10(3): 235-251.
- Chilvers, B.L. (2008). Foraging site fidelity of lactating New Zealand sea lions. *Journal of Zoology* 276(1): 28-36.
- Claridge, D.E., and Durban, J.W. (2009). Distribution, abundance and population structuring of beaked whales in the Great Bahama Canyon. *ONR Program Review*: 7-10.
- Clark, M. (2013). *Generalized additive models: getting started with additive models in R*. Center for Social Research, University of Notre Dame, 35.

- Clarke, M.R., Martins, H.R., Pacoe, P. (1993). The diet of sperm whales (*Physeter microcephalus* Linnaeus 1758) off the Azores. *Philosophical Transactions of the Royal Society, London B* 339: 67-82.
- Coll, M., Piroddi, C., Steenbeek, J., Kaschner, K., Lasram, F. B. R., Aguzzi, J. et al. (2010). The biodiversity of the Mediterranean Sea: estimates, patterns, and threats. *PloS One* 5(8) : e11842.
- Cotté, C., Guinet, C., Taupier-Letage, I., Mate, B., Petiau, E. (2009). Scale-dependent habitat use by a large free-ranging predator, the Mediterranean fin whale. *Deep Sea Research Part A. Oceanographic Research Papers* 56: 801-811.
- Cunningham, R.B., and Lindenmayer, D.B. (2005). Modeling Count Data of Rare Species. *Ecology* 86(5): 1135-1142.
- D'Amico, A., Gisiner, R.C., Ketten, D.R., Hammock, J.A., Johnson, C., Tyack, P.L., Mead, J. (2009). Beaked whale strandings and naval exercises. *Aquatic Mammals* 35: 452-472.
- Davoren, G.K., Montevecchi, W.A., Anderson, J.T. (2003). Search strategies of a pursuit-diving marine bird and the persistence of prey patches. *Ecological Monographs* 73(3): 463-481.
- De'ath, G. and Fabricius, K.E. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* 81: 3178-3192.
- De Ruijter, W.P., Van Aken, H.M., Beier, E.J., Lutjeharms, J.R., Matano, R.P., Schouten, M.W. (2004). Eddies and dipoles around South Madagascar: formation, pathways and large-scale impact. *Deep Sea Research Part I: Oceanographic Research Papers* 51(3): 383-400.
- DeRuiter, S. L., Southall, B. L., Calambokidis, J., Zimmer, W. M., Sadykova, D., Falcone, E. A. et al.. (2013). First direct measurements of behavioural responses by Cuvier's beaked whales to mid-frequency active sonar. *Biology letters* 9(4): 20130223.
- Dollhopf, S.L., Hashsham, S.A., Tiedje, J.M. (2001). Interpreting 16S rDNA T-RFLP data: application of self-organizing maps and principal component analysis to describe community dynamics and convergence. *Microbial Ecology* 42(4): 495-505.
- Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G. et al. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36(1): 27-46.
- Douve, F. (2008). The importance of marine spatial planning in advancing ecosystem-based sea use management. *Marine policy* 32(5): 762-771.
- Duffy, J.E. (2003). Biodiversity loss, trophic skew and ecosystem functioning. *Ecology Letters* 6: 680-687.
- Dunn, P.K., and Smyth, G.K. (2005). Series evaluation of Tweedie exponential dispersion model densities. *Statistics and Computing* 15(4): 267-280.
- Duque-Lazo, J., Van Gils, H.A.M.J., Groen, T.A., Navarro-Cerrillo, R.M. (2016). Transferability of species distribution models: The case of *Phytophthora cinnamomi* in Southwest Spain and Southwest Australia. *Ecological Modelling* 320: 62-70.
- Eakins, B.W. and Sharman, G.F. (2010). Volumes of the World's Oceans from ETOPO1, NOAA National Geophysical Data Center, Boulder, CO.
- Eckert, S.A. (2002). Distribution of juvenile leatherback sea turtle *Dermochelys coriacea* sightings. *Marine Ecology Progress Series* 230: 289-293.
- Edrén, S., Wisz, M.S., Teilmann, J., Dietz, R., Söderkvist, J. (2010). Modelling spatial patterns in harbour porpoise satellite telemetry data using maximum entropy. *Ecography* 33(4): 698-708.

- Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A. et al. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29: 129-151.
- Elith, J., Leathwick, J.R., Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology* 77: 802-813.
- Elith, J., and Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual review of ecology, evolution, and systematics* 40: 677-697.
- Elith, J., Kearney, M., Phillips, S. (2010). The art of modelling range-shifting species. *Methods in ecology and evolution* 1(4): 330-342.
- 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(1): 43-57.
- Elmhagen, B., Ludwig, G., Rushton, S. P., Helle, P., Lindén, H. (2010). Top predators, mesopredators and their prey: interference ecosystems along bioclimatic productivity gradients. *Journal of Animal Ecology* 79(4): 785-794.
- Engelhaupt, D., Rus Hoelzel, A., Nicholson, C., Frantzis, A., Mesnick, S., Gero, S. et al. (2009). Female philopatry in coastal basins and male dispersion across the North Atlantic in a highly mobile marine species, the sperm whale (*Physeter macrocephalus*). *Molecular Ecology* 18(20): 4193-4205.
- Engler, R., Guisan, A., Rechsteiner, L. (2004). An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology* 41(2): 263-274.
- ESRI, 2008. ArcGIS - A Complete Integrated System Environmental Systems Research Institute, Inc., Redlands, California. <<http://esri.com/arcgis>>.
- Evans, G.T., and Taylor, F.J.R. (1980). Phytoplankton accumulation in Langmuir cells. *Limnology and Oceanography* 25: 840-845.
- Evans, K., and Hindell, M. (2004). The diet of sperm whales (*Physeter macrocephalus*) in southern Australian waters. *ICES Journal of Marine Science* 61: 1313-1329.
- Fauchald, P., Erikstad, K.E., Skarsfjord, H. (2000). Scale-dependent predator-prey interactions: the hierarchical spatial distribution of seabirds and prey. *Ecology* 81(3): 773-783.
- Ferguson, M.C., Barlow, J., Reilly, S.B., Gerrodette, T. (2006). Predicting Cuvier's (*Ziphius cavirostris*) and Mesoplodon beaked whale population density from habitat characteristics in the eastern tropical Pacific Ocean. *Journal of Cetacean Research and Management* 7: 287-299.
- Fernández, E., and Pingree, R.D. (1996). Coupling between physical and biological fields in the North Atlantic subtropical front southeast of the Azores. *Deep Sea Research Part I: Oceanographic Research Papers* 43(9): 1369-1393.
- Fernández, A., Edwards, J.F., Rodríguez, F., Espinosa de los Monteros, A., Herráez, P. et al. (2005). "Gas and Fat Embolic Syndrome" Involving a Mass Stranding of Beaked Whales (Family *Ziphiidae*) Exposed to Anthropogenic Sonar Signals. *Veterinary Pathology* 42: 446-457.
- Flather, C.H., and Sieg, C.H. (2007). Species rarity: definition, causes and classification. *Conservation of rare or little-known species: Biological, social, and economic considerations*: 40-66.
- Fleishman, E., Nally, R.M., Fay, J.P., Murphy, D.D. (2001). Modeling and predicting species occurrence using broad-scale environmental variables: an example with butterflies of the Great Basin. *Conservation biology* 15(6): 1674-1685.



- Foster, S.D., and Bravington, M. V. (2013). A Poisson-Gamma model for analysis of ecological non-negative continuous data. *Environmental and ecological statistics* 20: 533–552.
- Franklin, J. (2010). *Mapping species distributions: spatial inference and prediction*. Cambridge University Press.
- Frantzis, A. (1998). Does acoustic testing strand whales? *Nature* 392.
- Game, E.T., Grantham, H.S., Hobday, A.J., Pressey, R.L., Lombard, A.T., Beckley, L.E et al. (2009). Pelagic protected areas: the missing dimension in ocean conservation. *Trends in ecology & evolution* 24(7): 360-369.
- Gaston, K.J. (1994). *Rarity*. London: Chapman and Hall.
- Goldbogen, J.A., Calambokidis, J., Croll, D.A., Harvey, J.T., Newton, K.M., Oleson, E.M., Shadwick, R.E. (2008). Foraging behavior of humpback whales: Kinematic and respiratory patterns suggest a high cost for a lunge. *Journal of Experimental Biology* 211: 3712-3719.
- Gormley, A.M., Forsyth, D.M., Griffioen, P., Lindeman, M., Ramsey, D.S., Scroggie, M.P., Woodford, L. (2011). Using presence-only and presence-absence data to estimate the current and potential distributions of established invasive species. *Journal of Applied Ecology* 48(1): 25-34.
- Griffin, R.B. (1999). Sperm whale distributions and community ecology associated with a warm-core ring off Georges Bank. *Marine Mammal Science* 15: 33-51.
- Guinan, J., Brown, C., Dolan, M.F., Grehan, A.J. (2009). Ecological niche modelling of the distribution of cold-water coral habitat using underwater remote sensing data. *Ecological Informatics* 4(2): 83-92.
- Guisan, A., Weiss, S.B. & Weiss, A.D. (1999). GLM versus CCA spatial modeling of plant distribution. *Plant Ecology* 143: 107–122.
- Guisan, A., and Zimmermann, N.E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling* 135: 147-186.
- Guisan, A., Edwards, T.C., Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling* 157: 89-100.
- Guisan, A., and Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology letters* 8(9): 993-1009.
- Gurevitch, J., Curtis, P.S., Jones, M.H. (2001). Meta-analysis in ecology. *Advances in ecological research* 32: 199-247.
- Halpern, B.S., Walbridge, S., Selkoe, K.A., Kappel, C.V., Micheli, F., D'agrosa, C et al. (2008). A global map of human impact on marine ecosystems. *Science* 319(5865): 948-952.
- Halpin, P.N., Read, A.J., Best, B.D., Hyrenbach, K.D., Fujioka, E., Coyne, M.S., Crowder, L.B., Freeman, S.A., Spoerri, C. (2006). OBIS-SEAMAP: developing a biogeographic research data commons for the ecological studies of marine mammals, seabirds, and sea turtles. *Marine Ecology Progress Series* 316: 239-246.
- Halpin, P.N., Read, A.J., Fujioka, E., Best, B.D., Donnelly, B., Hazen, L.J. et al. (2009). OBIS-SEAMAP: The world data center for marine mammal, sea bird, and sea turtle distributions. *Oceanography* 22(2): 104-115.
- Hamazaki, T. (2002). Spatiotemporal prediction models of cetacean habitats in the mid-western north Atlantic ocean (from Cape Hatteras, North Carolina, U.S.A. to Nova Scotia, Canada). *Marine Mammal Science* 18: 920-939.
- Hammond, P.S., Macleod, K., Berggren, P., Borchers, D.L., Burt, L., Cañadas, A. et al. (2013). Cetacean abundance and distribution in European Atlantic shelf waters to inform conservation and management. *Biological Conservation* 164: 107-122.

- Harris, P.T., Macmillan-Lawler, M., Rupp, J., Baker, E.K. (2014). Geomorphology of the oceans. *Marine Geology* 352: 4-24.
- Harvey, D., Leybourne, S., Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting* 13(2): 281-291.
- Hastie, T., and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science* 3: 297-313.
- Hastie, T., Tibshirani, R., Friedman, J.H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York.
- Hawkins, D.M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences* 44(1): 1-12.
- Hedley, S.L., and Buckland, S.T. (2004). Spatial models for line transect sampling. *Journal of Agricultural, Biological, and Environmental Statistics* 9(2): 181-199.
- Hegel, T.M., Cushman, S.A., Evans, J., Huettmann, F. (2010). Current state of the art for statistical modelling of species distributions. In *Spatial complexity, informatics, and wildlife conservation*. Springer Japan. pp. 273-311.
- Heikkinen, R.K., Marmion, M., Luoto, M. (2012). Does the interpolation accuracy of species distribution models come at the expense of transferability? *Ecography* 35(3): 276-288.
- Higgins, J., Thompson, S.G., Spiegelhalter, D.J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 172:137-159.
- Hirzel, A.H., Hausser, J., Chessel, D., Perrin, N. (2002). Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? *Ecology* 83: 2027-2036.
- Hohn, A.A., Rotstein, D.S., Harms, C.A. Southall, B.L. (2006). Multispecies mass stranding of pilot whales (*Globicephala macrorhynchus*), minke whale (*Balaenoptera acutorostrata*), and dwarf sperm whales (*Kogia sima*) in North Carolina on 15-16 January 2005. NOAA Technical Memorandum NMFS SEFSC 537: 222.
- Hooker, S.B., Rees, N.W., Aiken, J. (2000). An objective methodology for identifying oceanic provinces. *Progress in Oceanography* 45: 313-338.
- Hucke-Gaete, R., Moreno, C.A. Arata, J. (2004). Operational interactions of sperm whales and killer whales with the Patagonian toothfish industrial fishery off southern Chile. *CCAMLR Science* 11: 127-140.
- Hunt, G., and Schneider, D. (1987). Scale-dependent processes in the physical and biological environment of marine birds. *Seabirds: Feeding Ecology and Role in Marine Ecosystems*. Cambridge University Press. Croxall JP, pp. 7-42.
- Hutchinson, G.E. (1957). Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology* 22: 415-427.
- IUCN. (2001). *IUCN Red list categories and criteria: Version 3.1* IUCN Species Survival Commission. Gland, Switzerland, and Cambridge, UK: IUCN.
- Jacobsen, J.K., Massey, L., Gulland, F. (2010). Fatal ingestion of floating net debris by two sperm whales (*Physeter macrocephalus*). *Marine Pollution Bulletin* 60(5): 765-767.
- Jaquet, N. (1996). How spatial and temporal scales influence understanding of sperm whale distribution: a review. *Mammal Review* 26(1): 51-65.
- Jaquet, N., and Whitehead, H. (1996). Scale-dependent correlation of sperm whale distribution with environmental features and productivity in the South Pacific. *Marine ecology progress series*: 1-9.

- Jaynes, E.T. (1957). Information theory and statistical mechanics. *Physical review* 106: 620-630.
- Jennings, M.D. (2000). Gap analysis : concepts, methods, and recent results. *Landscape Ecology* 15: 5-20.
- Jensen, A.S., and G.K., Silber. (2003). Large Whale Ship Strike Database. U.S. Department of Commerce, NOAA Technical Memorandum. NMFS-OPR- , 37 pp.
- Johnson, C.N. (1998). Species extinction and the relationship between distribution and abundance. *Nature* 394:272-74.
- Jongman, R.H., Ter Braak, C.J., Van Tongeren, O.F. (1995). *Data analysis in community and landscape ecology volume 2*. Cambridge University Press, Cambridge.
- Kaschner, K., Watson, R., Trites, A.W., Pauly, D. (2006). Mapping world-wide distributions of marine mammal species using a relative environmental suitability (RES) model. *Marine Ecology Progress Series* 316: 285-310.
- Kawakami, T. (1980). A review of sperm whale food. *Scientific Reports of the Whales Research Institute* 32:199-218.
- King, G., and Zeng, L. (2007). When Can History Be Our Guide? The pitfalls of counterfactual inference. *International Studies Quarterly* 51 (1): 183-210.
- Kinney, S. K., and Dunson, D. B. (2008). Bayesian model uncertainty in mixed effects models. In *Random effect and latent variable model selection* (pp. 37-62). Springer New York.
- Kiszka, J., Macleod, K., Van Canneyt, O., Walker, D., Ridoux, V. (2007). Distribution, encounter rates, and habitat characteristics of toothed cetaceans in the Bay of Biscay and adjacent waters from platform-of-opportunity Data. *ICES Journal of Marine Science* 64: 1033-1043.
- Kleiber, C., and Zeileis, A. (2016). Visualizing count data regressions using rootograms. *The American Statistician* 70(3): 296-303.
- Laist, D.W., Coe, J.M., O'Hara, K.J. (1999). Marine debris pollution. In: J.R. Twiss and R.R. Reeves (eds). *Conservation and Management of Marine Mammals*.
- Lambert, D. (1992). Zero-Inflated Poisson Regression, With an Application To Defects in Manufacturing. *Technometrics* 34(1): 1-14.
- Lambert, C., Mannocci, L., Lehodey, P., Ridoux, V. (2014). Predicting cetacean habitats from their energetic needs and the distribution of their prey in two contrasted tropical regions. *PLoS One* 9(8): e105958.
- Lambert, C., Pettex, E., Dorémus, G., Laran, S., Stéphan, E., Van Canneyt, O., Ridoux, V. (2017a). How does ocean seasonality drive habitat preferences of highly mobile top predators? Part II: The eastern North-Atlantic. *Deep Sea Research Part II: Topical Studies in Oceanography* 141: 133-154.
- Lambert, C., Laran, S., David, L., Dorémus, G., Pettex, E., Van Canneyt, O., Ridoux, V. (2017b). How does ocean seasonality drive habitat preferences of highly mobile top predators? Part I: The north-western Mediterranean Sea. *Deep Sea Research Part II: Topical Studies in Oceanography* 141: 115-132.
- Lande, R. (1995). Mutation and conservation. *Conservation Biology* 9:782-91.
- Lanfredi, C., Azzellino, A., D'Amico, A., Centurioni, L., Ampolo Rella, M., Pavan, G., Podestà, M. (2016). Key Oceanographic Characteristics of Cuvier's Beaked Whale (*Ziphius cavirostris*) Habitat in the Gulf of Genoa (Ligurian Sea, NW Mediterranean). *Journal of Oceanography and Marine Research* 4: 145.

- Laran, S., Pettex, E., Authier, M., Blanck, A., David, L., Dorémus, G., et al. (2017a). Seasonal distribution and abundance of cetaceans within French waters. Part I: The North-Western Mediterranean, including the Pelagos sanctuary. *Deep Sea Research Part II* 141: 20-30.
- Laran, S., Authier, M., Van Canneyt, O., Dorémus, G., Watremez, P., Ridoux, V. (2017b). A comprehensive survey of pelagic megafauna: their distribution, densities and taxonomic richness in the tropical Southwest Indian Ocean. *Frontiers in Marine Science* 4: 139.
- Latimer, A.M., Wu, S., Gelfand, A.E., Silander, J. A. (2006). Building statistical models to analyze species distributions. *Ecological applications* 16(1): 33-50.
- Lawler, J.J., White, D., Sifneos, J.C., Master, L.L. (2003). Rare Species and the Use of Indicator Groups for Conservation Planning. *Conservation Biology* 17(3): 875-882.
- Leatherwood, S., and Reeves, R.R. (1983). The sperm, pygmy sperm, and dwarf sperm whales. In: Leatherwood, S., Reeves, R.R. (eds) *The Sierra Club Handbook of Whales and Dolphins*. Sierra Club Books, San Francisco, p. 302.
- Leathwick, J.R., Elith, J., Hastie, T. (2006). Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological modelling* 199(2): 188-196.
- Lehodey, P., Senina, I., Murtugudde, R. (2008). A spatial ecosystem and populations dynamics model (SEAPODYM) – Modeling of tuna and tuna-like populations. *Progress in Oceanography* 78(4): 304-318.
- 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): 69-84.
- Lehodey, P., Conchon, A., Senina, I., Domokos, R., Calmettes, B., Jouanno, J., Hernandez, O., Kloser, R. (2014). Optimization of a micronekton model with acoustic data. *ICES Journal of Marine Science* 72(5): 1399-1412.
- Lek, S., and Guégan, J.F. (1999). Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling* 120: 65-73.
- Levin, S.A. (1992). The problem of pattern and scale in ecology. *Ecology* 73: 1943-1967.
- Levin, L.A., and Gooday, A. (2003). The Deep Atlantic Ocean. *Ecosystems of the deep oceans*. (Tyler PA, Ed.), pp. 111-178. Amsterdam; New York: Elsevier.
- Lindén, A., and Mantyniemi, S. (2011). Using the negative binomial distribution to model overdispersion in ecological data. *Ecology* 92(7): 1414-1421.
- Liu, Z., Peng, C., Xiang, W., Tian, D., Deng, X., Zhao, M. (2010). Application of artificial neural networks in global climate change and ecological research: An overview. *Chinese science bulletin* 55(34): 3853-3863.
- Longhurst, A. (2007). *Ecological geography of the sea*. Academic Press, Oxford.
- Louzao, M., Hyrenbach, K.D., Arcos, J.M., Abelló, P., Sola, L.G., Oro, D. (2006). Oceanographic habitat of an endangered Mediterranean procellariiform: implications for marine protected areas. *Ecological Applications* 16(5): 1683-1695.
- Louzao, M., Bécares, J., Rodríguez, B., Hyrenbach, K.D., Ruiz, A., Arcos, J.M. (2009). Combining vessel-based surveys and tracking data to identify key marine areas for seabirds. *Marine Ecology Progress Series* 391: 183-198.
- Lumaret, J.P., and Jay-Robert, P. (2002). Modelling the species richness distribution of French dung beetles (*Coleoptera, Scarabaeidae*) and delimiting the predictive capacity of different groups of explanatory variables. *Global Ecology and Biogeography* 11(4): 265-277.

- Lyons, K.G., and Schwartz, M.W. (2001). Rare species loss alters ecosystem function—invasion resistance. *Ecology Letters* 4:1-8.
- Lyons, K.G., Brigham, C.A., Traut, B.H., Schwartz, M.W. (2005). Rare species and ecosystem functioning. *Conservation Biology* 19(4): 1019-1024.
- Mackenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, J.A., Langtimm, C.A. (2002). Estimating Site Occupancy Rates When Detection Probabilities Are Less Than One. *Ecology* 83(8): 2248-2255.
- MacLeod, C.D., Santos, M.B., Pierce, G.J. (2003). Review of data on diets of beaked whales: evidence of niche separation and geographic segregation. *Journal of the Marine Biological Association of the United Kingdom* 83(3): 651-665.
- MacLeod, C.D., Perrin, W.F., Pitman, R., Barlow, J., Ballance, L., D'Amico, A. et al. (2005). Known and inferred distributions of beaked whale species (Cetacea: *Ziphiidae*). *Journal of Cetacean Research and Management* 7(3): 271.
- MacLeod, C.D., and D'Amico, A. (2006). A review of beaked whale behaviour and ecology in relation to assessing and mitigating impacts of anthropogenic noise. *Journal of Cetacean Research and Management* 7(3): 211-221.
- MacLeod, C.D., Weir, C.R., Pierpoint, C., Harland, E.J. (2007). The habitat preferences of marine mammals west of Scotland (UK). *Journal of the Marine Biological Association of the United Kingdom* 87: 157-164.
- MacLeod, C.D., Weir, C.R., Santos, M.B., Dunn, T.E. (2008). Temperature-based summer habitat partitioning between white-beaked and common dolphins around the United Kingdom and Republic of Ireland. *Journal of the Marine Biological Association of the United Kingdom* 88: 1193-1198.
- MacLeod, C.D., Brereton, T., Martin, C. (2009). Changes in the occurrence of common dolphins, striped dolphins and harbour porpoises in the English Channel and Bay of Biscay. *Journal of the Marine Biological Association of the United Kingdom* 89(05): 1059.
- MacLeod, K., Brereton, T., Evans, P.G., Swift, R., Vázquez, J.A. (2011). Distribution and abundance of Cuvier's beaked whales in the canyons of southern Biscay. SC/63/SM7). In: 63st Annual Meeting of the International Whaling Commission, 1–13 June 2011, Tromsø, Norway.
- Mac Nally, R. (2002). Multiple regression and inference in ecology and conservation biology: further comments on identifying important predictor variables. *Biodiversity and Conservation* 11: 1397-1401.
- Madsen, P.T., Mohl, B., Nielsen, B.K. Wahlberg, M. (2002). Male sperm whale behaviour during exposures to distant seismic survey pulses. *Aquatic Mammals* 28(3): 231-240.
- Madsen, P.T., de Soto, N.A., Tyack, P.L., Johnson, M. (2014). Beaked whales. *Current Biology* 24(16): 728-730.
- Mannocci, L., Catalogna, M., Dorémus, G., Laran, S., Lehodey, P., Massart, W., Monestiez, P. et al. (2014a). Predicting cetacean and seabird habitats across a productivity gradient in the South Pacific gyre. *Progress in Oceanography* 120: 383-398.
- Mannocci, L., Laran, S., Monestiez, P., Dorémus, G., Van Canneyt, O., Watremez, P., Ridoux, V. (2014b). Predicting top predator habitats in the Southwest Indian Ocean. *Ecography* 37(3): 261-278.
- Mannocci, L., Monestiez, P., Spitz, J., Ridoux, V. (2015). Extrapolating cetacean densities beyond surveyed regions: habitat-based predictions in the circumtropical belt. *Journal of Biogeography* 42: 1267-1280.

- Mannocci, L., Boustany, A.M., Roberts, J.J., et al. (2017a). Temporal resolutions in species distribution models of highly mobile marine animals: Recommendations for ecologists and managers. *Diversity and Distributions* 23: 1098-1109.
- Mannocci, L., Roberts, J.J., Miller, D.L., Halpin, P.N. (2017b). Extrapolating cetacean densities to quantitatively assess human impacts on populations in the high seas. *Conservation Biology* 31: 601-614.
- Marmion, M., Luoto, M., Heikkinen, R. K., Thuiller, W. (2009). The performance of state-of-the-art modelling techniques depends on geographical distribution of species. *Ecological Modelling* 220(24): 3512-3520.
- Marques, F.F., and Buckland, S.T. (2003). Incorporating covariates into standard line transect analyses. *Biometrics* 59(4): 924-935.
- Marsh, H., and Sinclair, D.F. (1989). Correcting for visibility bias in strip transect aerial surveys of aquatic fauna. *The Journal of Wildlife Management*: 1017-1024.
- Martínez-Rincón, R.O., Ortega-García, S., Vaca-Rodríguez, J.G. (2012). Comparative performance of generalized additive models and boosted regression trees for statistical modeling of incidental catch of wahoo (*Acanthocybium solandri*) in the Mexican tuna purse-seine fishery. *Ecological Modelling* 233: 20-25.
- Master, L.L., Stein, B.A., Kutner, L.S., Hammerson, G.A. (2000). Vanishing assets: Conservation status of U.S. species. pp. 93-118 in *Precious heritage: The status of biodiversity in the United States*, ed. B. A Stein, L. S. Kutner, and J. S. Adams. New York: Oxford University Press.
- Matthies, D., Brauer, I., Maibom, W., Tschardtke, T. (2004). Population size and the risk of local extinction: Empirical evidence from rare plants. *Oikos* 105:481-88.
- Maxwell, S.M., Hazen, E.L., Lewison, R.L., Dunn, D.C., Bailey, H., Bograd, S.J. et al. (2015). Dynamic ocean management: Defining and conceptualizing real-time management of the ocean. *Marine Policy* 58: 42-50.
- McAlpine, D.F. (2002). Pygmy and dwarf sperm whales *Kogia breviceps* and *K. simus*. In: W.F. Perrin, B. Wursig and J.G.M. Thewissen (eds), *Encyclopedia of Marine Mammals*, pp. 1007-1009. Academic Press.
- McAlpine D.F. (2009). Pygmy and dwarf sperm whales. *Encyclopedia of marine mammals 2nd Edition*. pp. 936-938. Academic Press.
- McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models*, no. 37 in *Monograph on Statistics and Applied Probability*.
- McKinney, M.L. (1997). Extinction vulnerability and selectivity: Combining ecological and paleontological views. *Annual Review of Ecology and Systematics* 28:495-516.
- McSweeney, D.J., Baird, R.W., Mahaffy, S.D. (2007). Site fidelity, associations, and movements of Cuvier's (*Ziphius cavirostris*) and Blainville's (*Mesoplodon densirostris*) beaked whales off the island of Hawai'i. *Marine Mammal Science* 23: 666-687.
- Mead, J.G. (1984). Survey of reproductive data for the beaked whales (*Ziphiidae*). Report of the International Whaling Commission 6.
- Mead, J.G. (2009). Beaked whales, an overview. *Encyclopedia of marine mammals 2nd Edition*, pp. 94-97. Academic Press.
- Menges, E.S., and Kimmich, J. (1996). Microhabitat and time-since-fire: effects on demography of *Eryngium cuneifolium* (*Apiaceae*), a Florida scrub endemic plant. *American Journal of Botany* 83:185-191.
- Merow, C., Smith, M.J., Silander, J.A. (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography* 36(10): 1058-1069.

- Merrett, N.R. (1994). Reproduction in the North Atlantic oceanic ichthyofauna and the relationship between fecundity and species' sizes. In *Women in ichthyology: an anthology in honour of ET, Ro and Genie*, pp. 207-245. Springer Netherlands.
- Mesgaran, M.B., Cousens, R.D., Webber, B.L. (2014). Here be dragons: a tool for quantifying novelty due to covariate range and correlation change when projecting species distribution models. *Diversity and Distributions* 20(10): 1147-1159.
- Minami, M., Lennert-Cody, C.E., Gao, W., Roman-Verdesoto, M. (2007). Modeling shark bycatch: the zero-inflated negative binomial regression model with smoothing. *Fisheries Research* 84(2): 210-221.
- Mitchell, P.I., Newton, S.F., Ratcliffe, N. Dunn, T.E. (2004). *Seabird Populations of Britain and Ireland*. Poyser, London.
- Monk, J. Ierodionou, D., Versace, V.L., Bellgrove, A., Harvey, E., Rattray, A., Laurenson, L., Quinn, G.P. (2010). Habitat suitability for marine fishes using presence-only modelling and multibeam sonar. *Marine Ecology Progress Series* 420: 157-174.
- Monsarrat, S., Pennino, M.G., Smith, T.D., Reeves, R.R., Meynard, C.N., Kaplan, D.M., Rodrigues, A.S. (2015). Historical summer distribution of the endangered North Atlantic right whale (*Eubalaena glacialis*): a hypothesis based on environmental preferences of a congeneric species. *Diversity and Distributions* 21: 925-937.
- Monserud, R.A., and Leemans, R. (1992). Comparing global vegetation maps with the Kappa statistic. *Ecological modelling* 62(4): 275-293.
- Morin, X., and Thuiller, W. (2009). Comparing niche-and process-based models to reduce prediction uncertainty in species range shifts under climate change. *Ecology* 90(5): 1301-1313.
- Moura, A.E., Sillero, N., Rodrigues, A. (2012). Common dolphin (*Delphinus delphis*) habitat preferences using data from two platforms of opportunity. *Acta oecologica* 38: 24-32.
- Murphy, S., Pinn, E.H., Jepson, P.D. (2013). Review of New Information on Other Matters Relevant for Small Cetacean Conservation Population Size, Distribution, Structure and Causes of Any Changes Marine megavertebrates adrift: a framework for the interpretation of stranding data in a monitoring p. *Oceanography and Marine Biology* 51: 193-280.
- Nagelkerke, N.J. (1991). A note on a general definition of the coefficient of determination. *Biometrika* 78(3): 691-692.
- Naugle, D.E., Higgins, K.F., Nusser, S.M., Johnson, W.C. (1999). Scale-dependent habitat use in three species of prairie wetland birds. *Landscape ecology* 14(3): 267-276.
- Nielsen, J.B., Nielsen, F., Joergensen, P.J., Grandjean, P. (2000). Toxic metals and selenium in blood from pilot whales (*Globicephala melas*) and sperm whales (*Physeter catodon*). *Marine Pollution Bulletin* 40: 348-351.
- Oppel, S., Meirinho, A., Ramirez, I., Gardner, B., O'Connell, A. F., Miller, P.I., Louzao, M. (2012). Comparison of five modelling techniques to predict the spatial distribution and abundance of seabirds. *Biological Conservation* 156: 94-104.
- Oschlies, A., and Garçon, V. (1998). Eddy-induced enhancement of primary production in a model of the North Atlantic Ocean. *Nature* 394: 266-269.
- OSPAR Commission. (2000). *Quality Status Report 2000, Region V – Wider Atlantic*. OSPAR Commission, London.
- Österblom, H., Olsson, O., Blenckner, T., Furness, R.W. (2008). Junk-food in marine ecosystems. *Oikos* 117(7): 967-977.

- Panigada, S., Zanardelli, M., MacKenzie, M., Donovan, C., Mélin, F., Hammond, P.S. (2008). Modelling habitat preferences for fin whales and striped dolphins in the Pelagos Sanctuary (Western Mediterranean Sea) with physiographic and remote sensing variables. *Remote Sensing of Environment* 112(8): 3400-3412.
- Pärtel, M., Kalamees, R., Reier, D., Tuvi, E.L., Roosalu, E., Vellak, A., Zobel, M. (2005). Grouping and prioritization of vascular plant species for conservation: Combining natural rarity and management need. *Biological Conservation* 123: 271-78.
- Paxton, C.G.M., Scott-Hayward, L., Mackenzie, M., Rexstad, E., Thomas, L. (2016). Revised Phase III Data Analysis of Joint Cetacean Protocol Data Resources with Advisory Note, JNCC Report 517, ISSN 0963-8091
- Pearce, J., and Ferrier, S. (2000). Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling* 133(3): 225-245.
- Pearson, R.G., and Dawson, T.P. (2003). Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global ecology and biogeography* 12(5): 361-371.
- Pérez, C., López, A., Sequeira, M., Silva, M., Herrera, R., Gonçalves, J., Valdes, P., Mons, L. et al. (1997). Stranding and by-catch of cetaceans in the Northeastern Atlantic during 1996.
- Pérez, V., Fernández, E., Marañón, E., Serret, P., Varela, R., Bode, A., et al. (2005). Latitudinal distribution of microbial plankton abundance, production, and respiration in the Equatorial Atlantic in autumn 2000. *Deep Sea Research Part I: Oceanographic Research Papers* 52(5): 861-880.
- Perrin, W.F., Würsig, B., Thewissen, J. G. M. (2009). *Encyclopedia of marine mammals*. Academic Press.
- Phillips, S.J., Dudík, M., Schapire, R.E. (2004). A Maximum Entropy Approach to Species Distribution Modeling. Twenty-first international conference on Machine learning - ICML '04, p. 83.
- 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., and Dudík, M. (2008). Modeling of species distribution with Maxent: new extensions and a comprehensive evaluation. *Ecography* 31: 161-175.
- Pielou, E.C. (1984). *The Interpretation of Ecological Data: A Primer on Classification and Ordination*. Wiley, New York.
- Pinardi, N., and Masetti, E. (2000). Variability of the large scale general circulation of the Mediterranean Sea from observations and modelling: A review. *Palaeogeography, Palaeoclimatology, Palaeoecology* 158: 153-174.
- Pinaud, D., and Weimerskirch, H. (2005). Scale-dependent habitat use in a long-ranging central place predator. *Journal of Animal Ecology* 74(5): 852-863.
- 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.
- Plummer, M. (2016). rjags: Bayesian Graphical Models using MCMC. R package version 4-6. <https://CRAN.R-project.org/package=rjags>
- Praca, E., and Gannier, A. (2007). Ecological niche of three teuthophageous odontocetes in the northwestern Mediterranean Sea. *Ocean Science Discussions* 4: 49-59.



- 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.
- Price, M.C., Macdonald, G.J., Graham, S.L. Kirk, E.J. (1984). Treatment of a stranding whale (*Kogia breviceps*). *New Zealand Veterinary Journal* 32: 31-33.
- Pujo-Pay, M., Conan, P., Oriol, L., Cornet-Barthaux, V., Falco, C., Ghiglione, J. F. et al. (2011). Integrated survey of elemental stoichiometry (C, N, P) from the western to eastern Mediterranean Sea. *Biogeosciences* 8(4): 883-899.
- Purvis, A., Agapow, P.M., Gittleman, J.C., Mace, G.M. (2000). Nonrandom extinction and the loss of evolutionary history. *Science* 288: 328-330.
- R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Rabinowitz, D. (1981). Seven forms of rarity. *The biological aspects of rare plant conservation*, ed. H. Synge. New York: Wiley. pp. 205-217.
- Randin, C.F., Dirnböck, T., Dullinger, S., Zimmermann, N.E., Zappa, M., Guisan, A. (2006). Are niche-based species distribution models transferable in space? *Journal of biogeography* 33(10): 1689-1703.
- Read, A.J., and Wade, P.R. (2000). Status of marine mammals in the United States. *Conservation Biology* 14(4): 929-940.
- Redfern, J.V., Ferguson, M.C., Becker, E.A., Hyrenbach, K.D., Good, C. et al. (2006). Techniques for cetacean – habitat modeling. *Marine Ecology Progress Series* 310: 271-295.
- Redfern, J.V., Moore, T.J., Fiedler, P.C., de Vos, A., Brownell, R.L., Forney, K.A., Becker, E.A., Ballance, L.T. (2017). Predicting cetacean distributions in data-poor marine ecosystems. *Diversity and Distributions* 23: 394-408.
- Reeves, R.R., and Notarbartolo Di Sciara, G. (2006). The status and distribution of cetaceans in the Black Sea and Mediterranean Sea. IUCN Centre for Mediterranean Cooperation, Malaga, Spain.
- Reveal J.L. (1981). The concept of rarity and population threats in plant communities. *Rare plant conservation*, ed. L. E. Morse and M. S. Henefin. Bronx: New York Botanical Garden. pp. 41-46.
- Reygondeau, G., Maury, O., Beaugrand, G., Fromentin, J.M., Fonteneau, A., Cury, P. (2012). Biogeography of tuna and billfish communities. *Journal of Biogeography* 39: 114-129.
- Reynoldson, T.B., Norris, R.H., Resh, V.H., Day, K.E., Rosenberg, D.M. (1997). The reference condition: A comparison of multimetric and multivariate approaches to assess water-quality impairment using benthic macroinvertebrates. *Journal of the North American Benthological Society* 16: 833-52.
- Rice, D.W. (1989). Sperm whales (*Physeter macrocephalus*). In: Ridgway SH, Harrison R (eds) *Handbook of marine mammals*, Vol 4. Academic Press, London, pp. 177-233.
- Ridout, M., Demetrio, C.G., Hinde, J. (1998). Models for count data with many zeros. *International Biometric Conference*, pp.1-13.
- Roberts, J.J., Best, B.D., Dunn, D.C., Tremblay, E.A., Halpin, P.N. (2010). Marine Geospatial Ecology Tools: An integrated framework for ecological geoprocessing with ArcGIS, Python, R, MATLAB, and C++. *Environmental Modelling & Software* 25: 1197-1207.
- Roberts, J.J., Best, B.D., Mannocci, L., Fujioka, E., Halpin, P.N., Palka, D.L., Garrison, L.P., Mullin, K.D., Cole, T.V.N., Khan, C.B. et al. (2016). Habitat-based cetacean density models for the U.S. Atlantic and Gulf of Mexico. *Scientific Report* 6.

- Rogan, E., Cañadas, A., Macleod, K., Santos, M.B., Mikkelsen, B., Uriarte, A., Van Canneyt, O. et al. (2017). Distribution abundance and habitat use of deep diving cetaceans in the North-East Atlantic. *Deep Sea Research Part II: Topical Studies in Oceanography* 141: 8-19.
- Rosa, R., Dierssen, H.M., Gonzalez, L., Seibel, B.A. (2008). Large-scale diversity patterns of cephalopods in the Atlantic open ocean and deep sea. *Ecology* 89(12): 3449-3461.
- Rougerie, F. and Rancher, J. (1994). The Polynesian south ocean: features and circulation. *Marine Pollution Bulletin* 29: 14-25.
- Royle, J.A., and Dorazio, R.M. (2008). Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities. Academic Press.
- Rushton, S.P., Ormerod, S.J., Kerby, G. (2004). New paradigms for modelling species distributions? *Journal of applied ecology* 41(2): 193-200.
- Salah, M.M., and Boxer, B. (2009). Mediterranean Sea. *Encyclopedia Britannica*.
- Saltzman, J., and Wishner, K.F. (1997). Zooplankton ecology in the eastern tropical Pacific oxygen minimum zone above a seamount: 1. General trends. *Deep Sea Research Part I: Oceanographic Research Papers* 44(6): 907-930.
- Santos, M., Pierce, G., Boyle, P. (1999). Stomach contents of sperm whales *Physeter macrocephalus* stranded in the North Sea 1990-1996. *Marine Ecology Progress Series* 183:281-294.
- Santos, M., and Pierce, G. (2002). Additional notes on stomach contents of sperm whales *Physeter macrocephalus* stranded in the north-east Atlantic. *Journal of the Marine Biological Association of the United Kingdom*: 501-507.
- Santos, M.B., Pierce, G.J., Lopez, A., Reid, R.J., Ridoux, V., Mente, E. (2006). Pygmy sperm whales *Kogia breviceps* in the Northeast Atlantic: New information on stomach contents and strandings. *Marine Mammal Science* 22(3): 600-616.
- Sardā, F., Calafat, A., Flexas, M.M., Tselepidis, A., Canals, M., Espino, M., Tursi, A. (2004). An introduction to Mediterranean deep-sea biology. *Scientia Marina* 68(S3): 7-38.
- Schmitz, W.J., and McCartney, M.S. (1993). On the north Atlantic circulation. *Reviews of Geophysics* 31(1): 29-49.
- Schneider, D.C. (1994). *Quantitative Ecology: Spatial and Temporal Scaling*. San Diego: Academic Press.
- Schouten, M.W., De Ruijter W.P.M., Van Leeuwen, P.J., Ridderinkhof, H. (2003). Eddies and variability in the Mozambique Channel. *Deep Sea Research Part II: Topical Studies in Oceanography* 50: 1987-2003.
- Scott, J.M., Heglund, P.J., Haufler, J.B., Morrison, M., Raphael, M.G., Wall, W.B. et al. (2002). *Predicting Species Occurrences: Issues of Accuracy and Scale*. Island Press, Covelo, CA.
- Shirihai, H. (2006). *Mammifères marins du monde*. Delachaux and Niestlé, Paris.
- Shirihai, H., and Jarrett, B. (2006). *Whales Dolphins and Other Marine Mammals of the World*. Princeton: Princeton Univ. Press. pp. 155-158.
- Shmueli, G. (2010). To explain or to predict? *Statistical science* 25: 289–310.
- Shono, H. (2008). Application of the Tweedie distribution to zero-catch data in CPUE analysis. *Fisheries Research* 93(1): 154-162.
- Spitz, J., Cherel, Y., Bertin, S., Kiszka, J., Dewez, A., Ridoux, V. (2011). Prey preferences among the community of deep-diving odontocetes from the Bay of Biscay, Northeast Atlantic. *Deep Sea Research Part I: Oceanographic Research Papers* 58(3): 273-282.

- Staudinger, M.D., McAlarney, R.J., McLellan, W.A., Ann Pabst, D. (2014). Foraging ecology and niche overlap in pygmy (*Kogia breviceps*) and dwarf (*Kogia sima*) sperm whales from waters of the US mid-Atlantic coast. *Marine Mammal Science* 30(2): 626-655.
- Stockwell, D., and Peters, D. (1999). The GARP modelling system: problems and solutions to automated spatial prediction. *International journal of geographical information science* 13(2): 143-158.
- Stockwell, D.R., and Peterson, A.T. (2002). Effects of sample size on accuracy of species distribution models. *Ecological modelling* 148(1): 1-13.
- Stoll, H., King, G., Zeng, L. (2014). Whatif: Software for Evaluating Counterfactuals. R package version 1.5-6. <https://cran.r-project.org/web/packages/WhatIf/index.html>.
- Stone, C.J., and Tasker, M.L. (2006). The effects of seismic airguns on cetaceans in UK waters. *Journal of Cetacean Research and Management* 8: 255-263.
- Subramanian, J., and Simon, R. (2013). Overfitting in prediction models - Is it a problem only in high dimensions? *Contemporary Clinical Trials* 36(2): 636-641.
- Syphard, A.D., and Franklin, J., (2009). Differences in spatial predictions among species distribution modeling methods vary with species traits and environmental predictors. *Ecography* 32(6): 907-918.
- Talluto, M. V., Boulangeat, I., Ameztegui, A., Aubin, I., Berteaux, D., Butler, A., et al. (2016). Cross-scale integration of knowledge for predicting species ranges: a metamodeling framework. *Global ecology and biogeography* 25(2): 238-249.
- Tanhua, T., Hainbucher, D., Schroeder, K., Cardin, V., Álvarez, M., Civitarese, G. (2013). The Mediterranean Sea system: a review and an introduction to the special issue. *Ocean Science* 9(5).
- Taylor, B.L., Baird, R., Barlow, J., Dawson, S.M., Ford, J., Mead, J.G., Notarbartolo di Sciara, G. et al. (2008a). *Ziphius cavirostris*. The IUCN Red List of Threatened Species 2008: e.T23211A9429826.
- Taylor, B.L., Baird, R., Barlow, J., Dawson, S.M., Ford, J., Mead, J.G., Notarbartolo di Sciara, G. et al. (2008b). *Physeter macrocephalus*. The IUCN Red List of Threatened Species 2008: e.T41755A10554884.
- Taylor, B.L., Baird, R., Barlow, J., Dawson, S.M., Ford, J.K.B., Mead, J.G., Notarbartolo di Sciara, G. et al. (2012a). *Kogia breviceps*. The IUCN Red List of Threatened Species 2012: e.T11047A17692192.
- Taylor, B.L., Baird, R., Barlow, J., Dawson, S.M., Ford, J.K.B., Mead, J.G., Notarbartolo di Sciara, G. et al. (2012b). *Kogia sima*. The IUCN Red List of Threatened Species 2012: e.T11048A17695273.
- Theodose, T.A., Jaeger, C.H., Bowman, W.D., Schardt, J.C. (1996). Uptake and allocation of <sup>15</sup>N in alpine plants: implications for the importance of competitive ability in predicting community structure in a stressful environment. *Oikos* 75:59-66.
- Thiers, L., Louzao, M., Ridoux, V., Le Corre, M., Jaquemet, S., Weimerskirch, H. (2014). Combining methods to describe important marine habitats for top predators: application to identify biological hotspots in tropical waters. *PLoS one* 9(12): e115057.
- Thomas, L., Buckland, S.T., Rexstad, E.A., Laake, J.L., Strindberg, S., Hedley, S.L., Bishop, J.R.B. et al. (2010). Distance software: Design and analysis of distance sampling surveys for estimating population size. *Journal of Applied Ecology* 47: 5-14.
- Thorne, L.H., Johnston, D.W., Urban, D.L., Tyne, J., Bejder, L., Baird, R. W. et al. (2012). Predictive modeling of spinner dolphin (*Stenella longirostris*) resting habitat in the main Hawaiian Islands. *PLoS One* 7(8): e43167.

- Thuiller, W., Araújo, M. B., Lavorel, S. (2003). Generalized models vs. classification tree analysis: predicting spatial distributions of plant species at different scales. *Journal of Vegetation Science* 14(5): 669-680.
- Tomczak, M., and Godfrey, J.S. (2003). *Regional oceanography: an introduction*. Elsevier.
- Torres, L.G., Sutton, P.J., Thompson, D.R., Delord, K., Weimerskirch, H., Sagar, P.M. et al. (2015). Poor transferability of species distribution models for a pelagic predator, the grey petrel, indicates contrasting habitat preferences across ocean basins. *PLoS One* 10(3): e0120014.
- Tsoar, A., Allouche, O., Steinitz, O., Rotem, D., Kadmon, R. (2007). A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions* 13(4): 397-405.
- Unger, B., Rebolledo, E.L.B., Deaville, R., Gröne, A., IJsseldijk, L.L., Leopold, M.F. et al. (2016). Large amounts of marine debris found in sperm whales stranded along the North Sea coast in early 2016. *Marine pollution bulletin* 112(1): 134-141.
- Vanreusel, W., Maes, D., Van Dyck, H. (2007). Transferability of species distribution models: a functional habitat approach for two regionally threatened butterflies. *Conservation biology* 21(1): 201-212.
- Vilchis, L.I., Ballance, L.T., Fiedler, P.C. (2006). Pelagic habitat of seabirds in the eastern tropical Pacific: Effects of foraging ecology on habitat selection. *Marine Ecology Progress Series* 315: 279-292.
- Virgili, A., Authier, M., Boiseau, O., Cañadas, A., Claridge, D., Cole, T., Corkeron, P. et al. (in prep.). Combining visual surveys to model habitat of deep-diving cetaceans at the basin scale.
- Virgili A., Authier M., Monestiez P., Ridoux V. (in revision). How many sightings to model rare marine species distributions. *PLoS One*.
- Virgili, A., Racine, M., Authier, M., Monestiez, P., Ridoux, V. (2017a). Comparison of habitat models for scarcely detected species. *Ecological Modelling* 346: 88-98.
- Virgili, A., Lambert, C., Pettex, E., Dorémus, G., Van Canneyt, O., Ridoux, V. (2017b). Predicting seasonal variations in coastal seabird habitats in the English Channel and the Bay of Biscay. *Deep Sea Research Part II: Topical Studies in Oceanography* 141: 212-223.
- Walker, B., Kinzig, A., Langridge, J. (1999). Plant attribute diversity, resilience, and ecosystem function: the nature and significance of dominant and minor species. *Ecosystems* 2:95-113.
- Wallach, D., and Goffinet, B. (1989). Mean squared error of prediction as a criterion for evaluating and comparing system models. *Ecological Modelling* 44(3-4): 299-306.
- Wang, J.Y., and Yang, S.C. (2006). Unusual cetacean stranding events of Taiwan in 2004 and 2005. *Journal of Cetacean Research and Management* 8: 283-292.
- Warton, D.I. (2005). Many zeros does not mean zero inflation: Comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmetrics* 16(3): 275-289.
- Waring, G.T., Hamazaki, T., Sheehan, D., Wood, G., Baker, S. (2001). Characterization of beaked whale (*Ziphiidae*) and sperm whale (*Physeter macrocephalus*) summer habitat in shelf-edge and deeper waters off the Northeast U.S. *Marine Mammal Science* 17: 703-717.
- Watwood, S.L., Miller, P.J., Johnson, M., Madsen, P.T., Tyack, P.L. (2006). Deep-diving foraging behaviour of sperm whales (*Physeter macrocephalus*). *Journal of Animal Ecology* 75: 814-825.
- Weimerskirch, H. (2007). Are seabirds foraging for unpredictable resources? *Deep Sea Research Part II: Topical Studies in Oceanography* 54(3): 211-223.

- 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(1-3): 297-308.
- Wenger, S.J., and Freeman, M.C. (2014). Estimating Species Occurrence, Abundance, and Detection Probability Using Zero-inflated Distributions. *Ecology* 89(10): 2953-2959.
- Whitehead, H. (2002). Estimates of the current global population size and historical trajectory for sperm whales. *Marine Ecology Progress Series* 242: 295-304.
- Whitehead, H. (2003). Sperm whales: social evolution in the ocean. University of Chicago press, Chicago, US.
- Whitehead, H., MacLeod, C.D., Rodhouse, P. (2003). Differences in niche breadth among some teuthivorous mesopelagic marine mammals. *Marine Mammal Science* 19: 400-406.
- Whitehead, H. (2009). Sperm whales. *Encyclopedia of marine mammals* 2nd Edition, pp. 1091-1098. Academic Press.
- Whitehead, H. (2013). Trends in cetacean abundance in the Gully submarine canyon, 1988–2011, highlight a 21% per year increase in Sowerby's beaked whales (*Mesoplodon bidens*). *Canadian Journal of Zoology* 148: 141-148.
- Wilson, K.A., Westphal, M.I., Possingham, H.P., Elith, J. (2005). Sensitivity of conservation planning to different approaches to using predicted species distribution data. *Biological Conservation* 122(1): 99-112.
- Wimmer, T., and Whitehead, H. (2004). Movements and distribution of northern bottlenose whales, *Hyperoodon ampullatus*, on the Scotian Slope and in adjacent waters. *Canadian Journal of Zoology* 82: 1782-1794.
- Winiarski, K.J., Burt, M.L., Rexstad, E., Miller, D.L., Trocki, C.L., Paton, P.W.C., McWilliams, S.R. (2014). Integrating aerial and ship surveys of marine birds into a combined density surface model: a case study of wintering Common Loons. *Condor* 116: 149-161.
- Wisz, M.S., Hijmans, R. J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A. (2008). Effects of sample size on the performance of species distribution models. *Diversity and Distributions* 14(5): 763-773.
- Wong, S.N.P., and Whitehead, H. (2014). Seasonal occurrence of sperm whales (*Physeter macrocephalus*) around Kelvin Seamount in the Sargasso Sea in relation to oceanographic processes. *Deep Sea Research Part I: Oceanographic Research Papers* 91: 10-16.
- Wood, S.N. (2006a). On confidence intervals for generalized additive models based on penalized regression splines. *Australian and New Zealand Journal of Statistics* 48(4): 445-464.
- Wood, S. (2006b). *Generalized Additive models: An Introduction with R*. Chapman & Hall/CRC. 422 p.
- Wood, S. (2013). mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation. Retrieved 7 July 2014, from <http://cran.r-project.org/web/packages/mgcv/index.html>.
- Woodson, C.B., and Litvin, S.Y. (2015). Ocean fronts drive marine fishery production and biogeochemical cycling. *Proceedings of the National Academy of Sciences* 112(6): 1710-1715.
- Wu, J., Shen, W., Sun, W., Tueller, P.T. (2002). Empirical patterns of the effects of changing scale on landscape metrics. *Landscape Ecology* 17(8): 761-782.
- Wu, J., and Li, H. (2006). Concepts of scale and scaling. *Scaling and uncertainty analysis in ecology*: 3-15.
- Yackulic C.B., Chandler R., Zipkin E.F., Royle A., Nichols J.D., Campbell Grant E.H., Veran S. (2013). Presence-only modelling using MAXENT: when can we trust the inferences? *Methods in Ecology and Evolution* 4: 236-243.

- Yamada, M. (1954). Some remarks on the pygmy sperm whale *Kogia*. Scientific Report of the Whales Research Institute, Tokyo 9: 37-58.
- Yang, W.C., Chou, L.S., Jepson, P.D., Brownell, R.L., Cowan, D., Chang, P.H., Chiou, H.I., Yao, C.J. et al. (2008). Unusual cetacean mortality events in Taiwan, possibly linked to naval activities. *Veterinary Record* 162: 184-186.
- Yee, T.W., and Mitchell, N.D. (1991). Generalized additive models in plant ecology. *Journal of vegetation science* 2: 587-602.
- Yen, P.P., Sydeman, W.J., Hyrenbach, K.D. (2004). Marine bird and cetacean associations with bathymetric habitats and shallow-water topographies: implications for trophic transfer and conservation. *Journal of Marine systems* 50(1): 79-99.
- Zador, S.G., Parrish, J.K., Punt, A.E., Burke, J.L., Fitzgerald, S.M. (2008). Determining spatial and temporal overlap of an endangered seabird with a large commercial trawl fishery. *Endangered Species Research* 5(2-3): 103-115.
- Zaniewski, A.E., Lehmann, A., Overton, J.M. (2002). Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling* 157(2-3): 261-280.
- Zeileis, A., Kleiber, C., Jackman, S. (2007). Regression models for count data in R. Research Report Series / Department of Statistics and Mathematics, 53.
- Zerbini, A.N., and Kotas, J.E. (1998). A note on cetacean bycatch in pelagic driftnetting off southern Brazil. Report of the International Whaling Commission. Cambridge, UK.
- Zurell, D., Berger, U., Cabral, J.S., Jeltsch, F., Meynard, C.N., Münkemüller, T. et al. (2010). The virtual ecologist approach: simulating data and observers. *Oikos* 119(4): 622-635.
- Žydelis, R., Lewison, R.L., Shaffer, S.A., Moore, J.E., Boustany, A.M., Roberts, J.J. et al. (2011). Dynamic habitat models: using telemetry data to project fisheries bycatch. *Proceedings of the Royal Society of London B: Biological Sciences* 278(1722): 3191-3200.



# Annexes

---



© Laura Hedon

## CONTENTS

---

ANNEX A: COMPARISON OF HABITATS MODELS FOR SCARCELY DETECTED SPECIES.....	123
ANNEX B: HOW MANY SIGHTINGS TO MODEL RARE MARINE SPECIES DISTRIBUTIONS.....	153
ANNEX C: COMBINING VISUAL SURVEYS TO MODEL HABITAT OF DEEP-DIVING CETACEANS AT THE BASIN SCALE.....	187
ANNEX D: DATA-ASSEMBLING: A MATTER OF ECOSYSTEMS SIMILARITY – SUPPORTING INFORMATION.....	217
ANNEX E: WOULD MODELS BE IMPROVED IF PREY DISTRIBUTIONS WERE INCLUDED? .....	221





# Annex A

---

## COMPARISON OF HABITAT MODELS FOR SCARCELY DETECTED SPECIES

---

Auriane Virgili, Mélanie Racine, Matthieu Authier, Pascal Monestiez, Vincent  
Ridoux

*Ecological Modelling*, 2017, 346: 88-98

# Comparison of habitat models for scarcely detected species

Auriane VIRGILI <sup>1\*</sup>, Mélanie RACINE <sup>2</sup>, Matthieu AUTHIER <sup>2</sup>, Pascal MONESTIEZ <sup>1,3</sup>, Vincent RIDOUX <sup>1,2</sup>

<sup>1</sup>Centre d'Etudes Biologiques de Chizé - La Rochelle, UMR 7372 CNRS - Université de La Rochelle, Institut du Littoral et de l'Environnement, 17000 La Rochelle, France ; <sup>2</sup>Observatoire PELAGIS, UMS 3462 CNRS - Université de La Rochelle, Systèmes d'Observation pour la Conservation des Mammifères et des Oiseaux Marins, 17000 La Rochelle, France ; <sup>3</sup>BioSP, INRA, 84914 Avignon, France

## Abstract

When performing habitat models, modellers have to choose between presence-absence and presence-only models to estimate the habitat preferences of a species. Primarily, this choice depends on the data that are available and whether effort data are recorded in parallel to sighting data. For species that are rare or scarce, the models have to address a great number of zeros (*i.e.* no animal seen) that weakens the ability to make sound ecological inferences. We tested two types of habitat models (presence-absence vs. presence-only) to determine which type best dealt with datasets containing an excess of zeros, and we applied our models to a sighting dataset that included the common (*Delphinus delphis*) and striped (*Stenella coeruleoalba*) dolphin (approximately 92% zeros). We used two types of presence-absence models (Generalised Additive models – GAMs, Generalised Linear Model – GLM) and one presence-only model, a MaxEnt model, and we used various criteria to compare these models (*i.e.* AIC, deviances, rootograms and distribution patterns predicted by the models). Overall, we observed that the presence-absence models made better predictions than the presence-only model. Among the presence-absence models, the GAM with a Negative Binomial distribution was better at predicting small delphinids habitats, even though the GAM with a Tweedie distribution exhibited similar results. However, the zero-inflated Poisson distributions exhibited less convincing results and was contrary to what was expected. Finally, despite 92% zeros, our dataset was not zero-inflated. Our study demonstrates the importance of selecting appropriate models to make reliable predictions of habitat use for species that are rare or scarce.

**Keywords:** Habitat modelling; scarce species; GAM; GLM; MaxEnt; Poisson; Negative Binomial; Tweedie; zero-inflated Poisson

## A.1. INTRODUCTION

Identifying habitat needed and used by species is important for wildlife management and conservation (Cañadas et al., 2005; Bailey and Thompson, 2009). One means of identifying habitat is with statistical models that correlate the spatial distributions of animal sightings with environmental inputs (Austin, 2002; Guisan and Thuiller, 2005; Redfern et al., 2006). Such models allow the habitat of a species and presence to be estimated. They also allow for predictions in areas that have not been previously surveyed (Segurado et al., 2004).

Species distribution models have recently undergone rapid development and have been used for diverse applications (*e.g.* Elith et al., 2006; Elith and Leathwick, 2009; Mannocci et al., 2014a; 2014b;

2015). There are generally two categories of habitat models: presence-absence and presence-only models; the chosen model depends on the type of data used and, notably, whether effort data are recorded in parallel to sighting data (Guisan and Zimmermann, 2000).

The first group of models requires presence and effort data that are recorded during planned surveys, where each on-effort sighting represents a detection of the target species. Such presence-absence models include, among others, generalised linear models (GLM), generalised additive models (GAM), regression trees analyses such as boosted regression trees (BRT) (Guisan and Zimmerman, 2000; Brotons et al., 2004), or occupancy models (MacKenzie et al., 2002). Some of these models allow estimating detection probability, and consequently, prediction of habitat suitability of a species (Gormley et al., 2011). They also allow functional relationships to be fitted between species locations and local environmental conditions (Guisan and Zimmerman, 2000). The models of the second group only require detection data, such as opportunistic data, where the absence data are missing because effort data were not documented and non-detection data are not prospected and informed (Hirzel et al., 2002). These include Ecological Niche Factor Analysis (ENFA) or Maximum Entropy Modelling (MaxEnt), and allows for the identification of potentially suitable sites by evidencing the environmental conditions that are similar to the sites where animals were recorded (Elith et al., 2006). Nevertheless, the accuracy of presence-only model outputs is conditional on random or representative sampling of the habitat at the data collection stage (Yackulic et al., 2013). Presence-only data are the default option when data on absence (that is effort data) are not available (Zaniewski et al., 2002).

Except for presence-only models, which do not consider the zeros, choosing among presence-absence models might be difficult depending on the studied species, particularly when focusing on scarcely detected species, because of the inherent difficulty of models to accommodate a large number of absences. Due to restricted habitat range, low density and poor detection even in favourable habitats (Martin et al., 2005), the number of absences in some datasets (*i.e.* the zeros) can be large. True (or structural - the taxon is really absent from an area), and false (or sampling - the taxon is present but poorly detected) absences become particularly challenging to tell apart (Ridout et al., 1998).

Due to their discrete probability distribution, count data are basically modelled with a Poisson regression, but when compared to this Poisson distribution, ecological data are often over-dispersed (*i.e.* the variance is greater than the mean) and require specific treatment to avoid biased results (Ridout et al., 1998; Dobbie and Welsh, 2001). Failure to accommodate over-dispersion leads to the selection of a model that is more complex than necessary (Richards, 2008), where the model does not generalise outside the sample used to calibrate it. One reason for over-dispersion that has attracted much attention is zero-inflation (Deng and Paul, 2005), where a large abundance of zeros in a dataset needs to be adequately analysed to prevent model misspecification and misleading ecological conclusions due to the under- or over-estimation of some functional relationships. Too many zeros can also increase biases and uncertainties in the estimated model parameters (MacKenzie et al., 2002; Martin et al., 2005). Hence, habitat modellers face two main issues: first, they have to define if their data are under-, equi- or over-dispersed, and second, depending on their data, they have to find an appropriate model for the dispersion (for example, zero-inflated models).

The selection of a good enough (that is accurate) model is critical for habitat models to fulfil their potential for management and conservation purposes. Habitat models can reveal areas of high densities of organisms; they can help to define or confirm key areas of conservation in order to meet stakeholder

expectations (Cañadas et al., 2005). This is even more important when focusing on scarcely detected species because these areas of high densities are more difficult to identify.

Consequently, the aim of our study was to help habitat modellers find an appropriate model when working with data with many zeros. To do that, we compared the predictive performance of both presence-absence and presence-only models and tested their ability to address an apparently zero-inflated dataset. We used a small delphinids sightings dataset; which pooled the common *Delphinus delphis* and the striped *Stenella coeruleoalba* dolphin. These data include approximately 92% zeros. Small delphinids show distribution patterns that are easily identified by habitat modelling, and thus they allow a comparison of different models. They are typical top predators in that they are sparsely distributed *in natura*. Associated datasets are characterised by the presence of many zeros even within favourable habitats (Redfern et al., 2006). However, they provide sufficient data to fit various distribution models and statistically compare their outputs. Using this dataset, we tested different models: GAMs with a Poisson, a Negative Binomial, a Tweedie and a zero-inflated Poisson distribution; a GLM with a zero-inflated Poisson distribution and a presence-only model; the MaxEnt model. Due to their ability to model separately the absences and the presences (Lambert, 1992), we assumed *a priori* that a zero-inflated Poisson model would perform best. However, the Negative Binomial and Tweedie distributions can also provide good fits (Warton, 2005; Dunn and Smyth, 2005; Lindén and Mantyniemi, 2011). In addition, with its multiple applications (Yackulic et al., 2013), including those by managers, and its ability to take into account the complex interactions between response and predictor variables (Elith et al., 2006; 2011; Phillips et al., 2004; Phillips and Dudik, 2008), the MaxEnt model appears to be a relevant tool for modelling habitats of rare species (Wisz et al., 2008). Therefore, we also tested the model to assess its efficiency. This study aims to pragmatically answer some questions commonly asked by habitat modellers, such as those regarding the effective zero-inflation of their data and the relevance of the chosen model depending on their data.

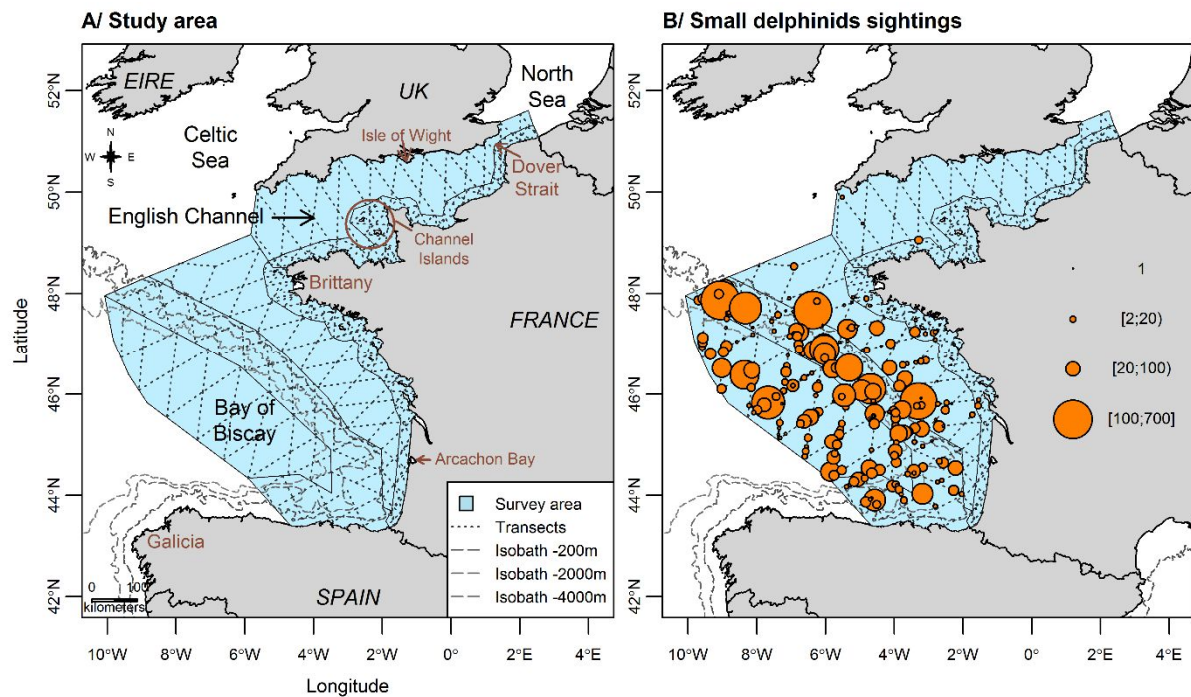
## A.2. MATERIALS AND METHODS

### A.2.1 Datasets

#### Aerial surveys and data collection

The small delphinids sighting data were recorded during the SAMM survey (*Suivi Aérien de la Mégafaune Marine*; Aerial Census of Marine Megafauna), which was dedicated to the observation of marine megafauna and conducted in the northeast Atlantic Ocean and the northwest Mediterranean Sea (Laran et al., in review; Lambert et al., in press). In the present work we focused on data collected in the summer of 2012 in the entire English Channel and the Bay of Biscay from the tip of Brittany to the Dover Strait in the north, and to the Spanish coast in the south (Fig. A.1). The survey was carried out from mid-May to early August along 31,427 km of transect lines. A standard methodology for cetacean surveys was applied (Hammond et al., 2013) using twin-engine high-wing aircrafts equipped with bubble windows. The flights followed a zig-zag pattern, at a speed of 167 km/h and an altitude of 183 m. Observation conditions (Beaufort seastate, turbidity, cloud cover and glare severity) and sightings with group size were recorded following a line-transect methodology (Buckland et al., 2001). This implies that the angle between every sighting and the track line was recorded to estimate the Effective Strip Width (ESW; see the small delphinids detection function and estimated ESW in Laran et al., in review).

The common and striped dolphins were pooled because it was most often impossible to tell apart the two species from the plane. During the survey, 277 sightings of small delphinids were recorded in good observation conditions, corresponding to 14,477 individuals (Fig. A.1).



**Fig. A.1. Study area (A) showing small delphinids sightings (B) recorded during the survey.** The study area covers the English Channel and the Bay of Biscay. Surveys were carried out along transects (dotted lines) following a zig-zag pattern, and sightings were classified by group size with each orange point representing a group of individuals (1, 2 to 20, 20 to 100, and 100 to 700 individuals).

### Environmental predictors

To model the relationships between small delphinids and their environment, we used eight environmental predictors (Lambert et al., in press; Virgili et al., in review), of which there were two physiographic variables (depth and slope) and six oceanographic variables (mean, variance and gradient of Sea Surface Temperature – SST, mean and standard deviation of Sea Surface Height – SSH, and the maximum velocity of tidal currents (Table A.1)). All oceanographic variables were computed at a seven day resolution, *i.e.* averaged over 6 days prior to the sampled day. Physiographic variables are static and relate to the bathymetry, and oceanographic variables are dynamic and describe water masses. These variables have been considered in the model selection procedure because they are all candidate drivers of small delphinids distribution via their effect on the functioning of pelagic ecosystems (Table A.1; Appendix A.1).

### A.2.2 Statistical models

#### Analytical strategy

In a first step, we arbitrarily chose a baseline model, a GAM with a Poisson distribution appropriate for equi-dispersed data, for comparison with the other models. For this baseline model, relationships between the abundances of small delphinids and environmental variables were investigated. Next, we fitted GAMs with a Negative Binomial and a Tweedie distribution, which are suitable for over-dispersed data, a GAM and a GLM with a zero-inflated Poisson distribution, which are suitable when over-

dispersion is due to zero-inflation, and a MaxEnt model, which is specific to presence-only data. Using these models, we applied the variables previously selected in the baseline model. Even if a model performance is largely determined by its selected variables (Syphard and Franklin, 2009), in this study, we applied the same variables for each tested model to assess how the results were affected by the model alone when using the same dataset. To finish, we compared all models by using different criteria such as the Akaike Information Criterion (AIC), the deviances, the rootograms (Kleiber and Zeilis, 2016) and the predicted density maps to evaluate the predictive performance of each model.

**Table A.1. Environmental predictors used for habitat modelling.** \* A: Depth and slope were computed from the GEBCO-08 30 arc-second database (<http://www.gebco.net/>). B: The mean and variance and gradient of Sea Surface Temperature (SST) were calculated from the ODYSSEA product (from My Ocean project <http://www.myocean.eu/>). C: The MARS 3D model from Previmer (2014, [www.previmer.org](http://www.previmer.org)) was used to compute mean and standard deviation of Sea Surface Height (SSH). D: The daily maximum intensity of the currents was computed from the MARS 2D model (Previmer, 2014, [www.previmer.org](http://www.previmer.org)).

Environmental predictors	Sources*	Effects on pelagic ecosystems of potential interest to top predators
<b>Physiographic</b>		
Depth (m)	A	Shallow waters could be associated with high primary production
Slope (°)	A	Associated with currents, high slope induce prey aggregation and/or primary production increasing
<b>Oceanographic</b>		
Mean of SST (°C)	B	Variability over time and horizontal gradients of SST reveal front locations, potentially associated to prey aggregations
Variance of SST (°C)	B	
Mean gradient of SST (°C)	B	
Mean of SSH (m)	C	High SSH is associated with high mesoscale activity and prey aggregation and/or primary production increase
Standard deviation of SSH (m)	C	
Daily maximum intensity of the currents (m.s <sup>-1</sup> )	D	High currents induce water mixing and prey aggregation

#### Baseline model: GAM with a Poisson distribution

A Generalised Additive Model (GAM; Hastie and Tibshirani, 1986) with a Poisson distribution (variance equal to the mean), hereafter called PO-GAM, was retained as the baseline model (Table A.2). The response variable was linked to the additive predictors using a log-link function. We included, as an offset, the effort per segment (Hastie and Tibshirani, 1986). This offset was calculated as the segment linear length multiplied by twice the ESW (Effective Strip Width estimated from Conventional Distance Sampling, see Laran et al., in review). The model was fitted using R-3.1.2 (R Core Team, 2016) with the mgcv package (Wood, 2006; 2013) by restricting polynomial smoothness to three degrees of freedom (Ferguson et al., 2006).

In the selection procedure, all models with a combination of one to four variables were tested and the combinations of variables with a correlation coefficient higher than |0.7| were excluded. A maximum of four covariates was implemented to avoid excessive complexity and difficulty of interpretation (Mannocci et al., 2014a; 2014b). The Akaike Information Criterion (AIC) was used to select the best models, where the lower the AIC the better the model (Akaike, 1974). Finally, we extracted the

explained, null and residual deviances and the residuals to assess the goodness-of-fit of the selected models.

**Table A.2. Details of the models used in the study.** GAM: Generalised Additive Model; GLM: Generalised Linear Model; PO: Poisson; NB: Negative Binomial; TW: Tweedie; ZIP: Zero-Inflated Poisson; PA: Presence-Absence data; and AIC: Akaike Information Criterion. \* R Core Team (2016)

Generic models	Used names	Data	Settings and details
Generalised Additive Model with Poisson distribution	PO-GAM	Equi-dispersed PA	Used R-3.1.2*, package <b>mgcv</b> , function GAM, <i>Poisson</i> distribution, log-link function, included an offset, 3 degrees of freedom for the smoothing curve functions, used AIC to select the better model
Generalised Additive Model with Negative Binomial distribution	NB-GAM	Over-dispersed PA	Used R-3.1.2*, package <b>mgcv</b> , function GAM, <i>Negative Binomial</i> distribution, log-link function, included an offset, 3 degrees of freedom for the smoothing curve functions, no selection of the better model
Generalised Additive Model with Tweedie distribution	TW-GAM	Over-dispersed PA	Used R-3.1.2*, package <b>mgcv</b> , function GAM, <i>Tweedie</i> distribution, log-link function, included an offset, 3 degrees of freedom for the smoothing curve functions, no selection of the better model
Generalised Additive Model with Zero-Inflated Poisson distribution	ZIP-GAM	Zero-inflated PA	Used R-3.1.2*, package <b>mgcv</b> , function GAM, <i>ZIP</i> distribution, log-link function, included an offset, 3 degrees of freedom for the smoothing curve functions, no selection of the better model
Generalised Linear Model with Zero-Inflated Poisson distribution	ZIP-GLM	Zero-inflated PA	Used R-3.1.2*, package <b>mgcv</b> , function GAM, <i>ZIP</i> distribution log-link function, included an offset, no smooth, no selection of the better model
Maximum Entropy Modelling	MaxEnt	Presence-only	Used MaxEnt software version 3.3.3, no selection of the better model, <i>hinge</i> feature, default prevalence of 0.5, logistic output format

### Challenger models

For all other models (Table A.2), we used the variables associated with the best selected model following 2.3.1, and there was no covariate selection procedure. As for the baseline model, we extracted the explained, null and residual deviances and checked the distribution of residuals for each model, except for the MaxEnt model. In this part, the models are briefly described, but they are more developed in Appendix A.2.

For the GAM with a Negative Binomial and a Tweedie distribution, we used the R package **mgcv** (Wood, 2006; 2013) with the `gam`, `nb` and `tw` functions to estimate the model parameters and the relationship between environmental variables and small delphinids densities. Hereafter, the fitted model will be called NB-GAM and TW-GAM.



Next, we tested two types of zero-inflated models, with linear (GLM) and nonlinear (GAM) relationships between the response variable and the predictors and considering a zero-inflated Poisson distribution. We used the *mgcv* package (Wood, 2006; 2013) with the *gam* and *ziP* functions to estimate the parameter of the models. Hereafter, the fitted models will be called ZIP-GAM and ZIP-GLM.

Finally, we fitted a presence-only model with Maximum Entropy (MaxEnt), in which relationships with the environment are estimated using background samples of the environment instead of absence locations (Elith et al., 2011). We used the MaxEnt version 3.3.3 (<http://www.cs.princeton.edu/~schapire/maxent/>; Phillips et al., 2006). The input file was the same as the baseline model, but we removed all absences; hence, each line corresponded to one observation of small delphinids and for the environmental predictors, we used the four covariates selected by the baseline model. Regarding model parameters, we used the “*hinge*” feature to generate models with smooth functions similar to GAM ones, with a default prevalence of 0.5 and a logistic output format to compare it to the probability of presence (Phillips and Dudík, 2008; Elith et al., 2011; Merow et al., 2013).

### Predictions

For each fitted model, except for MaxEnt, which directly provides a prediction map, we computed the predictions and their associated coefficients of variation for each day of the survey (85 days) on a 0.05°x0.05° resolution grid. Next, daily predictions were averaged over the entire period to produce maps of averaged density of small delphinids. Finally, we provided uncertainty maps that corresponded to the standard errors of the predictions. To limit extrapolation, all predictions were constrained within the envelope of sampled values of covariates used to fit the model.

In addition, we assessed whether a prediction was an extrapolation or an interpolation using the non-parametric Gower’s distance (King and Zeng, 2007). An extrapolation is a prediction for a combination of covariate values that falls outside the convex hull which is defined by the covariate data used to calibrate the model (King and Zeng, 2007; Authier et al., 2016). However, even if a prediction falls outside this convex hull, this extrapolation can nevertheless be informed by calibration data lying in its neighbourhood. The neighbourhood of a prediction was defined as the calibration covariate data within a radius of one geometric mean Gower’s distance of the prediction (King and Zeng, 2007). The geometric mean was computed from all pairs of calibration data point. The results from this extrapolation analysis were mapped to visually assess how trustworthy the predictions were.

#### A.2.3 Model comparison

Evaluating the predictive performance of a model requires demonstrating its consistency with raw observation data and comparing the outputs of several models (Pearce and Ferrier, 2000). Each assessment criterion quantifies a particular aspect of a model performance and several criteria must be used in combination (Elith and Graham, 2009). We calculated different selection measures to improve the relevance of model comparison.

First, an Akaike Information Criterion was computed for each model to assess model relative fit: the lower the AIC, the better the model (Akaike, 1974). Second, we examined several deviance-based quantities (null, residual and explained) as a proxy of the model reliability to predict the frequencies of species occurrence (Elith and Graham, 2009). A high explained deviance can indicate a good fit, whereas a high null deviance and a high residual deviance can indicate a bad one. Finally, to evaluate the absolute

goodness-of-fit of the models and how they handled the excess of zeros, we plotted rootograms that compared, with histograms, the raw data frequencies to the frequencies fitted with the models (Kleiber and Zeileis, 2016).

The methods cannot be readily applied with presence-only models, which leads to some complexity in the methods of model comparison. To evaluate the predictive performance of MaxEnt, we used the Area Under the receiver operating characteristic Curve (AUC; Elith et al., 2006). This method works only on binary data (not on count data) and measures how a model can differentiate the sites where the species is present and the sites where it is absent. A perfect discrimination of the sites is revealed by a score of 1, a discrimination equivalent to a random distribution is indicated by a score of 0.5 and for a score lower than 0.5, the model performance is worse than a random guess (Elith et al., 2006). This AUC is directly provided by the MaxEnt software. However, with this method, we cannot compare the model performance to the fitted baseline model. We thus transformed the PO-GAM prediction maps (only this one) to probability of presence with the formula  $presence\ probability = 1 - e^{(-predicted\ density)}$ .

## A.3. RESULTS

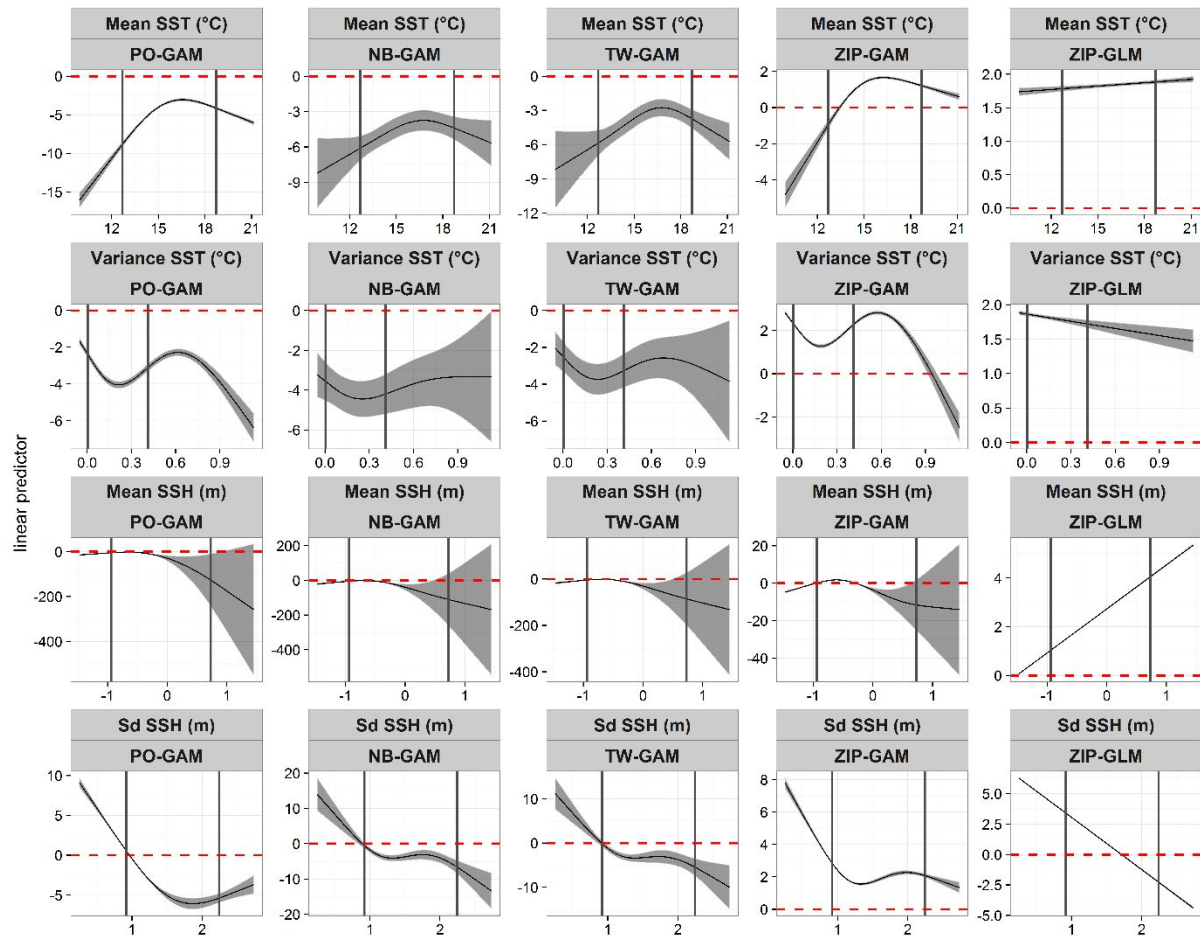
### A.3.1 Model selection

Among the eight environmental predictors, the variables selected by the best PO-GAM, defined as the baseline model, were the mean and variance of SST and the mean and standard deviation of SSH (Fig. A.2). The highest densities of delphinids were predicted for stable temperatures at approximately 16°C (variance around 0°C), and a rather stable low average altimetry (SSH, around -0.5 m, standard error around 0.5 m).

The NB-GAM ( $k=0.028$ ), TW-GAM ( $p=1.573$ ) and ZIP-GAM ( $\vartheta=(-4.861, 0.26)$ ) showed fairly similar smooth functions compared to the PO-GAM, except for sd SSH (Fig. A.2). However, confidence intervals around the functional relationships were significantly smaller and the predicted densities were higher in the case of the ZIP-GAM. The smooth functions of the ZIP-GLM showed increasing small delphinids densities with increasing SST mean and SSH mean and decreasing densities with decreasing SST variance and sd SSH, which was an opposite trend compared to the other models for SSH mean.

To complete the comparison of the models, we analysed the residuals of each fitted model (Appendix A.3). In all cases, there was an accumulation of residuals at zero and an over-dispersion of positive values, but it was less important for the TW-GAM. In addition, we calculated Cook's distances (Appendix A.4) to determine if some values highly influenced the fitted models (Cook's distance > 1). It appeared that some values greatly influence the PO-GAM and ZIP-GLM. However, for NB-GAM, TW-GAM and ZIP-GAM, no value appeared to affect the models (Cook's distance < 1) and the only values that could influence them correspond to non-extreme values of covariates. Consequently, that strengthened the results provided by the fitted models, especially for NB-GAM, TW-GAM and ZIP-GAM.

Finally, with an AUC of 0.822, the MaxEnt model predicted delphinids presence probabilities much better than a random prediction would do (AUC of 0.5).



**Fig. A.1.** Forms of smooth functions for the selected covariates for each presence-absence model. The solid line in each plot is the smooth function estimate and shaded regions represent approximate 95% confidence intervals. The y-axis indicates the logarithm of the abundance in individual/km<sup>2</sup>. The x-axis indicates the values of the covariates and zero on the x-axis indicates no effect of the covariate. Best model fits are between the vertical lines indicating the 10<sup>th</sup> and 90<sup>th</sup> quantiles of the data.

### A.3.2 Predictions

Prediction maps of the PO-GAM showed a concentration of delphinids in offshore waters, from the continental shelf to the oceanic waters, with the highest densities over the slope (Fig. A.3). The highest densities, which reached 30 individuals·km<sup>-2</sup>, were predicted in the north of Galicia, which is outside the survey area. In addition, we noticed a good match between observations and predictions of the model (Fig. A.1). Within the survey area, predictions were associated with low uncertainties (Appendix A.6), which strengthened the results. In contrast, outside the survey area, patches of high densities were associated with higher uncertainties and needed to be considered with caution.

The TW-GAM predicted exactly the same distribution as the previous model but with slightly higher densities (maximum at 35 individuals·km<sup>-2</sup>; Fig. A.3; Appendix A.5). The NB-GAM also predicted the same distribution as PO-GAM but with higher densities (maximum at 73 individuals·km<sup>-2</sup>, Fig. A.3; Appendix A.5). The ZIP-GAM showed the same distribution patterns in the Bay of Biscay but with lower densities (maximum at 11 individuals·km<sup>-2</sup>, Fig. A.3; Appendix A.5) and more individuals predicted near the coasts. However, contrary to the PO-GAM, this model predicted delphinids in the western English Channel, with a concentration of individuals around the Channel Islands (Appendix A.5). Regarding the ZIP-GLM,

densities were also predicted in offshore waters, approximately 5 and 10 individuals·km<sup>-2</sup> and similar to the other models, but a larger patch was identified and located west of the Isle of Wight with more than 2,000 individuals·km<sup>-2</sup> (Fig. A.3; Appendix A.5). Similarly to the PO-GAM, high predicted densities of the NB-GAM and TW-GAM were associated with high uncertainties outside the survey area but low uncertainties in the survey area. For ZIP-GAM and ZIP-GLM, patches of high densities in the survey area were associated with uncertainties, making the predictions less reliable (Appendix A.6).

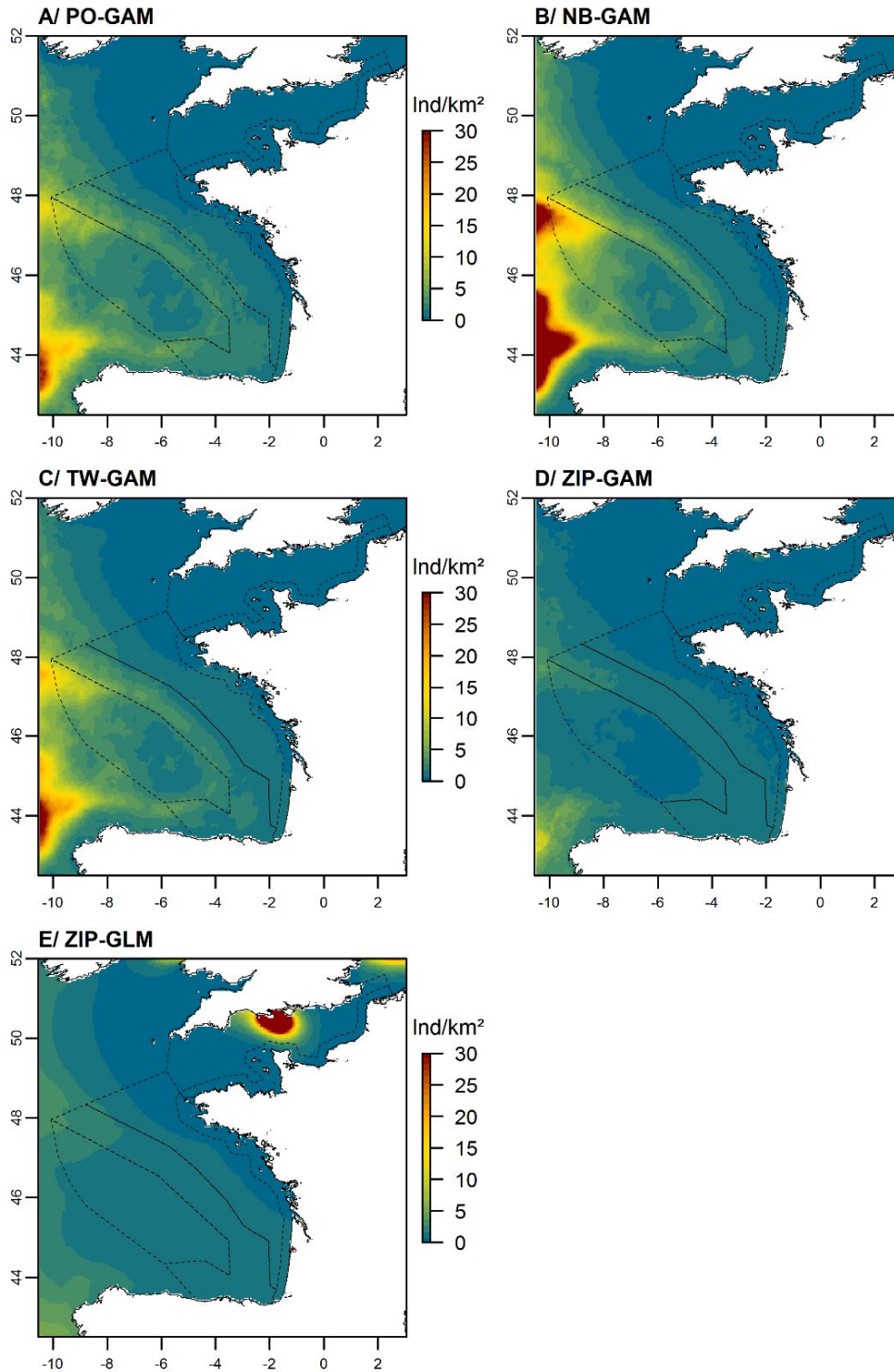
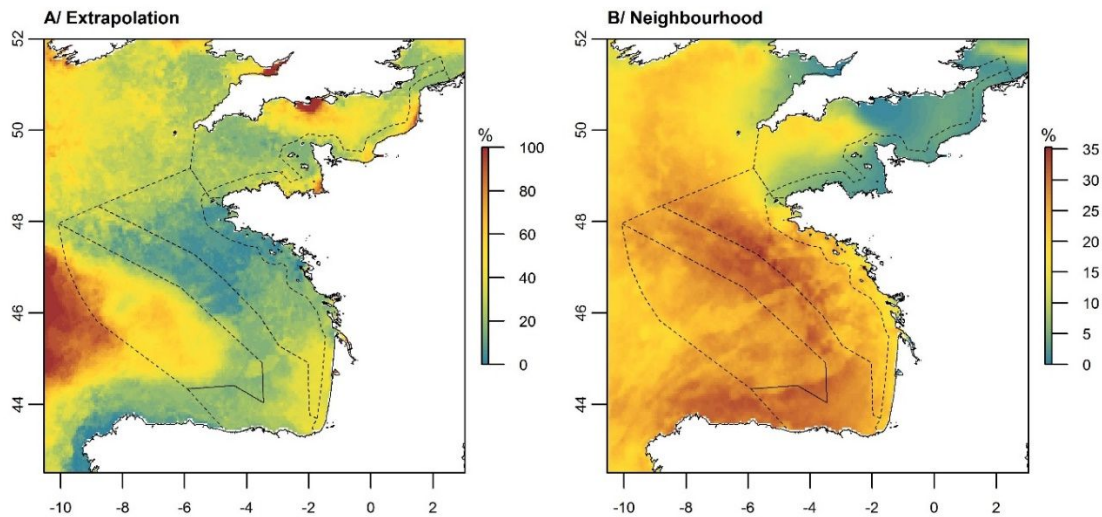


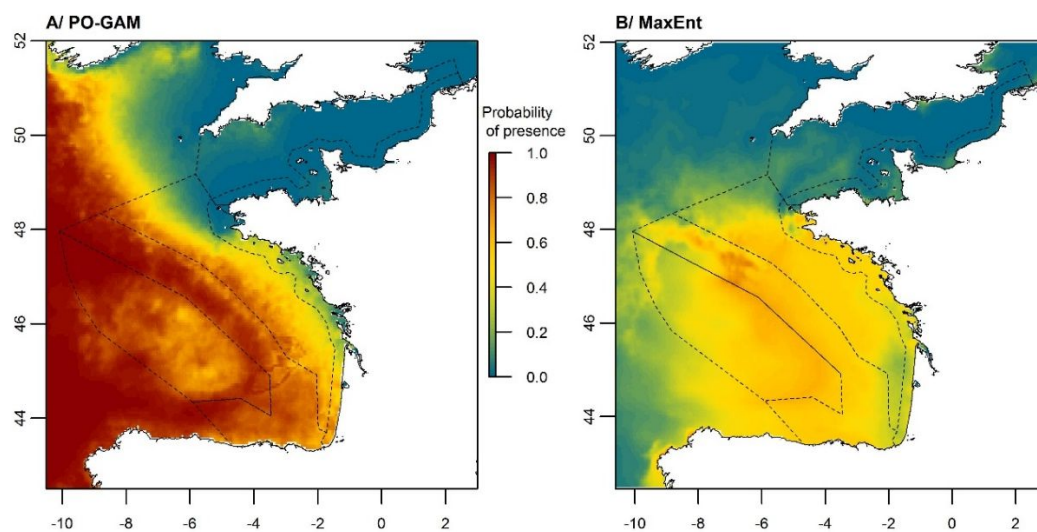
Fig. A.3. Predicted distributions of small delphinids in individuals·km<sup>-2</sup> (Ind/km<sup>2</sup>) for each presence-absence model in the Bay of Biscay and the English Channel. Dotted lines represented the survey area. The scale of the PO-GAM was applied on all maps to facilitate the comparison, and Appendix A.5 shows the maps with their own scale.

Extrapolation and neighbourhood maps (Fig. A.4) allowed us to assess the reliability of the predictions obtained with the fitted models. Overall, a high percentage of extrapolation and a low percentage of calibration data used to inform neighbouring cells (neighbourhood) indicated unreliable predictions. Hence, a model that predicted densities in the English Channel was inconsistent, which was particularly the case with ZIP-GLM and to a lesser extent, the case with ZIP-GAM. PO-GAM, NB-GAM and TW-GAM all predicted high densities outside the survey area, but according to Fig. A.4, these predictions were reliable because they were informed by approximately 20% of the data used to calibrate the model. However, NB-GAM made more extreme extrapolations than the other models in the Bay of Biscay.



**Fig. A.4.** Extrapolation analysis using Gower's distance (King and Zeng, 2007). The extrapolation map (A) assesses whether a prediction was an extrapolation (100%) or an interpolation (0%) and the neighbourhood map (B) represents the percentage of calibration covariate data which informed each cell.

The MaxEnt model predicted higher probabilities of occurrence in the Bay of Biscay, particularly over the slope but also fairly evenly spread along the coasts of the Bay of Biscay and southwest England (Fig. A.5). Compared to the PO-GAM (see prediction map in probability of presence, Fig. A.5), MaxEnt hardly extrapolated beyond the sampled area and the predicted probabilities of presence were lower.



**Fig. A.5.** Distributions predicted by PO-GAM (A) versus MaxEnt (B) in the Bay of Biscay and the English Channel. Dotted lines represent the survey area. The same scale was applied for the two model to facilitate the comparison.

### A.3.3 Evaluation and comparison of models

The NB-GAM showed the lowest AIC and was followed by TW-GAM, whereas the PO-GAM showed the highest (Table A.3). The explained deviances varied between 7.3% for the ZIP-GLM and 39.1% for the TW-GAM (Table A.3). In addition, the lowest null and residual deviances were computed for the NB-GAM, which indicated a better fit of the model compared to TW-GAM that, despite a high explained deviance, showed very high null and residual deviances (Table A.3). The ZIP models performed worse than NB-GAM or TW-GAM but better than PO-GAM. Overall, the NB-GAM showed a better predictive performance than the other models.

**Table A.3. Indices used for the comparison of the presence-absence models.** AIC: Akaike Information Criterion.

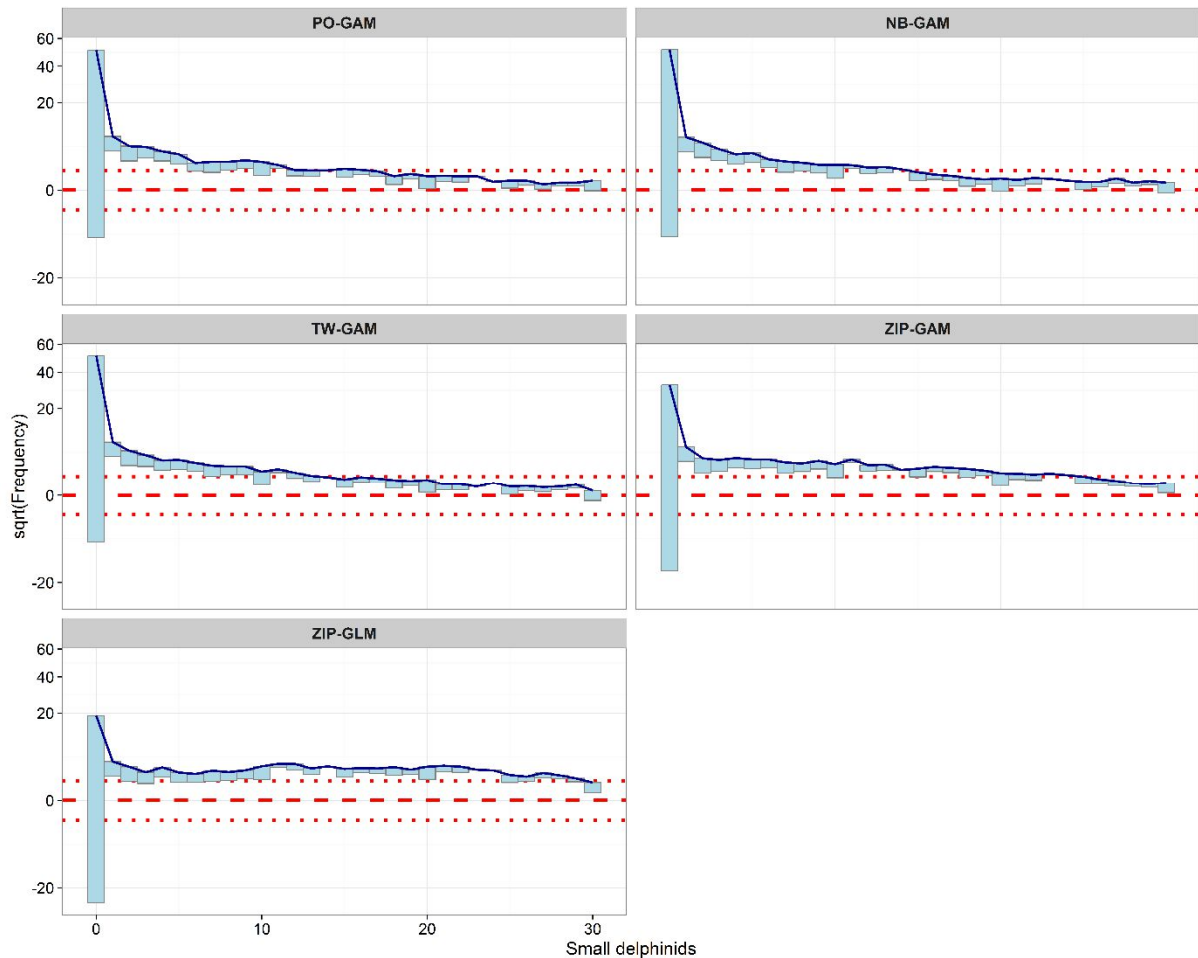
	PO-GAM	NB-GAM	TW-GAM	ZIP-GAM	ZIP-GLM
<b>AIC</b>	70,082	4,284	5,869	25,438	28,001
<b>Explained deviance</b>	28.6 %	38.4 %	39.1 %	17.1 %	7.3 %
<b>Null deviance</b>	96,266	1,001	41,341	27,305	27,146
<b>Residual deviance</b>	68,742	616	25,168	22,635	25,164

Visible with rootograms (Fig. A.6), all fitted models were not adequate for counts between 0 and 5 individuals. In all cases, the observed frequencies (blue bars) were not included in the confidence intervals (red dots). PO-GAM, NB-GAM and TW-GAM did not predict enough zeros and predicted too many sightings from 1 to 5 individuals. Likewise, ZIP-GAM and ZIP-GLM did not predict enough zeros and predicted too many sightings beyond 5 individuals. Albeit all fitted models tended to over-predict the frequencies between 1 and 5 individuals, the highest number of observed frequencies included in the confidence interval was observed for the NB-GAM thus making it the best fitted model.

## A.4. DISCUSSION

### A.4.1 General

We compared different types of habitat models, particularly presence-absence and presence-only models, to choose one that would be the most suitable for a scarce species. We found that a GAM with a Negative Binomial distribution was the most appropriate model to fit the data, even though the GAM with a Tweedie distribution also showed good predictions. In contrast, the zero-inflated Poisson distributions showed less convincing results, which was contrary to what was expected. We also found that MaxEnt provided quite good results compared to PO-GAM. These findings suggest that habitat for species that are rare or seldom seen are best described using presence-absence models such as GAM with a Negative Binomial distribution (Warton, 2005). However, it is important to also recognise that we used a particular biological system and we did not test all possible models.



**Fig. A.6. Rootograms obtained for each model.** Bars represent the observed frequencies, and solid lines represent the fitted frequencies. Frequencies are in a square-root scale. Red dot lines represent the confidence intervals, in which the blue bars have to be included to indicate a good fit of the models. The blue bars only have to intersect the confidence intervals to consider the fitted model as adequate.

### Biological system

Small delphinids, defined here as common and striped dolphins taken together, were selected as study material for several reasons. First, they provided a large enough dataset to allow proper statistical analyses. The present dataset included 277 sightings corresponding to approximately 14,477 individuals; consequently, the models tested here could be fitted without generating error messages during optimisation. Second, this dataset included more than 90% zero observations, which could be suggestive of zero-inflation. Third, in a previous investigation, Lambert et al., (in press) showed that small delphinids have well-defined patterns of distribution over the Bay of Biscay, which helped in evaluating predictive accuracy.

### Baseline model

In this study, we chose a panel of eight environmental predictors, both static and dynamic, as they are considered proxies for primary production and consequently, prey distribution (Austin, 2002). To compare the habitat models of small delphinids, we used a GAM with a Poisson distribution (PO-GAM) as a baseline model. The effectiveness of GAMs has been repeatedly demonstrated (Ferguson et al., 2006; Vilchis et al., 2006; Becker et al., 2010; Mannocci et al., 2014a; 2014b) and we chose a Poisson

distribution as the baseline because it characterises equi-dispersed data. Equi-dispersion is expected in the idealised situation where each detection event is independent of the others. In addition, GAMs are commonly used because they are more flexible than GLMs thanks to their semi-parametric functions that can accommodate non-linear relationships between animal densities and environmental predictors (Becker et al., 2010). We then established a variable selection procedure to define the best fitted model to be used as the baseline. This fitted model showed a relatively high explained deviance (28.6%) and interpolated rather than extrapolated (Fig. A.4). The selected model appeared ecologically consistent, which strengthened our model choice. A preference of eastern North Atlantic small delphinids for waters warmer than 15°C and depths between 400 and 1000 m and with concentrations along the shelf edge and lower densities in the western Channel and in coastal waters of the Bay of Biscay has been shown (Figs. A.2 and A.3) and already mentioned by Cañadas et al., (2009), MacLeod et al., (2009), Murphy et al., (2013) and Lambert et al., (in press).

### Challenger models choice

The choice of the challenger models was an important step in the comparison process. For the presence-absence models, the aim was not to test all existing models but to answer the pragmatic question: “What type of model should we use if the dataset contains more than 90% zeros?” To answer this question, we built realistic models that included linear (GLM) or non-linear (GAM) relationships between the response variable and predictors and tested different structural choices for the data likelihood (Tweedie, Negative Binomial and ZIP). All these models handle differently the datasets with extra zeros (Appendix A.2). A ZIP model links two sub-models: a binomial model for the zero count that distinguishes between true and false absences; and a Poisson count model for non-zero observations. Conditional on an observation not being a true absence, equi-dispersion is assumed. True absences are in this case the only source of over-dispersion. Tweedie and Negative Binomial models directly include an assumption on the relationship between the mean and the variance, in that they address over-dispersion in a more phenomenological way because the micro-level process generating over-dispersion is not explicit (Dunn and Smyth, 2005; Ridout et al., 1998; Zeileis et al., 2007, Wenger and Freeman, 2014).

Variable selection was only done on the baseline model. This could have led to sub-optimal models for the other likelihood choices (Tweedie, Negative Binomial, etc...): the performance of a model is largely determined by its selected variables (Syphard and Franklin, 2009). We decided to hold the set of covariates constant over models to assess how the results were affected by the structural choice in the model only. This corresponds to the idealised situation where the habitat of a species is known from previous investigations, but there is uncertainty in the exact model structure to predict its habitat. In practice, variable selection depends also on the model structure.

To assess the robustness of our results, we also ran a variable selection procedure for each model (not shown). For ZIP-GAM, the same variables were selected, but for NB-GAM and TW-GAM, the gradient of SST was selected over the variance of SST (two variables for which the potential effect on the pelagic ecosystems is quite similar; Table A.1). All other variables were identical. The model with the four variables selected by the baseline model (PO-GAM) was the second model in the two cases. All GAMs were almost identical with respect to the set of selected covariates. The biggest difference was observed in ZIP-GLM where the mean and variance of SST were replaced by the slope and gradient of SST; the model with the four variables selected in the baseline model was the 13<sup>th</sup> model. A complete comparison (Appendix A.7) revealed that the predictions of all the “best” GAMs were similar, but the



predictions of the ZIP-GLM were greatly different from the rest and they were more likely to be extrapolations.

### MaxEnt

We also wanted to test a presence-only model because it is not expected to be bad for scarce species as it does not see the zeros and it is easier to fit since corresponding data are more readily available (Tsoar et al., 2007). Indeed, most available data of species distribution are presence-only records because they are easier to collect, and contrary to presence-absence data, do not require recording effort data (Tsoar et al., 2007). In the case of rare and elusive species, opportunistic data, a common example of presence-only data, often represent the largest set of available data (Pearce and Boyce, 2006). Hence, it appeared necessary for habitat modellers, who have to choose between several models, to know if presence-only models could provide good predictions of species distribution relative to other more comprehensive methods based on presence-absence data (which requires effort data to be properly recorded). Among all presence-only models (Elith et al., 2006; Tsoar et al., 2007; Monk et al., 2010), we chose the MaxEnt model because it appeared more suitable to model the predictions of species distribution with complex interactions between the response and the predictor variables (Elith et al., 2006; 2011; Phillips et al., 2004; Phillips and Dudik, 2008) and seemed to manage datasets characterised by scarce data well (Wisiz et al., 2008).

#### A.4.2 Pragmatic habitat modelling of scarce species

Although environmental variables used in the models were identical, each fitted model showed a different predictive performance based on its own characteristics. Overall, NB-GAM and TW-GAM were very similar in the improvement they provided over PO-GAM (Appendix A.3) and estimated similar non-linear relationships with environmental covariates. NB-GAM exhibited the best predictive performances with the smallest AIC and a moderate explained deviance. Habitat predictions from models PO-GAM, TW-GAM and NB-GAM were qualitatively similar, suggesting robustness with respect to extrapolation (Fig. A4) and consistency in the results but predicted densities were larger in magnitude with NB-GAM. The overall bad performance of the only GLM among the candidate set of models stressed the importance of non-linear relationships in habitat modelling of small delphinids in the Bay of Biscay. Thanks to their flexibility, GAMs are appropriate for modelling the distribution of sparsely distributed megafauna either marine or terrestrial (Wood, 2006; Becker et al., 2010; Hegel et al., 2010). Thus, NB-GAM and TW-GAM were able to fit the data well despite the huge number of zeros, as seen on the rootograms (Fig. A.6).

Fitted ZIP models showed lower explained deviances (17.1% for ZIP-GAM and 7.3% for ZIP-GLM), lower predictive performances (higher AIC) and less ecologically consistent predictions with extrapolation of the predicted densities. Indeed, ZIP-GAM and ZIP-GLM predicted large densities of small delphinids in the English Channel where no sightings were recorded. Moreover, previous studies evidenced that these species generally avoid this area (Cañadas et al., 2009; MacLeod et al., 2009; Murphy et al., 2013; Lambert et al., in press). The disappointing performance of ZIP-GAM was somewhat surprising. We expected, following Barry and Welsh (2002), a better performance of this model because it mixes a zero-inflated model with the non-parametric functions of a GAM. In fact, the results were less convincing than NB-GAM or TW-GAM results because of the lower explained deviance, higher AIC and unrealistic densities predicted in the English Channel. This is likely due to the current parametrisation

of the ZIP family in mgcv. In fact, the current parametrisation uses the linear predictors and linearly scales them on a logit scale to generate extra-zero observations (see the help pages in mgcv v1.8-9; Wood, 2013). This parametrisation implicitly assumes that areas with lower densities have a higher probability of non-detection, which is *a priori* reasonable. However, it does not allow for incorporating detection-specific covariates which may better explain non-detection patterns. Despite 92% zeros in the data, ZIP models showed worse results than NB-GAM or TW-GAM; over-dispersion was not mainly due to zero-inflation. Even though the best model we selected did not completely accommodate all the zero observations, suggesting some zero-inflation (Fig. A.6), the latter was arguably less prevalent than initially thought.

MaxEnt showed a fairly high predictive performance (AUC=0.82) and distribution patterns relatively similar to those of PO-GAM, albeit more spread out in the whole study area. However, this model underestimated the probabilities of presence compared to PO-GAM and did not extrapolate beyond the study area. This presence-only model appeared relatively efficient to establish distribution patterns in a given survey area (Tsoar et al., 2007) and to identify areas of high probabilities of presence when only presence data were available (Zaniewski et al., 2002; Gormley et al., 2011). However, this may result more from the sampling design than from MaxEnt modelling *per se*. Data were collected with a standard protocol that ensured almost uniform coverage over the Bay of Biscay; no area was over- or under-sampled. The main issue with presence-only models is that they cannot account for uneven effort and must assume that the sampling of the habitat was random in order to interpret MaxEnt predictions correctly (Yackulic et al., 2013). Thus, our results might give too much of an optimistic outlook of the performance of MaxEnt. To moderate that optimism, cross-validation with portions of the study area removed “in block” would have been useful but was beyond the scope of this study.

Finally, as Warton (2005) warned, “many zeros does not mean zero-inflation” of the data, and even 92% zeros does not necessarily mean zero-inflation. We would recommend to habitat modellers, even if they study scarce species, to first test over-dispersed models such as GAMs with Tweedie or Negative Binomial distributions before testing zero-inflated models. Obviously, the predictive performance of the model has to be assessed for the tested model. A useful visual method to assess whether a model adequately addresses many zeros is the rootogram (Minami et al., 2007; Kleiber and Zeileis, 2016).

#### A.4.3 Management applications

Habitat models can be useful to delineate marine protected areas (Cañadas et al., 2005; Lambert et al., in review). These models allow investigating species habitat preferences (Austin, 2002) and revealing contiguous areas of high predicted densities, thereby highlighting potential areas of conservation (Cañadas et al., 2005). They can accommodate low sampling effort and remain useful in identifying suitable habitat despite few recorded sightings. For example, due to the zig-zag pattern of the SAMM survey transects, areas on the continental slope were not entirely prospected but habitat models predicted relatively high densities throughout the stratum without extrapolating. Habitat models can also help with sampling gaps that might necessitate further extensive effort to validate predictions (Bailey and Thompson, 2009).

Scarcely detected species present additional challenges for habitat modelling. Due to the small number of sightings compared to the deployed effort, it can be difficult to obtain reliable predictions and establish conservation plans for these species. However, these sparsely distributed species may

face many threats and require conservation action plans. We outlined in this study a pragmatic approach to build habitats models when focusing on scarce or rare species.

## A.5. CONCLUSIONS

Modelling the habitats of cetaceans or large predators in general is challenging because these organisms are by nature sparsely distributed compared to lower trophic levels, and their detection is often imperfect. This situation results in scarce datasets when survey effort is low to modest, and heavy zero-rich datasets when the amount of survey effort is large. However, statistical models generally require large presence-absence datasets to fit count data to environmental predictors. It is arguably easier for managers to use presence-only data rather than presence-absence data because effort data are not required.

Thanks to a homogeneous subadjacent effort, the MaxEnt model, a presence-only model, provided relatively good results in this study. Despite its lower accuracy, it would provide good enough predictions for small delphinids presence in the Bay of Biscay, although its ability to predict outside the survey area seemed limited. Among the presence-absence models, non-linear models predicted best small delphinids habitat. Contrary to what was expected, zero-inflated models were not the best predictive models; we thought that with 92% zeros, the data would be zero-inflated, but a thorough analysis revealed that they were mostly over-dispersed and not zero-inflated. Our study shows the importance of selecting appropriate models (beyond variable selection) to make reliable predictions of habitat use for species that are rare or scarce and that an abundance of zeros does not necessarily mean zero-inflation.

### Acknowledgments

We are grateful to the French Ministry in charge of the environment (*Ministère de l'Environnement, de l'Énergie et de la Mer*) and the Agency for marine protected areas (*Agence des aires marines protégées*) for funding the project. We thank Hélène Falchetto for processing the survey data that were used in the study. We are grateful to PREVIMER for providing us outputs from MARS-2D and MARS-3D models. We thank the *Direction Générale de l'Armement* (DGA) for a financial support to this study and funding A. Virgili's PhD. We thank Charlotte Lambert and Andrew Trites for their very helpful comments on this manuscript. We are grateful to reviewers for their comments on the article.

### References

- Akaike, H., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), pp.716–723.
- Austin, M.P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, 157, pp.101–118.
- Authier, M., Saraux, C. & Péron, C., 2016. Variable Selection and Accurate Predictions in Habitat Modelling: a Shrinkage Approach. *Ecography*, 39, pp.001-012.
- Bailey, H. & Thompson, P.M., 2009. Using marine mammal habitat modelling to identify priority conservation zones within a marine protected area. *Marine Ecology Progress Series*, 378, pp.279–287.
- Barry, S.C. & Welsh, A.H., 2002. Generalized additive modelling and zero inflated count data. *Ecological Modelling*, 157(2-3), pp.179–188.

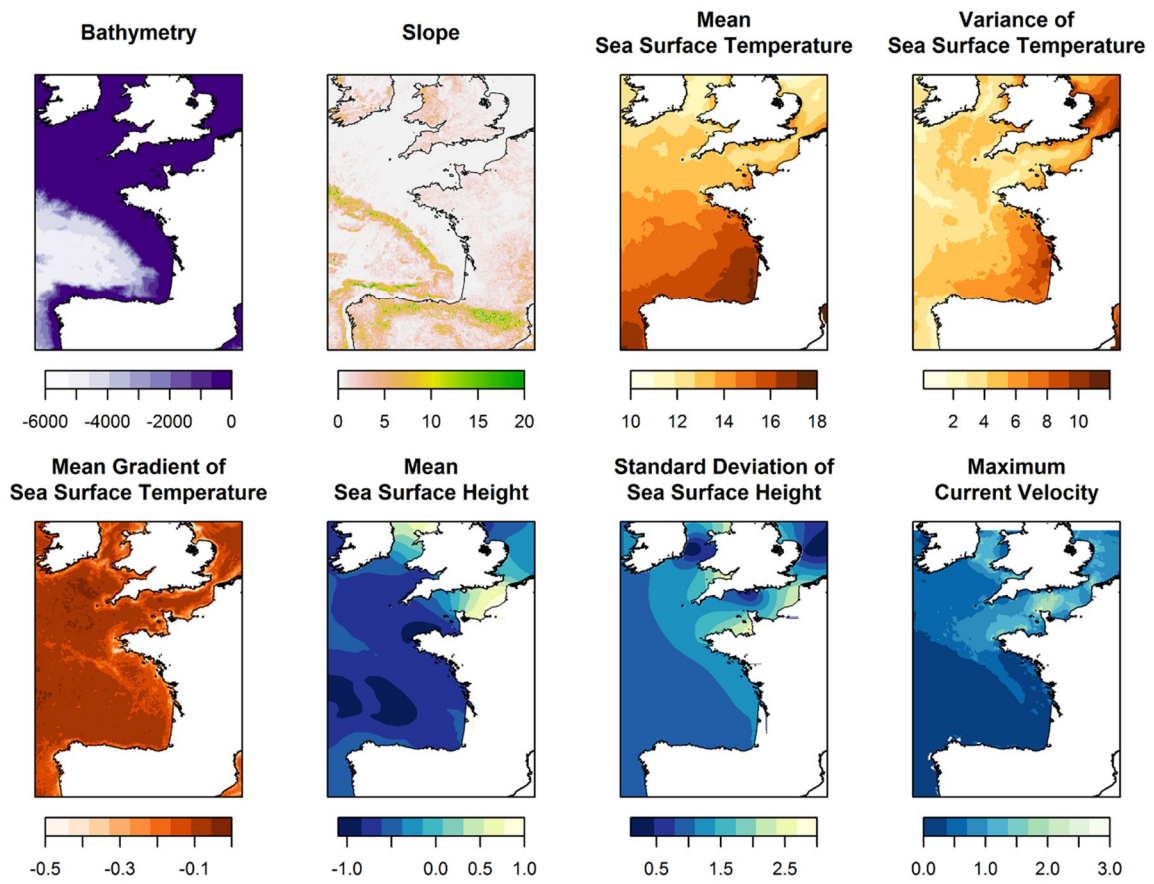
- Becker, E. A., Forney, K. A., Ferguson, M. C., Foley, D. G., Smith, R. C., Barlow, J., & Redfern, J. V., 2010. Comparing California current cetacean-habitat models developed using in situ and remotely sensed sea surface temperature data. *Marine Ecology Progress Series*, 413, pp.163–183.
- Brotons, L., Thuiller, W., Araújo, M. B., & Hirzel, A. H., 2004. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, 4, pp.437–448.
- Buckland, S.T., Anderson, D.R., Burnham, H.P., Laake, J.L., Borchers, D.L. & Thomas, L., 2001. *Introduction to Distance Sampling: Estimating Abundance of Biological Populations*. Oxford University Press, Oxford.
- Cañadas, A., Sagarminaga, R., De Stephanis, R., Urquiola, E., Hammond, P.S., 2005. Habitat preference modelling as a conservation tool: proposals for marine protected areas for cetaceans in southern Spanish waters. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 15, 495–521.
- Cañadas, A., Donovan, G. P., Desportes, G., & Borchers, D. L., 2009. A short review of the distribution of short beaked common dolphins (*Delphinus delphis*) in the central and eastern North Atlantic with an abundance estimate for part of this area. *NAMMCO Scientific Publications*, 7, pp.201–220.
- Deng, D., & Paul, S. R., 2005. Score tests for zero-inflation and over-dispersion in generalized linear models. *Statistica Sinica*, 257-276.
- Dobbie, M. J., & Welsh, A. H., 2001. Theory & Methods: Modelling Correlated Zero-inflated Count Data. *Australian & New Zealand Journal of Statistics*, 43(4), 431-444.
- Dunn, P. K., & Smyth, G. K. 2005. Series evaluation of Tweedie exponential dispersion model densities. *Statistics and Computing*, 15(4), 267-280.
- 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(1), pp.43–57.
- Elith, J. et al., 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29, pp.129–151.
- Elith, J. & Graham, C.H., 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography*, 32(1), pp.66–77.
- 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.
- Ferguson, M. C., Barlow, J., Reilly, S. B., & Gerrodette, T., 2006. Predicting Cuvier's (*Ziphius cavirostris*) and Mesoplodon beaked whale population density from habitat characteristics in the eastern tropical Pacific Ocean. *Journal of Cetacean Research and Management*, 7(3), pp.287–299.
- Gormley, A. M., Forsyth, D. M., Griffioen, P., Lindeman, M., Ramsey, D. S., Scroggie, M. P., & Woodford, L., 2011. Using presence-only and presence-absence data to estimate the current and potential distributions of established invasive species. *Journal of Applied Ecology*, 48(1), pp.25–34.
- Guisan, A. & Thuiller, W., 2005. Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, 8(9), pp.993–1009.
- Guisan, A. & Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*, 135(2-3), pp.147–186.
- Hammond, P.S. et al., 2013. Cetacean abundance and distribution in European Atlantic shelf waters to inform conservation and management. *Biological Conservation*, 164, pp.107–122.
- Hastie, T., & Tibshirani, R., 1986. Generalized Additive Models. *Statistical Science*, 3, pp.297-313.

- Hegel, T. M., Cushman, S. A., Evans, J., & Huettmann, F., 2010. Current state of the art for statistical modelling of species distributions. In *Spatial complexity, informatics, and wildlife conservation* (pp. 273-311). Springer Japan.
- Hirzel, A. H., Hausser, J., Chessel, D., & Perrin, N., 2002. Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? *Ecology*, 83(7), pp.2027–2036.
- King, G. & Zeng, L., 2007. When Can History Be Our Guide? The pitfalls of counterfactual inference. *International Studies Quarterly*, 51 (1), 183-210.
- Kleiber, C. & Zeileis, A., 2016. Visualizing Count Data Regressions Using Rootograms. *The American Statistician*, (accepted) pp.1–25.
- Lambert, D., 1992. Zero-Inflated Poisson Regression, With an Application To Defects in Manufacturing. *Technometrics*, 34(1), pp.1–14.
- Lambert, C., Virgili, A., Pettex, E., Delavenne, J., Toison, V., Blanck, A., Ridoux, V. (in review) Habitat modelling predictions highlight seasonal relevance of Marine Protected Areas for marine megafauna. *Deep Sea Research II*, Special Issue "European Marine Megafauna".
- Lambert, C., Pettex, E., Dorémus, G., Laran, S., Stephan, E., Van Canneyt, O., Ridoux, V. (in press). How does ocean seasonality drive habitat preferences of highly mobile top predators? Part II: the eastern North-Atlantic. *Deep-Sea Research II*, Special Issue "European Marine Megafauna".
- Laran, S., Authier, M., Blanck, A., Dorémus, G., Falchetto, H., Monestiez, P., Pettex, E., Stephan, E., Van Canneyt, O., Ridoux, V. (in review). Using large scale survey to investigate seasonal variations in seabird distribution and abundance. Part II: the Bay of Biscay and the English Channel, *Deep Sea Research Part II*, Special Issue "European Marine Megafauna".
- Lindén, A. & Mantyniemi, S., 2011. Using the negative binomial distribution to model overdispersion in ecological data. *Ecology*, 92(7), pp.1414–1421.
- MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J., & Langtimm, C. A., 2002. Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8), pp.2248-2255.
- MacLeod, C.D., Brereton, T. & Martin, C., 2009. Changes in the occurrence of common dolphins, striped dolphins and harbour porpoises in the English Channel and Bay of Biscay. *Journal of the Marine Biological Association of the United Kingdom*, 89(05), p.1059.
- Mannocci, L., Monestiez, P., Spitz, J., & Ridoux, V., 2015. Extrapolating cetacean densities beyond surveyed regions: habitat-based predictions in the circumtropical belt. *Journal of Biogeography*, 42, pp.1267–1280.
- Mannocci, L., Catalogna, M., et al., 2014a. Predicting cetacean and seabird habitats across a productivity gradient in the South Pacific gyre. *Progress in Oceanography*, 120, pp.383–398.
- Mannocci, L., Laran, S., et al., 2014b. Predicting top predator habitats in the Southwest Indian Ocean. *Ecography*, 37(3), pp.261–278.
- Martin, T.G. et al., 2005. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology Letters*, 8(11), pp.1235–1246.
- Merow, C., Smith, M.J. & Silander, J.A., 2013. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, 36(10), pp.1058–1069.
- Minami, M., Lennert-Cody, C. E., Gao, W., & Roman-Verdesoto, M., 2007. Modeling shark bycatch: the zero-inflated negative binomial regression model with smoothing. *Fisheries Research*, 84(2), pp.210–221.

- Monk, J. et al., 2010. Habitat suitability for marine fishes using presence-only modelling and multibeam sonar. *Marine Ecology Progress Series*, 420, pp.157–174.
- Murphy, S., Pinn, E H.; Jepson, P.D., 2013. Review of New Information on Other Matters Relevant for Small Cetacean Conservation Population Size, Distribution, Structure and Causes of Any Changes Marine megavertebrates adrift: a framework for the interpretation of stranding data in a monitoring p. *Oceanography and Marine Biology*, 51, pp.193–280.
- Pearce, J. & Ferrier, S., 2000. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, 133(3), pp.225–245.
- Pearce, J.L. & Boyce, M.S., 2006. Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, 43, pp.405–412.
- Phillips, S.J., Anderson, R.P. & Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190, pp.231–259.
- Phillips, S.J. & Dudík, M., 2008. Modeling of species distribution with Maxent: new extensions and a comprehensive evaluation. *Ecography*, 31, pp.161–175.
- Phillips, S.J., Dudík, M. & Schapire, R.E., 2004. A Maximum Entropy Approach to Species Distribution Modeling. Twenty-first international conference on Machine learning - ICML '04, p.83
- Previmer. (2014). Previmer - Observation et prévisions côtières. Catalogue version 2.1.
- R Core Team., 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from <http://www.R-project.org/>.all
- Redfern, J. V et al., 2006. Techniques for cetacean - habitat modeling. *Marine Ecology Progress Series*, 310, pp.271–295.
- Richards, S.A., 2008. Dealing with overdispersed count data in applied ecology. *Journal of Applied Ecology*, 45(1), pp.218–227.
- Ridout, M., Demetrio, C.G. & Hinde, J., 1998. Models for count data with many zeros. *International Biometric Conference*, pp.1–13.
- Segurado, P., Araujo, M.B. & Arau, M.B., 2004. An evaluation of methods for modelling species distributions. *Journal of Biogeography*, 31(10), pp.1555–1568.
- Syphard, A.D. & Franklin, J., 2009. Differences in spatial predictions among species distribution modeling methods vary with species traits and environmental predictors. *Ecography*, 32(6), pp.907–918.
- Tsoar, A., Allouche, O., Steinitz, O., Rotem, D., & Kadmon, R., 2007. A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions*, 13(4), pp.397–405.
- Vilchis, L.I., Ballance, L.T. & Fiedler, P.C., 2006. Pelagic habitat of seabirds in the eastern tropical Pacific: Effects of foraging ecology on habitat selection. *Marine Ecology Progress Series*, 315, pp.279–292.
- Virgili, A., Lambert, C., Pettex, E., Dorémus, G., Van Canneyt, O., Ridoux, V. (in review). Predicting seasonal variations in coastal seabird habitats in the English Channel and the Bay of Biscay. *Deep Sea Research II*, Special Issue “European Marine Megafauna”.
- Warton, D.I., 2005. Many zeros does not mean zero inflation: Comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmetrics*, 16(3), pp.275–289.

- Wenger, S.J. & Freeman, M.C., 2014. Estimating Species Occurrence, Abundance, and Detection Probability Using Zero-inflated Distributions. *Ecology*, 89(10), pp.2953–2959.
- Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., & Guisan, A., 2008. Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, 14(5), pp.763–773.
- Wood, S. 2006. Generalized Additive models: An Introduction with R. Chapman & Hall/CRC. 422 p.
- Wood, S. 2013. mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation. Retrieved 7 July 2014, from <http://cran.r-project.org/web/packages/mgcv/index.html>
- Yackulic C.B., Chandler R., Zipkin E.F., Royle A., Nichols J.D., Campbell Grant E.H. & Veran S. 2013. Presence-only modelling using MAXENT: when can we trust the inferences? *Methods in Ecology and Evolution*, 4, pp.236–243.
- Zaniewski, A. E., Lehmann, A., & Overton, J. M., 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, 157(2-3), pp.261–280.
- Zeileis, A., Kleiber, C. & Jackman, S., 2007. Regression models for count data in R. Research Report Series / Department of Statistics and Mathematics, 53.

## Appendix A.1. Maps of average covariates over the entire survey.





**Appendix A.2.** Some key concepts about the models used in the study.

- Negative Binomial distribution

The Negative Binomial distribution is extended from the Poisson regression and is defined by two parameters, the arithmetic mean and an exponent  $k$ . By modulating this exponent  $k$ , this distribution can be adapted to over-dispersed data (Bliss & Fisher 2016). If the response variable  $Y$  obeys to a Negative Binomial distribution, the variance  $V(Y)$  and the mean  $E(Y)$  are related by the relationship  $V(Y) = E(Y) + kE(Y)^2$  (Ver Hoef & Boveng 2007).

In this study, we wanted to fit a GAM with a Negative Binomial distribution so we used the R package `mgcv` (Wood 2006; 2013) and the `gam` function specifying the “Negative Binomial” family. Besides, we used the `nb` function to estimate the parameter  $k$  during the fitting.

- Tweedie distribution

The Tweedie distribution is useful to model continuous positive data because, compared to the Poisson distribution, it includes an additional parameter  $p$  which defines the model distribution. Indeed, if  $p=0$ , it is a normal distribution, if  $p=1$ , it is a Poisson distribution and if  $p=2$ , it is a Gamma distribution. Tweedie models can handle zero-inflated data (*i.e.* data with many zeros), because when  $1 < p < 2$ , they are a Poisson mixture of Gammas distributions (Arcuti et al. 2013). Besides, for a response variable  $Y$  that obeys a Tweedie distribution, the variance  $V(Y)$  and the mean  $E(Y)$  are related by the relationship  $V(Y) = \varphi E(Y)^p$  where  $\varphi$  represents the dispersion parameter (Dunn & Smyth 2005).

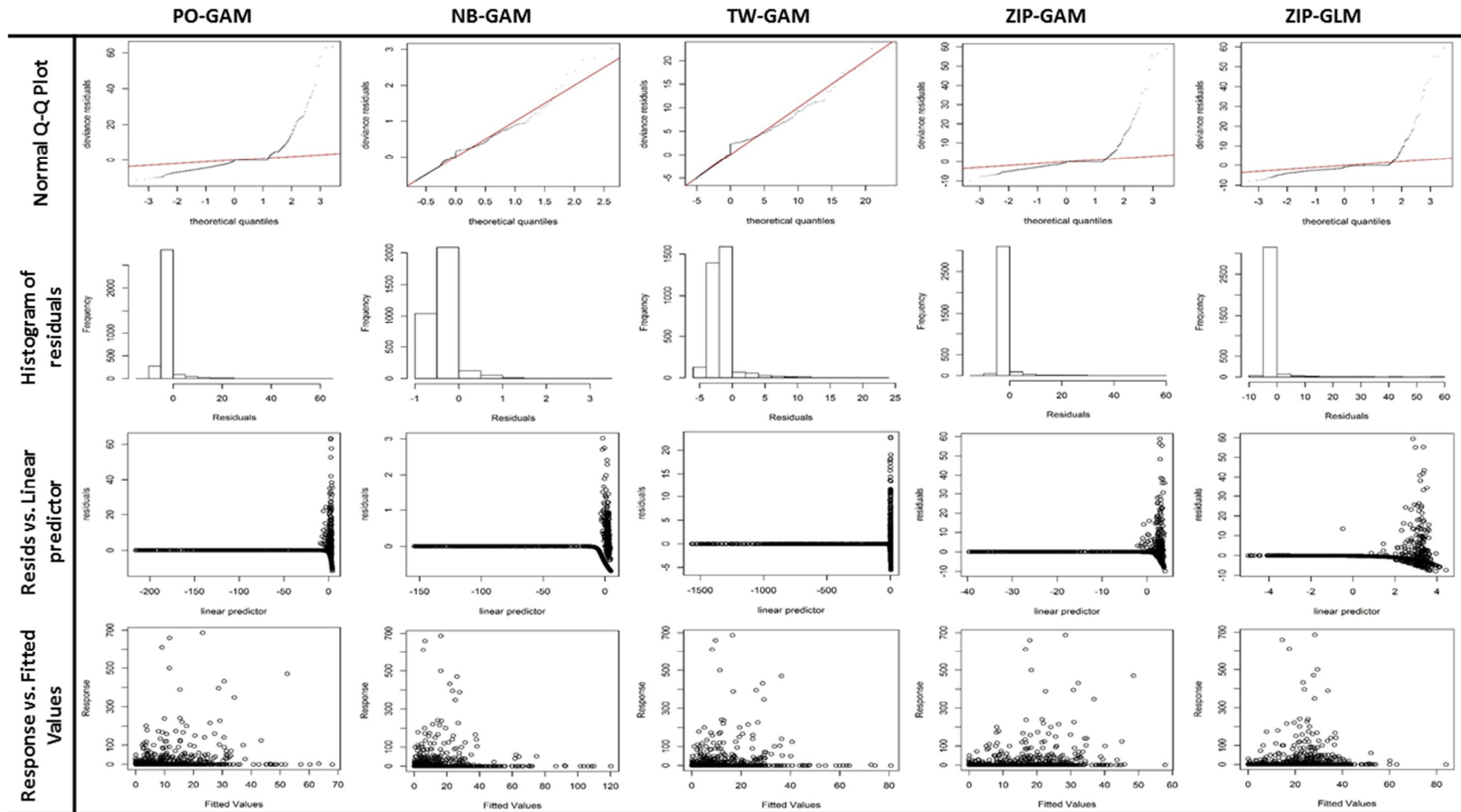
To fit a GAM with a Tweedie distribution, we used the R package `mgcv` (Wood 2006; 2013) and the `gam` function specifying the “Tweedie” family. Besides, as we ignored the value of the parameter  $p$ , we used the `tw` function which estimates this parameter during the fitting.

- Zero-inflated Poisson distribution

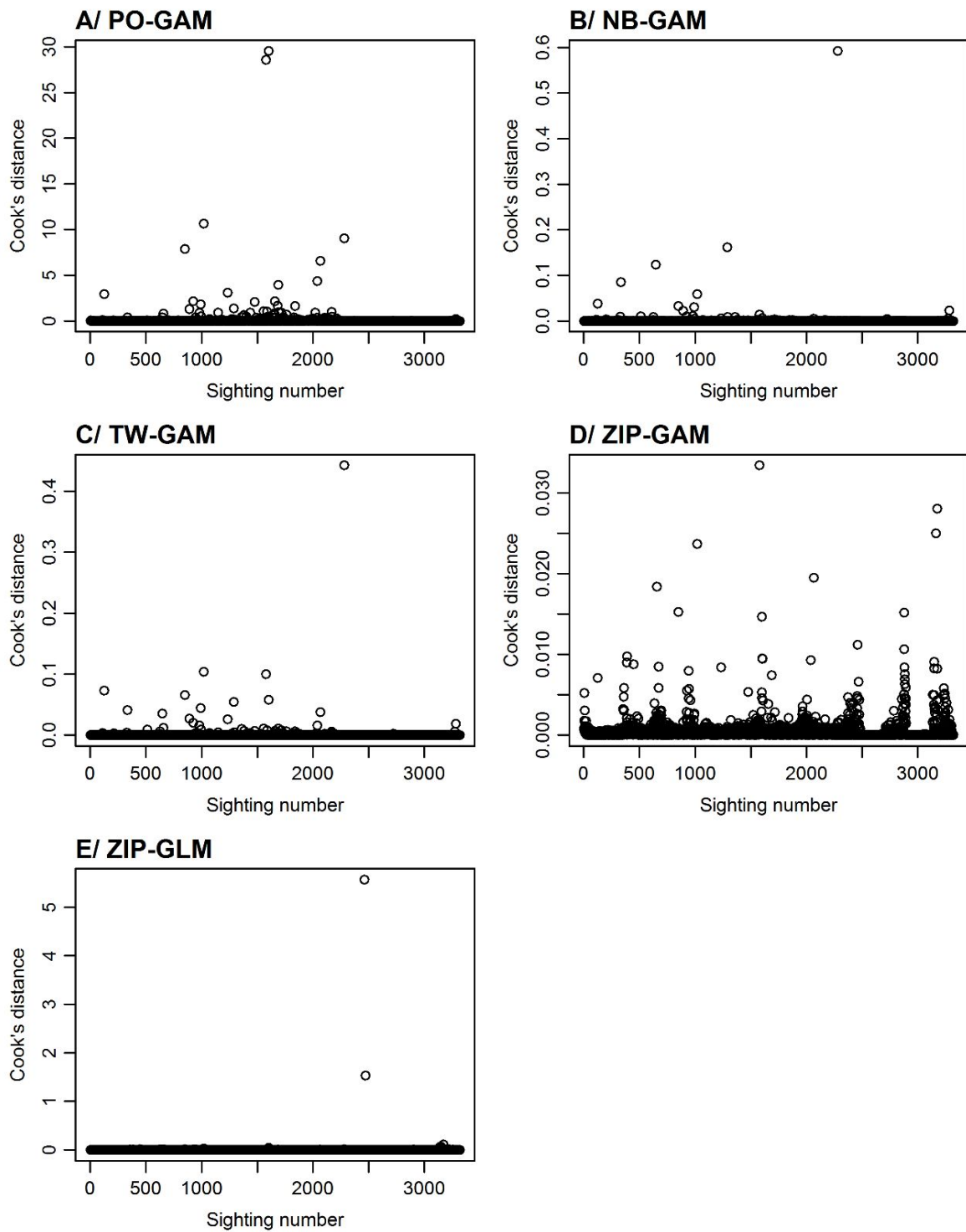
Zero-inflated Poisson distribution is used to model count data with extra zero counts by modelling independently the count values, with a Poisson distribution (Zeileis et al. 2007), and the excess of zeros (Lambert 1992). Thus, the ZIP regression is divided into two parts in which the species probability of presence and, given the presence, the species abundance are modelled sequentially (Ridout et al. 1998; Wenger & Freeman 2014).

In the study, we tested the ZIP distribution with a GAM which showed nonlinear relationships between the response variable and the environmental predictors. To fit the model, we used the `mgcv` package and the `gam` function but we specified the ZIP family and we used the `ziP` function to estimate the  $\vartheta$  parameter. This parameter includes two parameters which control the slope and the intercept of the zero model (Wood 2006; 2013). To fit smooth functions for the GAM we introduced a  $k$  parameter which worth 4, for 4 degrees of freedom.

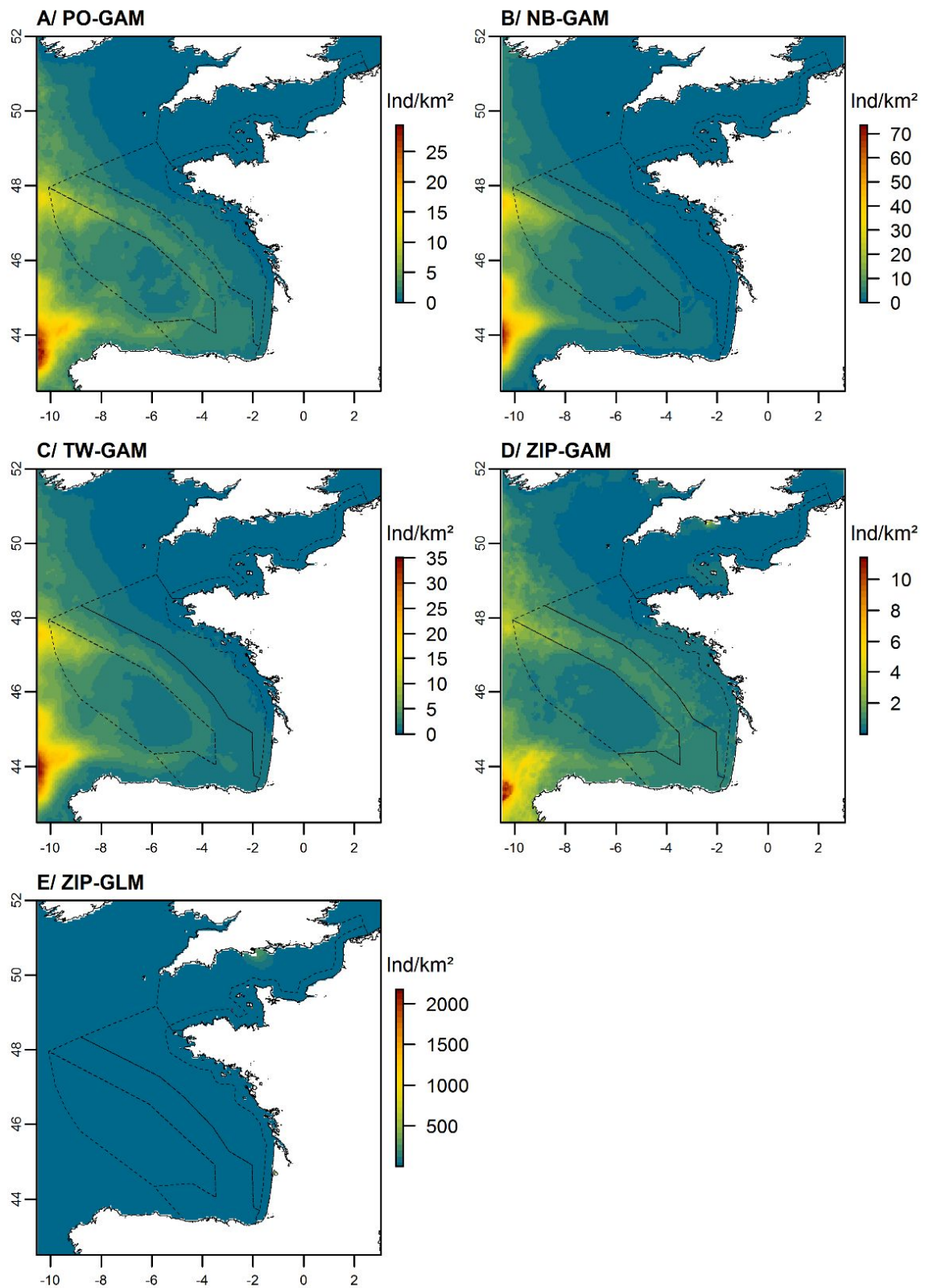
Appendix A.3. Residuals of the models obtained for each fitted model.



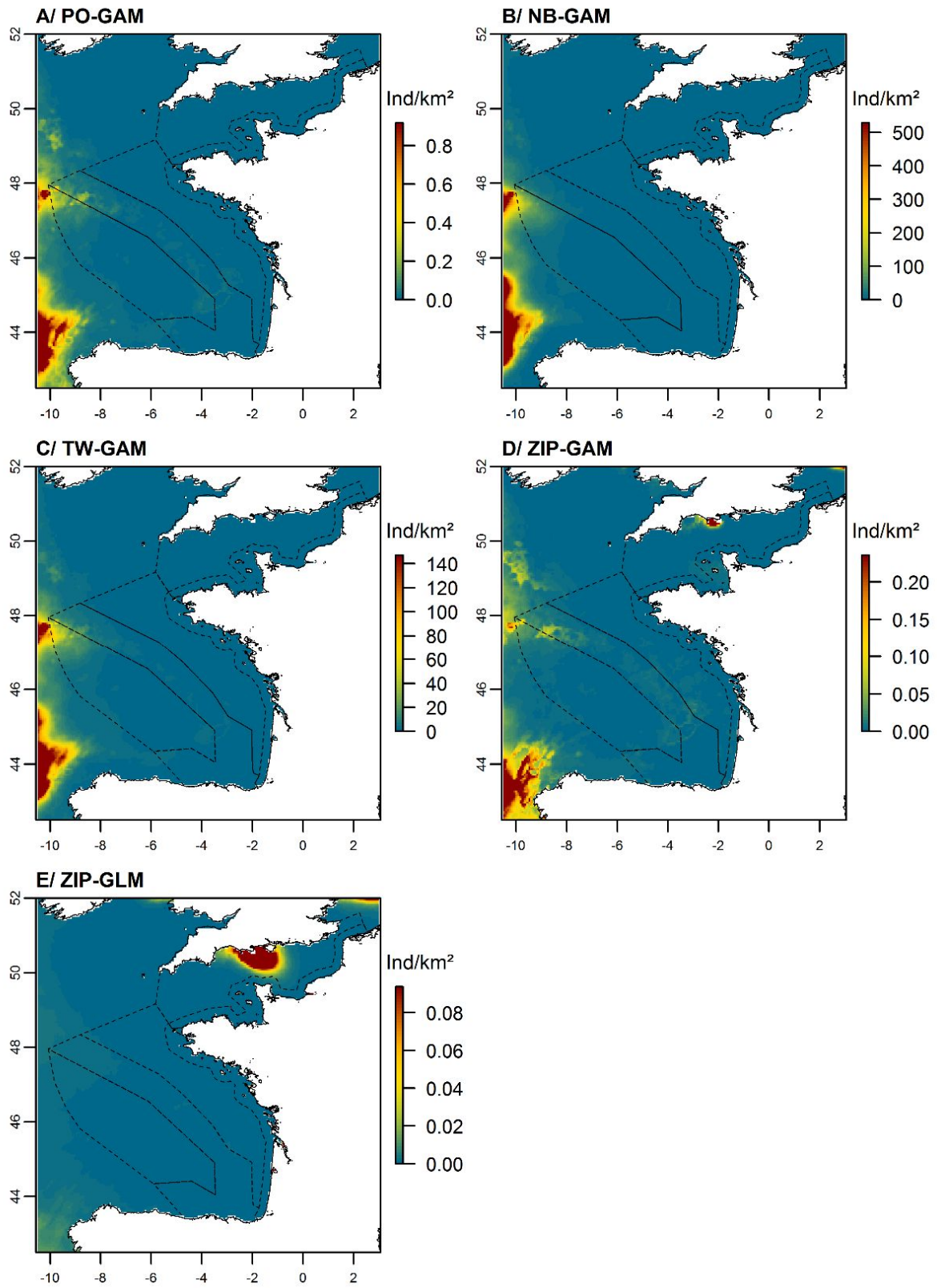
**Appendix A.4.** Cook's distances associated with each fitted model. A distance higher than 1 indicates a high influence of the sighting.



**Appendix A.5.** Predicted distributions of small delphinids in individuals·km<sup>-2</sup> (Ind/km<sup>2</sup>) for each presence-absence model in the Bay of Biscay and the English Channel. Contrary to Fig. A.3, the scale was not constrained to the one of the PO-GAM. Dotted lines represented the survey area.

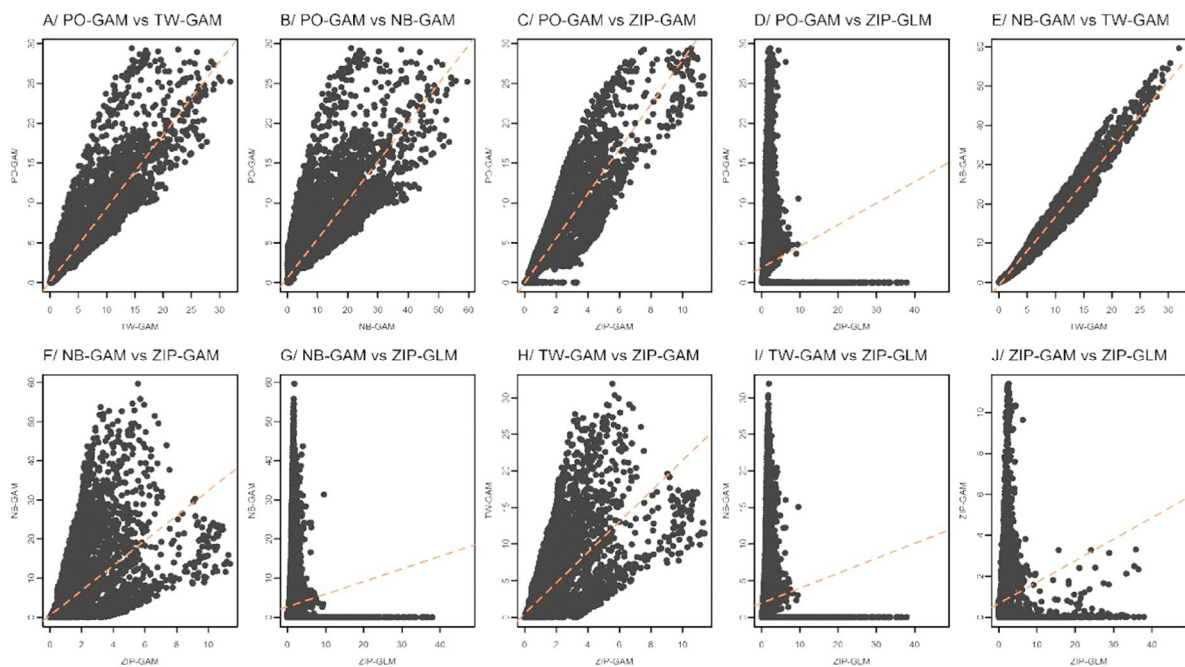


**Appendix A.6.** Uncertainty maps represented the standard error (in individuals- $\text{km}^{-2}$ ) associated with the predictive relative density of small delphinids groups. Dotted lines represent the survey area.



**Appendix A.7.** Predictions of the baseline model versus predictions of each “best” generic model. To assess the effect of the selection of the relevant explanatory variables by the models, we compared the predictions of the baseline model to the predictions of each “best” model (Table at the top). The comparison was done with simple linear regressions (Figure in the middle); if two models make similar prediction, then we expect a slope of 1 and an intercept of 0 when running a linear regression between the two predictions (Table at the bottom). Predictions of all the “best” GAMs were similar, but the predictions of the zero-inflated GLM greatly differed (with slope between 0.1 and 0.32 and intercepts between 0.67 and 2.72).

	Selected variables of the “best” models				D*
<b>PO-GAM</b>	SST <sub>mean</sub>	SST <sub>var</sub>	SSH <sub>mean</sub>	SSH <sub>sd</sub>	28.6 %
<b>NB-GAM</b>	SST <sub>mean</sub>	SST <sub>grad</sub>	SSH <sub>mean</sub>	SSH <sub>sd</sub>	39.4 %
<b>TW-GAM</b>	SST <sub>mean</sub>	SST <sub>grad</sub>	SSH <sub>mean</sub>	SSH <sub>sd</sub>	39.1 %
<b>ZIP-GAM</b>	SST <sub>mean</sub>	SST <sub>var</sub>	SSH <sub>mean</sub>	SSH <sub>sd</sub>	17.1 %
<b>ZIP-GLM</b>	Slope	SST <sub>grad</sub>	SSH <sub>mean</sub>	SSH <sub>sd</sub>	9.5 %



	Slope →				
Intercept ↓	PO-GAM	NB-GAM	TW-GAM	ZIP-GAM	ZIP-GLM
<b>PO-GAM</b>	1	0.49	0.92	2.8	0.27
<b>NB-GAM</b>	0.64	1	1.73	3.17	0.32
<b>TW-GAM</b>	0.14	-0.67	1	2.09	0.2
<b>ZIP-GAM</b>	-0.054	0.57	0.52	1	0.1
<b>ZIP-GLM</b>	1.86	2.72	1.94	0.67	1



# Annex B

---

## HOW MANY SIGHTINGS TO MODEL RARE MARINE SPECIES DISTRIBUTIONS

---

Auriane Virgili, Matthieu Authier, Pascal Monestiez, Vincent Ridoux

In revision in *PLoS ONE*



# How many sightings to model rare marine species distributions

Auriane VIRGILI <sup>1\*</sup>†, Matthieu AUTHIER <sup>2&</sup>, Pascal MONESTIEZ <sup>1,3&</sup>, Vincent RIDOUX <sup>1,2&</sup>

<sup>1</sup> Centre d'Etudes Biologiques de Chizé - La Rochelle, UMR 7372 CNRS - Université de La Rochelle, Institut du Littoral et de l'Environnement, La Rochelle, France

<sup>2</sup> Observatoire PELAGIS, UMS 3462 CNRS - Université de La Rochelle, Systèmes d'Observation pour la Conservation des Mammifères et des Oiseaux Marins, La Rochelle, France

<sup>3</sup> BioSP, INRA, Avignon, France.

## Abstract

Despite large efforts, datasets with few sightings are often available for rare species of marine megafauna that typically live at low densities. This paucity makes modelling the habitat of these taxa particularly challenging. We tested the predictive performance of different types of species distribution models fitted to decreasing numbers of sightings. Generalised additive models (GAMs) with three different residual distributions and the presence only model MaxEnt were tested on two megafauna case studies differing in both the number of sightings and ecological niches. From a dolphin (277 sightings) and an auk (1,455 sightings) datasets, we simulated rarity with a sighting thinning protocol by random sampling (without replacement) of a decreasing fraction of sightings. Better prediction of the distribution of a rarely sighted species occupying a narrow habitat (auk dataset) was expected compared to the distribution of a rarely sighted species occupying a broad habitat (dolphin dataset). We used the original datasets to set up a baseline model and fitted additional models on fewer sightings but keeping effort constant. Model predictive performance was assessed with mean squared error and area under the curve. Predictions provided by the models fitted to the thinned-out datasets were better than a homogeneous spatial distribution down to a threshold of approximately 30 sightings for a GAM with a Tweedie distribution and approximately 130 sightings for the other models. Thinning the sighting data for the taxon with narrower habitats seemed to be less detrimental to model predictive performance than for the broader habitat taxon. To generate reliable habitat modelling predictions for rarely sighted marine predators, our results suggest (1) using GAMs with a Tweedie distribution with presence-absence data and (2) implementing, as a conservative empirical measure, at least 50 sightings in the models.

## B.1. INTRODUCTION

The rarity of a species can be described in many different ways depending on a combination of criteria such as the extent of its geographic range, the specificity of its habitat and its local abundance (Table B.1; [1,2]). According to these criteria, only species that are widely distributed, live in diversified habitats and are abundant, are considered common. Other species are defined as rare because they show a restricted range, a specific habitat, low abundance, or any combination of these criteria.

Many species are naturally rare, but others become rare as a result of man induced pressures; in any case a species rarity contributes to its vulnerability. Therefore, rare species often benefit of a variety of management, conservation or recovery plans to maintain or restore their populations and habitats [3]. Determining the abundance, distribution and habitat use of these species are generally key

elements of these plans [4], yet gathering enough high quality data (*e.g.* sighting and effort data) is often a challenge.

**Table B.2. The three characteristics that defined the rarity of a species: the habitat specificity, the abundance and the geographic range (from [1,2]).** Each cell defines a form of species rarity except for the top left cell, which characterises a common species.

		Habitat specificity			
		Non-specialist		Specialist	
Abundance	High	Common species	Abundant but localised population in several habitats	Abundant and widespread population in specific habitats	Abundant and localised population in specific habitats
	Low	Scarce and widespread population in several habitats	Scarce and localised population in several habitats	Scarce and widespread population in specific habitats	Scarce and localised population in specific habitats
		Large	Limited	Large	Limited
		Geographic range			

Rare species usually result in a low number of sightings per unit effort [2]. This scarcity of sighting data renders difficult to fit species distribution models (SDM) because the reliability predictions largely depends on the number of sightings on which the models are fitted [2,5,6]. Although some studies have addressed the use of models for rare species datasets [2,5,7], the reliability of the predictions produced by these models, and the uncertainty associated with these predictions remain pending issues. To address these issues, one option is to examine how the performance of a species distribution model changes when sighting data becomes scarcer.

The aim of the present study was to suggest an empirical rule-of-thumb for the minimum sighting number needed to provide reliable predictions for different types of SDMs. This number is expected to be lower for specialist species using a narrow habitat than for more generalist species. It may also vary with the type of residual distribution functions (that is, the likelihood) used when fitting SDM. We thus conducted a sighting thinning experiment using two large datasets (with respect to effort) of marine megafauna collected in the eastern North Atlantic Ocean: small delphinid and auk datasets. Small delphinids are a generalist taxon, and are present at depths between 50-5000 m. In contrast, auks represent a more specialised taxon, as they are present at depths between 10-150 m. Hence, thinning the number of sightings of small dolphins would generate datasets of a rare, non-specialist species living in a large geographic range, and the thinning the number of sightings of auks would simulate a rare, more specialized species living in a more restricted geographic range. These datasets represent two forms of rare species defined by Rabinowitz ([1]; Table B.1). By thinning real datasets, this approach aimed to help habitat modellers circumvent the difficulty associated with assessing the predictive capacity of models fitted to rare species datasets.

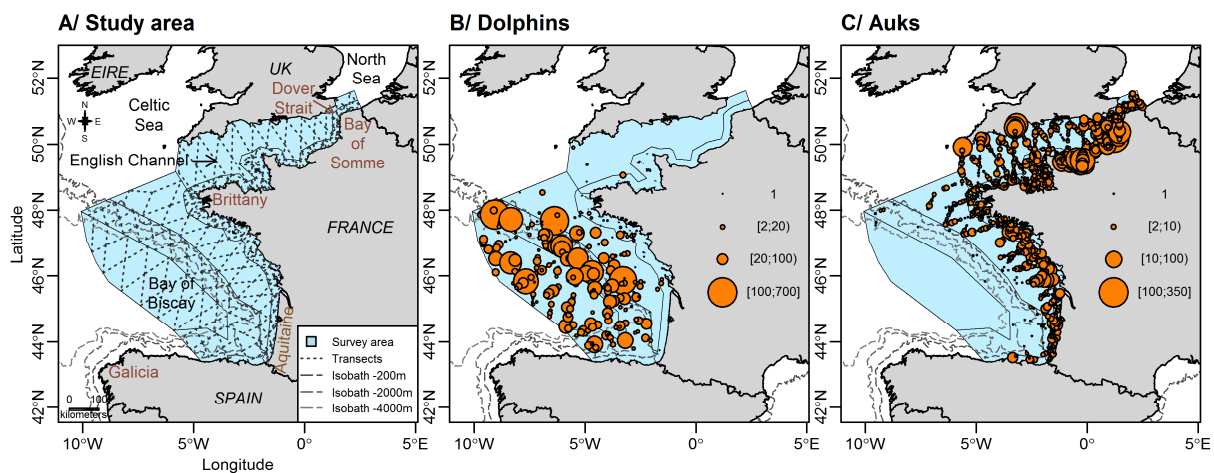
## B.2. MATERIALS AND METHODS

### B.2.1 Datasets

#### Data collection

Marine megafauna sightings were recorded during the two aerial SAMM surveys (Suivi Aérien de la Mégafaune Marine – Aerial Census for Marine Megafauna) conducted in the English Channel and the Bay of Biscay (Fig. B.1). Two taxa with abundant sightings (> 250) and contrasted distributions were selected. The first taxon was composed of small delphinids (hereafter called “dolphins”) including the common (*Delphinus delphis*) and striped (*Stenella coeruleoalba*) dolphins, both of which showing overall offshore distributions. The second taxon was composed of auks (hereafter called “auks”) and mostly consisted of the common guillemot (*Uria aalge*) and, to a much lower extent, the razorbill (*Alca torda*) and the Atlantic puffin (*Fratercula arctica*), all of which showing a more coastal distribution (Fig. B.1).

The surveys were conducted during the winter of 2011-2012 (from mid-November to early February; 28,068 km of transects) and the summer of 2012 (from mid-May to early August; 31,427 km of transects). A line-transect methodology was used to record all cetacean sightings [8], while seabird sightings were recorded using a strip-transect methodology [9]. In the line-transect methodology, the angle between the sighting and the track line was measured to determine the effective strip width (ESW; see the detection functions and estimated ESW in [10]) on each side of the plane. In the strip-transect methodology, the sightings were gathered from a 200-m strip on either side of the plane, and it was assumed that all animals within the strip were detected.



**Fig. B.1.** Study area (A) with dolphin (B) and auk (C) sightings recorded during the survey. The study area expands through the Bay of Biscay and the English Channel. The surveys were carried out along transects (dotted lines) following a zig-zag pattern across bathymetric strata. The sightings are classified by group sizes (1; 2-20; 20-100 and 100-700 individuals for dolphins and 1; 2-10; 10-100 and 100-350 individuals for auks), with each point representing one group of individuals.

In this study, we only used sighting data recorded in the summer for the dolphins and in the winter for the auks (Fig. B.1): the large number of sightings (>250) allowed for the sighting thinning approach to be implemented in a realistic and meaningful way. A total of 277 dolphin sightings accounting for 14,477 individuals and 1,455 auk sightings representing 16,658 individuals were recorded in good observation conditions (seastate <4 and medium to excellent observation conditions, as defined in [10]).

## Environmental predictors

Two categories of environmental predictors at a 10 km resolution were used to model the habitats of the two taxa (Table B.2). Static (or physiographic) predictors relate to the bathymetry and included depth and slope, whereas dynamic (or oceanographic) predictors describe the water masses and included the mean, variance and gradient of sea surface temperature (SST); the mean and standard deviation of sea surface height (SSH), and the maximum intensity of general currents (mostly referring to tidal currents in the study area; Appendix B.1). To avoid gaps in remotely sensed oceanographic variables, we used a 7-day resolution. All available data were averaged over the 6 days prior to each sampled day (details in [11, 12]).

**Table B.2. Environmental predictors used for habitat modelling.** A: Depth and slope were computed from the GEBCO-08 30 arc-second database (<http://www.gebco.net/>). B: Mean, variance and gradient of sea surface temperature (SST) were calculated from the ODYSSEA products (My Ocean project <http://www.myocean.eu/>). C: The MARS 3D model from Previmer ([13]; [www.previmer.org](http://www.previmer.org)) was used to compute mean and standard deviation of sea surface height (SSH). D: Daily maximum current intensity was computed from the MARS 2D model ([13]; [www.previmer.org](http://www.previmer.org)).

Environmental predictors	Sources	Effects on pelagic ecosystems of potential interest to top predators
<b>Physiographic</b>		
Depth (m)	A	Shallow waters could be associated with high primary production
Slope (°)	A	Associated with currents, high slope induce prey aggregation and/or primary production increasing
<b>Oceanographic</b>		
Mean of SST (°C)	B	Variability over time and horizontal gradients of SST reveal front locations, potentially associated to prey aggregations
Variance of SST (°C)	B	
Mean gradient of SST (°C)	B	
Mean of SSH (m)	C	High SSH is associated with high mesoscale activity and prey aggregation and/or primary production increase
Standard deviation of SSH (m)	C	
Daily maximum intensity of the currents (m.s <sup>-1</sup> )	D	High currents induce water mixing and prey aggregation

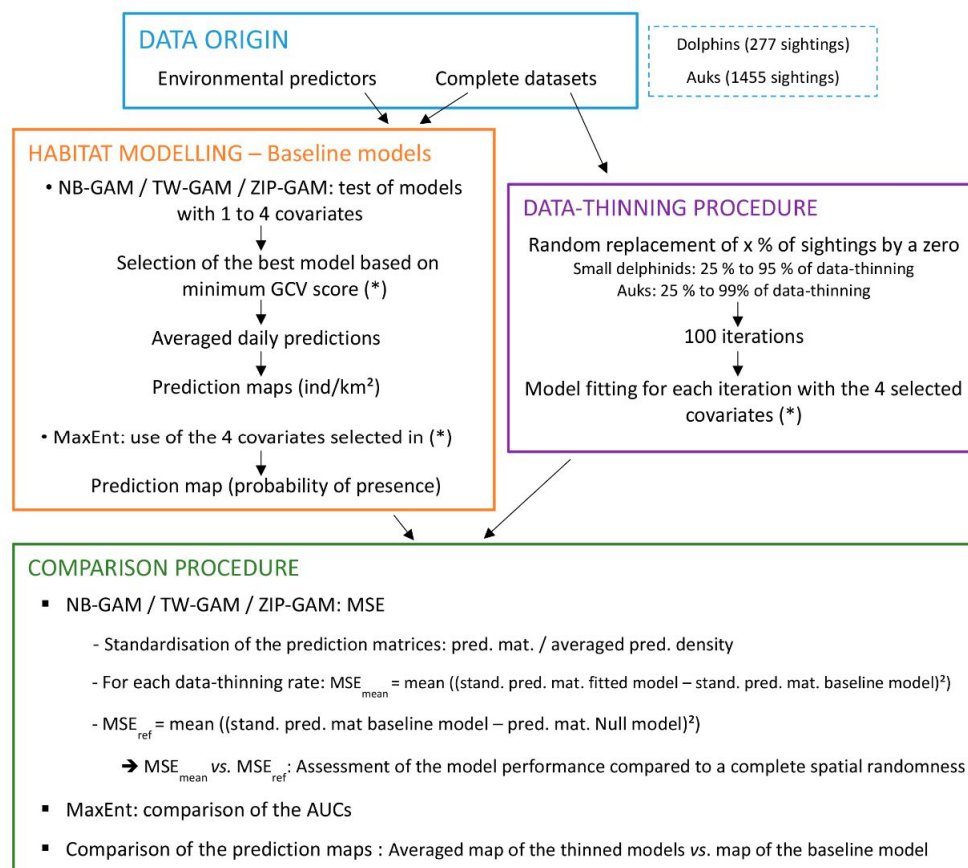
### B.2.2 Statistical analyses

#### Analytical strategy

We tested the predictive capacity of various SDMs fitted on rarely sighted species datasets (Fig. B.2). Two categories of SDMs can be used to predict a species distribution and model its habitats: presence-absence and presence-only models [14]. By establishing the functional relationships between sightings and environmental conditions, presence-absence models (*e.g.* Generalised Linear Models, GLMs, or Generalised Additive Models, GAMs) can predict areas of high species occurrence [14-16]. In contrast, with presence-only models (*e.g.* Ecological Niche Factor Analysis, ENFA, or Maximum Entropy Modelling, MaxEnt), only sites with environmental conditions similar to those of the sites where the taxon was recorded can be identified [17,18]. Presence-only models are the default option when data on absence (effort data) are not available [19], but the accuracy of presence-only model outputs largely relies on the representativeness of the sampled habitats [20]. Because of its ease-of-use, MaxEnt model

is widely used by managers and environmental agencies to help prioritising conservation areas [21-23]. Consequently, assessing the predictive performance of these presence-only models, compared to that of the presence-absence models, is relevant for rare species for which few sightings are typically available unless a considerable amount of effort is deployed.

Three presence-absence models and one presence-only model were tested. We used a GAM with a negative binomial distribution (NB-GAM), a GAM with a Tweedie distribution (TW-GAM), a GAM with a zero-inflated Poisson distribution (ZIP-GAM), and a MaxEnt model. These SDMs were first fitted to the original dolphin and auk datasets in order to select the 4 most important predictors for each taxon (hereafter referred to as 'baseline models'). These baseline models served as a reference to compare models fitted to the thinned-out datasets (hereafter referred to as the 'experimental models'). The original datasets were thinned of sightings by randomly removing 25-99% of the sightings. For each thinning-out level, the four SDMs were fitted with the same explanatory variables as in the baseline model. Finally, predictions from experimental models were compared to those of the baseline models to determine the minimum number of sightings to reliably predict rare species distribution. Although model performance is largely determined by its selected variables [24], we used the same specification for each experimental model in this study to assess how the results were affected by the sighting thinning alone. This choice reflects current practice in marine spatial planning in which the same SDM specification is frequently used by end-users (e.g. managers) but updated at a much lower frequency by researchers.



**Fig. B.2. Flowchart of the methods used in the study.** NB-GAM: generalised additive model with a negative binomial distribution; TW-GAM: generalised additive model with a Tweedie distribution; ZIP-GAM: generalised additive model with a zero-inflated Poisson distribution; MaxEnt: maximum entropy model; GCV: generalised cross-validation; ind: individuals; MSE: mean squared error; stand.: standardised; pred.: prediction; mat.: matrix; ref: reference; AUC: area under the curve.

### Baseline models

To fit GAMs, we used the ‘gam,’ ‘nb,’ ‘tw’ and ‘zip’ functions within the ‘mgcv’ package [25,26] (See Appendix B.2 for more details about the models). A log function linked the response variable to the additive predictors; the curve smoothing functions were restricted to three degrees of freedom [27]; finally an offset that considered the variation of effort per segment [28] was included and calculated as segment length multiplied by 2\*ESW (see Laran et al. [10]). After removing all combinations of variables with correlation coefficients higher than |0.7|, the models with combinations of 1 to 4 variables were tested [29,30], and the best models were selected, *i.e.* the models with the lowest generalised cross-validation score (GCV; [31]), which estimates the mean prediction error using a leave-one-out cross-validation process [32]. For both taxa, the selected variables for NB-GAM, TW-GAM and ZIP-GAM were identical so that it was straightforward to compare the different models.

For each fitted model, predicted densities (in individuals per km<sup>2</sup>) were mapped on a 0.05°x0.05° resolution grid. We computed the predictions for each day of the surveys and averaged the predictions over the entire survey period. To limit extrapolation, the covariates were constrained within the range of the covariate values used when fitting the models. Finally, we provided uncertainty maps by computing the variance around the predictions as the sum of the variance around the mean prediction and the mean of the daily variances. Then, the coefficient of variation was calculated as

$$CV = 100 \times \sqrt{(\text{variance over the survey period})/\text{mean over the survey period}}.$$

We also tested the effect of thinning-out on a presence-only model: Maxent (version 3.3.3, <http://www.cs.princeton.edu/~schapire/maxent/>; [22]). In this model, environmental relationships are estimated using the background samples of the environment instead of absence locations [33]. For both taxa, we first removed all absences from the input files used for the presence-absence models to obtain a file with only presence locations that would be compatible with the software. We used the four environmental variables determined by the selection procedure for the GAMs to allow for comparisons between the different models. Finally, we selected a “hinge” feature as a model parameter to generate models with smooth functions similar to GAMs with a default prevalence of 0.5 and a logistic output format to obtain a probability of presence of the species groups [33-35]. Table B.3 summarises the tested models and their characteristics.

### Thinning-out of the sightings

To generate datasets of rare species, we thinned the original auk and dolphin sightings at different rates. We aimed to obtain a decreasing number of sightings, simulating thereby an increasing rarity of the two taxa. In the dolphin dataset, we randomly replaced 25, 50, 75, 90, 92 and 95% of the sightings with zeros, and in the auk dataset, we randomly replaced 25, 50, 75, 90, 92, 95, 97 and 99% of the sightings with zeros (Table B.4; Appendices B.3 and B.4 show examples of thinning-out). For each thinning rate, sightings to be replaced with zero were randomly sampled without replacement, and the procedure was reiterated 100 times, hence producing 100 randomly thinned or experimental datasets for each thinning rate. This procedure simulates different levels of species rarity as observed under a constant sampling effort. Removing part of survey effort (*e.g.* whole transects) would not have generated a greater rarity of the species but only a lower sighting effort; and would have led to similar results of the baseline models because encounter rates had remained similar on average.

**Table B.3. Details of the models used in the study.** GAM: generalised additive model; GLM: generalised linear model; PO: Poisson; NB: negative binomial; TW: Tweedie; ZIP: zero-inflated Poisson; PA: presence-absence data; AIC: Akaike information criterion. \* R Core Team [34]

Models	Used names	Data	Settings and details
Generalised Additive Model with Negative Binomial distribution	NB-GAM	Over-dispersed PA	Used R-3.1.2*, package <b>mgcv</b> , function GAM, <i>Negative Binomial</i> distribution, log-link function, included an offset, 3 degrees of freedom for the smoothing curve functions
Generalised Additive Model with Tweedie distribution	TW-GAM	Over-dispersed PA	Used R-3.1.2*, package <b>mgcv</b> , function GAM, <i>Tweedie</i> distribution, log-link function, included an offset, 3 degrees of freedom for the smoothing curve functions
Generalised Additive Model with Zero-Inflated Poisson distribution	ZIP-GAM	Zero-inflated PA	Used R-3.1.2*, package <b>mgcv</b> , function GAM, <i>ZIP</i> distribution, log-link function, included an offset, 3 degrees of freedom for the smoothing curve functions
Maximum Entropy Modelling	MaxEnt	Presence-only	Used MaxEnt software version 3.3.3, <i>hinge</i> feature, default prevalence of 0.5, logistic output format

**Table B.4. Number of sightings contained in the thinned or experimental datasets for each sighting thinning rate and each species group.**  $n_{\text{sigh}}$ : number of sightings;  $n_z$ : number of segments with a zero;  $\%_z$ : percentage of zeros; Original: initial (and complete) datasets. “–” indicates that the data thinning was not performed.

Species groups		Sighting thinning rates								
		Original	25%	50%	75%	90%	92%	95%	97%	99%
Dolphins	$n_{\text{sigh}}$	277	208	139	69	28	23	14	-	-
	$n_z$	3043	3112	3181	3250	3292	3297	3306	-	-
	$\%_z$	91.7	93.7	95.8	97.9	99.2	99.3	99.6	-	-
Auks	$n_{\text{sigh}}$	1455	1091	728	364	146	116	73	44	15
	$n_z$	2046	2409	2773	3137	3355	3384	3428	3457	3486
	$\%_z$	56	66	76	85.9	91.9	92.7	94	94.7	95.5

### Assessment of the predictive performance of the model

The baseline SDMs were selected using the minimum GCV score, and a leave-one-out cross-validation process was used to estimate mean prediction error and explained deviances [31,32]. However, for experimental models, we based the assessment of the predictive performance of the presence-absence models on two criteria: mean squared error (MSE; [37,38]) and maps of the predicted densities. The MSE directly compared the prediction matrices of the experimental models to the prediction matrix of the baseline model. Each cell of the matrices provides the densities predicted by the model over the entire prediction area. The MSE is given by  $MSE = \text{mean} \left( \sum (\hat{Y}_{\text{exp}} - \hat{Y}_{\text{baseline}})^2 \right)$

[37,38]. Here, " $\hat{Y}_{\text{exp}}$ " represents the prediction matrix of an experimental model, and " $\hat{Y}_{\text{baseline}}$ " represents the prediction matrix of the baseline model. For each type of model and thinning rate, we averaged the MSEs of all the experimental models to obtain an averaged MSE (called  $\text{MSE}_{\text{mean}}$ ). Then, we investigated whether the predictions provided by the models fitted to sighting thinned-out datasets were better than those from a homogeneous process. For this purpose, we compared the MSE of each fitted model and the  $\text{MSE}_{\text{mean}}$  to a reference threshold, called the  $\text{MSE}_{\text{ref}}$ , which was calculated as the MSE between the prediction matrix of the baseline model (NB-GAM, TW-GAM or ZIP-GAM) and the prediction matrix of a null NB-GAM, TW-GAM or ZIP-GAM (which described a homogeneous spatial distribution). We assumed that if the MSE was higher than the  $\text{MSE}_{\text{ref}}$ , it was more appropriate to consider a homogeneous spatial distribution rather than taking into account the predictions provided by the experimental model.

To assess the predictive performance of the MaxEnt models, we used the area under the receiver operating characteristic curve (AUC; [17]). AUC allows for the direct comparison of SDM predictive performance but can only be used on binary data. An AUC of 1 indicates a perfect discrimination between the sites where the species is present and absent, an AUC of 0.5 indicates a discrimination equivalent to a random distribution, and an AUC lower than 0.5 indicates that the model performance is worse than a random guess [17]. We compared the AUC of each fitted model and the  $\text{AUC}_{\text{mean}}$  (averaged over the 100 fitted models) to the AUC of the baseline model and used a threshold value of 0.5 to assess the performance of the experimental models.

Finally, we compared the prediction maps of the models fitted to the thinned datasets to the prediction maps of the baseline models in order to determine the lowest sample size that did not change predicted distribution patterns. For each model type and each thinning rate, we averaged the predictions over the 100 models fitted to the thinned datasets and produced averaged prediction maps that we compared to the prediction maps of the baseline models. We averaged the predictions over the 100 fitted models to ensure a uniform data deletion throughout the area. In practice, habitat modellers only have one real dataset (not 100); hence, we compared the MSE or AUC of each fitted model to the  $\text{MSE}_{\text{ref}}$  or  $\text{AUC}_{\text{ref}}$  to determine the proportion of the model that provided good predictions.

## B.3. RESULTS

### B.3.1 Model selection and predictions of the baseline models

#### Small delphinids

The explained deviances in the dolphin dataset were fairly high: 38.4% for the NB-GAM, 37.3% for the TW-GAM and 17.1% for the ZIP-GAM. Densities were best predicted from SST mean and variance and SSH mean and standard deviation (Fig. B.3). All models showed similar smooth functions with the highest densities of delphinids predicted at temperatures of approximately 16°C (variance close to 0°C) and low average altimetry (SSH, approximately -0.5 m, standard error approximately 0.5 m). Small delphinids were predicted to be distributed in offshore waters from the continental shelf to oceanic waters with higher densities along the slope and a peak north of Galicia (Fig. B.3). There was a strong match between sightings and model predictions (Figs. B.1 and B.3) with high predicted densities associated with low coefficients of variation (Appendix B.5). With an AUC of 0.822, the MaxEnt model correctly predicted the presence probabilities of delphinids. Similar to the other fitted models, the



highest presence probabilities were predicted along the slope of the Bay of Biscay and were evenly distributed elsewhere (Fig. B.3).

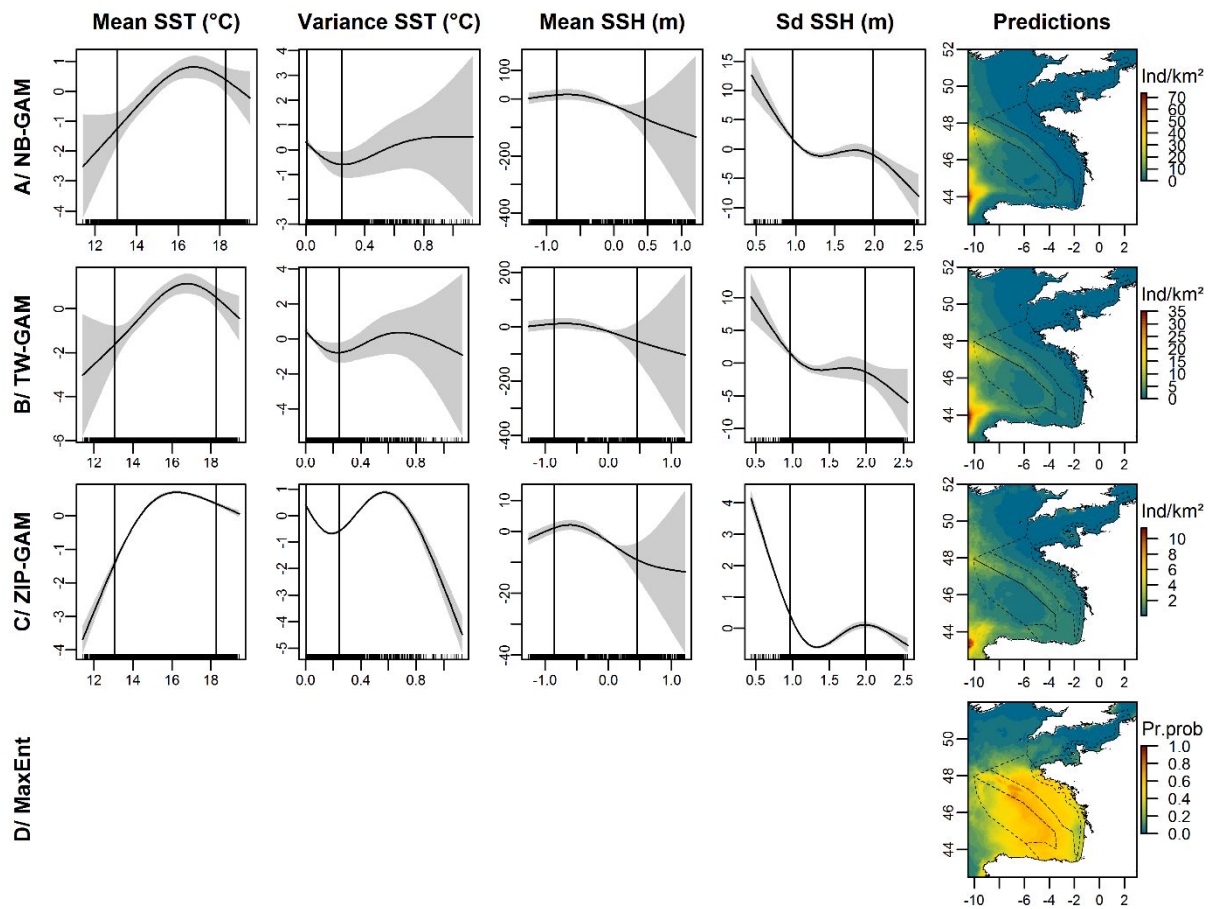
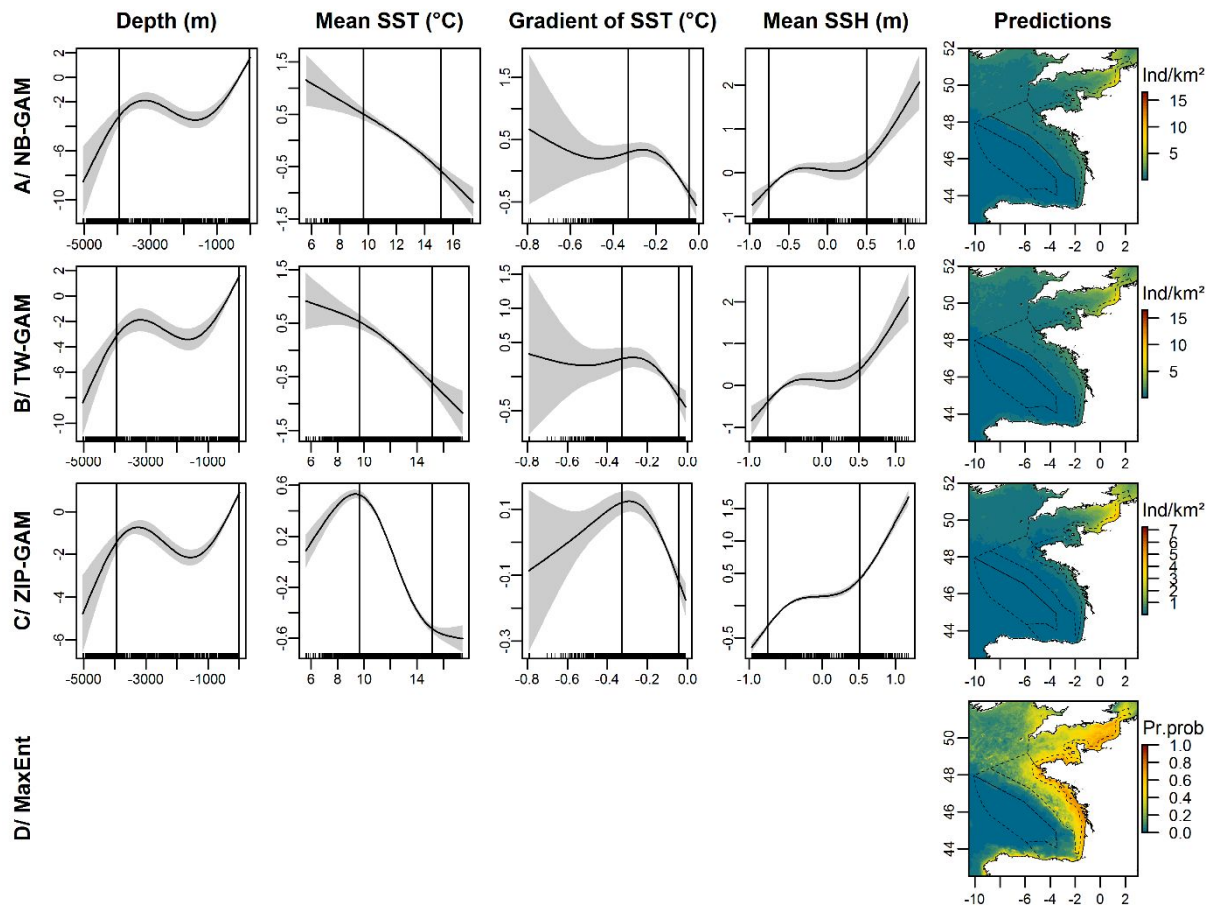


Fig. B.3. Forms of smooth functions for the selected covariates and predicted distribution of dolphins in individuals.km<sup>-2</sup> (Ind/km<sup>2</sup>) for each presence-absence model and in presence probabilities (Pr.prob) for the Maxent model. The solid line in each plot is the estimated smooth function, and the shaded regions represent the approximate 95% confidence intervals. The y-axis indicates the number of individuals on a log scale, and a zero indicates no effect of the covariate. The best model fits are between the vertical lines indicating the 10<sup>th</sup> and 90<sup>th</sup> quantiles of the data. The dotted lines represent the bathymetric strata of the survey area. The white areas on certain maps represent the absence of predictions beyond the range of covariates used in fitted models.

## Auks

The explained deviances in the auk dataset reached 44.9% for the NB-GAM, 40.9% for the TW-GAM and 33.6% for the ZIP-GAM. The variables selected by the three baseline models were depth, mean and gradient of SST and mean SSH (Fig. B.4). Greater auk densities were associated with colder and shallower waters, stronger gradients of temperature and higher positive altimetry. The predicted distribution ranged from the coast to the edge of the continental shelf and predicted densities were particularly high in the eastern English Channel (Fig. B.4). There was a good match between the sightings and the predictions of the model (Figs. B.1 and B.4) with high predicted densities associated with low coefficients of variation (Appendix B.5). The MaxEnt model, with an AUC of 0.842, generally predicted the same distribution as the other models with higher concentrations along the coast (Fig. B.4).



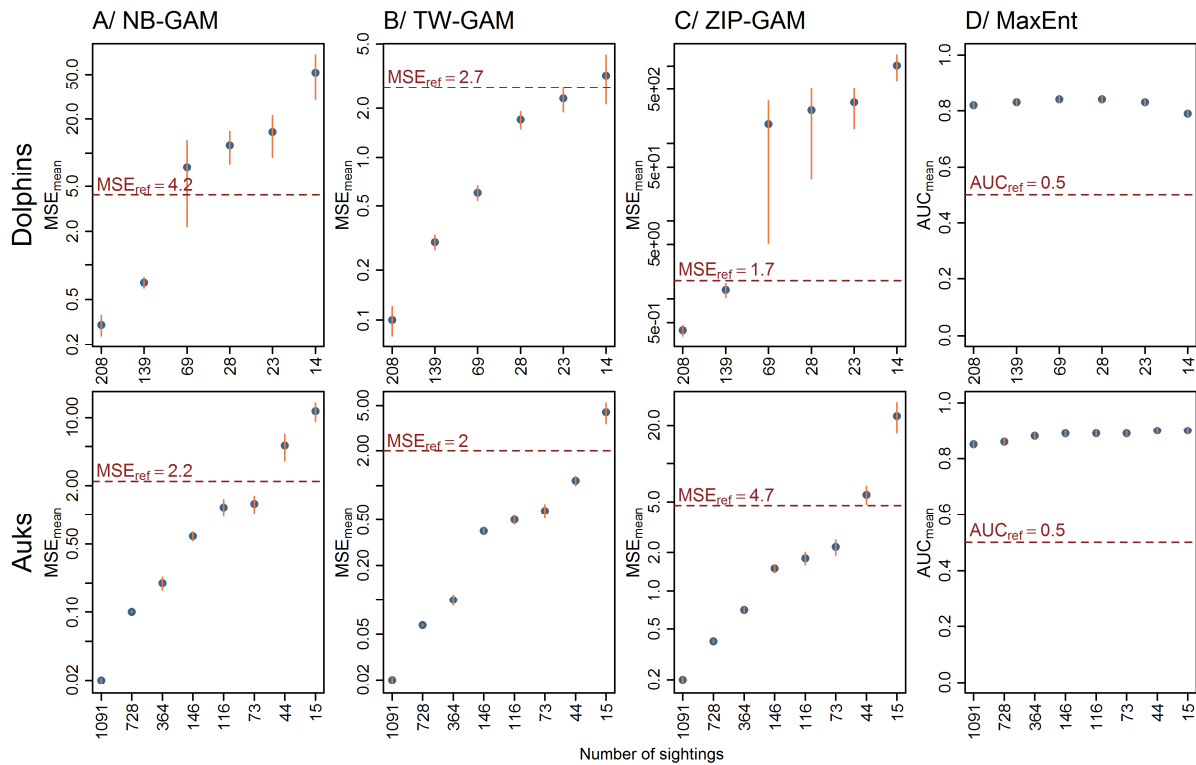
**Fig. B.4.** Forms of smooth functions for the selected covariates and predicted distribution of auks in individuals.km<sup>-2</sup> (Ind/km<sup>2</sup>) for each presence-absence model and in presence probabilities (Pr.prob) for the MaxEnt model. The solid line in each plot is the estimated smooth function, and the shaded regions represent the approximate 95% confidence intervals. The y-axis indicates the number of individuals on a log scale, and a zero indicates no effect of the covariate. The best model fits are between the vertical lines indicating the 10<sup>th</sup> and 90<sup>th</sup> quantiles of the data. The dotted lines represent the bathymetric strata of the survey area.

### B.3.2 Predictive performance of the experimental models

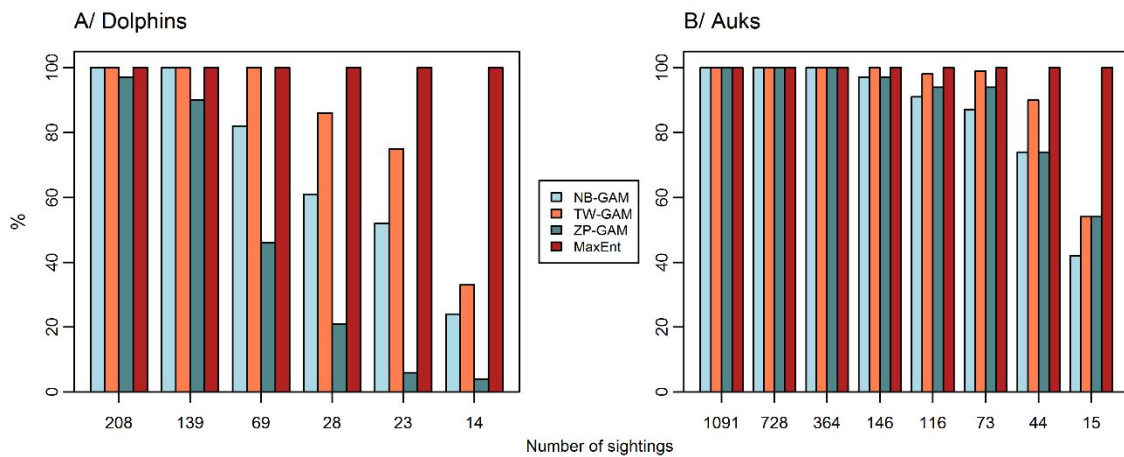
#### Small delphinids

As expected, a decrease in the number of sightings led to an increase in  $MSE_{mean}$  (Fig. B.5). Predictions with 208 sightings (the lowest thinning rate) were closer to those of the baseline models than the predictions with only 14 sightings (the highest thinning rate). The comparison of  $MSE_{mean}$  with  $MSE_{ref}$  (representing the MSE between the baseline predictions and the null models), suggested that for less than 139 sightings,  $MSE_{mean}$  values for NB-GAMs and ZIP-GAMs were higher than  $MSE_{ref}$ . In contrast,  $MSE_{mean}$  values for TW-GAMs were lower than  $MSE_{ref}$ , except for the most extreme thinning rate that yielded as few as 14 sightings. Consequently, below 139 sightings, it was better to predict a homogeneous spatial distribution rather than to use the predictions provided by the NB-GAMs and the ZIP-GAMs. For the TW-GAMs, this threshold was under 23 sightings. Furthermore, the number of experimental models in which the MSE was higher than the  $MSE_{ref}$  varied among model types (Fig. B.6). With a decrease in the number of sightings, the proportion of experimental models in which predictions were better than a homogeneous spatial distribution decreased ( $MSE < MSE_{ref}$ ; Fig. B.6). For example, with 23 sightings, only 51% NB-GAMs and 6% ZIP-GAMs predicted better than a homogeneous spatial distribution compared to 75% TW-GAMs. For MaxEnt,  $AUC_{mean}$  values of the experimental models were

high ( $>0.82$ ) and very similar and higher than the  $AUC_{ref}$ , which predicted a homogeneous distribution of the sites occupied by the species.



**Fig. B.5. Evaluation of the predictive performance of the models using MSE and AUC.**  $MSE_{mean}$ : mean squared error averaged over 100 models;  $AUC_{mean}$ : area under the curve averaged over 100 models; Ref: reference index (*i.e.* a random spatial distribution). A log scale is applied on the y-axis. The vertical bars on each point represent the standard error calculated from 100 models.

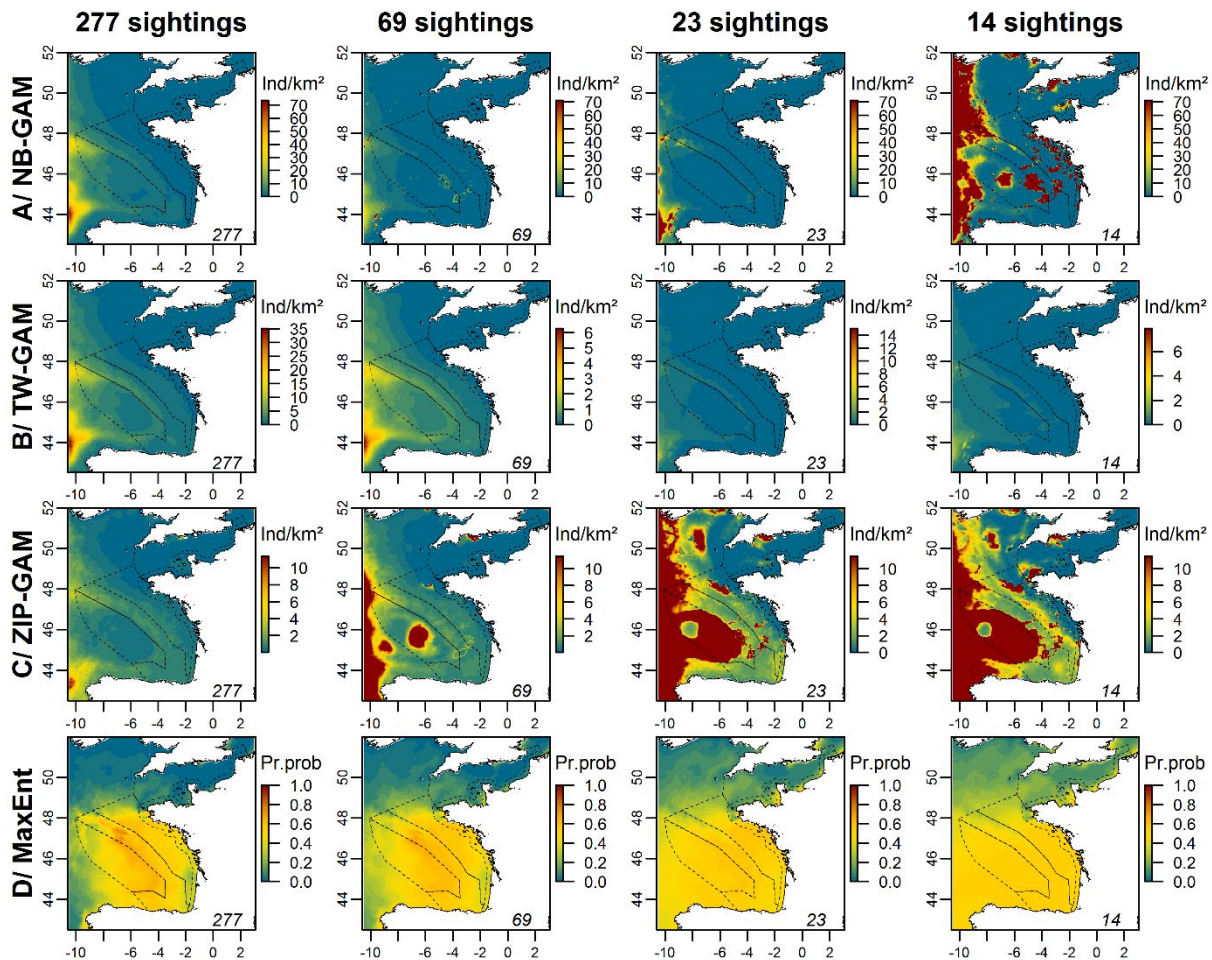


**Fig. B.6. Proportion of experimental models better than a random spatial distribution.** Each bar represents, the proportion of the experimental models out of the 100 fitted in which the MSE is lower than the  $MSE_{ref}$  for each number of sightings, *i.e.* the model that is better than a random spatial distribution. Each colour represents a different model type.

We noticed an important variation in the prediction maps among experimental models (Fig. B.7; Appendices B.6 and B.7). Despite a decrease in the number of sightings, the distribution patterns of the baseline models were maintained down to 139 sightings for NB-GAMs and ZIP-GAMs. Beyond this threshold, the pattern disappeared or became unrealistic. Predictions from TW-GAM were similar to the distribution pattern of the baseline model with as few as 28 sightings. Beyond this threshold, the

pattern started to fade out. When compared to the baseline, the highest densities predicted by NB-GAM, TW-GAM and ZIP-GAM were associated with the highest uncertainties (Appendices B.8 and B.9).

Presence probability predicted by MaxEnt model became more uniform in area when the number of sightings decreased, with high probability areas located along the slope, and low probability areas located near the Aquitaine coast gradually fading out (Fig. B.7).



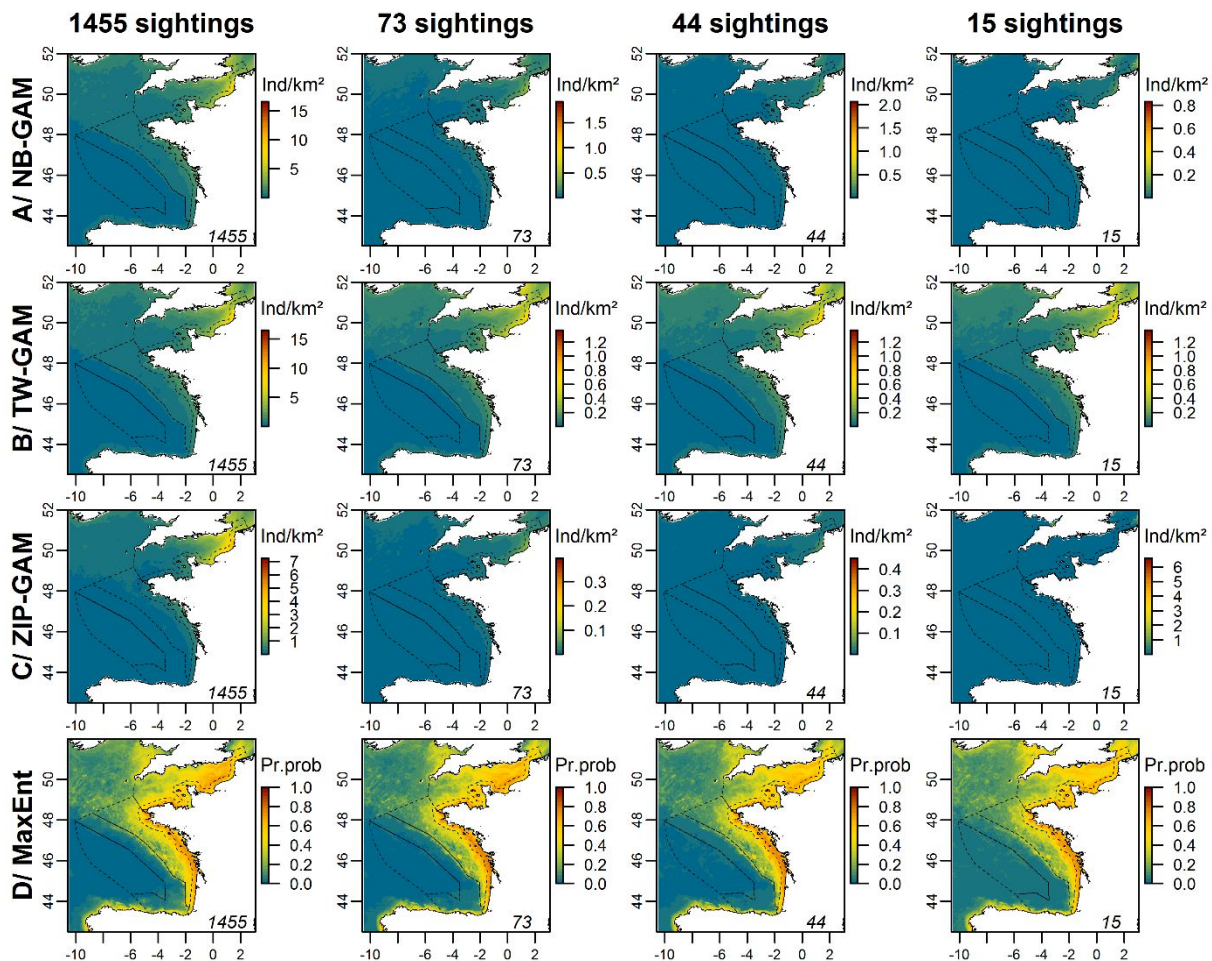
**Fig. B.7.** Prediction maps of dolphins averaged over 100 models fitted to thinned datasets for each type of model in the Bay of Biscay and the English Channel. The rows represent the different types of generic models, and the columns represent the number of sightings used to fit the models. The numbers in the right corner of each map represent the number of sightings used to fit the model. The scale is in individuals.km<sup>-2</sup> (Ind/km<sup>2</sup>) for the NB-GAM, the TW-GAM and the ZIP-GAM and in the probability of presence (Pr.prob) for MaxEnt. This figure only shows the results for which a change was observed compared with the other predictions. All maps are presented in Appendices B.6 and B.7. The dotted lines represent the bathymetric strata of the survey area.

## Auks

MSE<sub>mean</sub> values increased with decreasing numbers of sightings (Fig. B.5). As expected, predictions with 1,091 sightings (the lowest thinning level) were closer to those of the baseline model (1,455 sightings) than were the predictions with only 15 sightings (the highest thinning level). When the number of sightings was lower than 73, MSE<sub>mean</sub> values of NB-GAMs and ZIP-GAMs were higher than MSE<sub>ref</sub>, whereas MSE<sub>mean</sub> for TW-GAMs was higher than MSE<sub>ref</sub> with only 15 sightings. Consequently, with less than 73 sightings, the predictions provided by NB-GAMs and ZIP-GAMs were worse than a homogeneous spatial distribution. For TW-GAMs, this threshold was below 44 sightings. Similar to the results for dolphins, the number of models in which the MSE was higher than the MSE<sub>ref</sub> varied (Fig.

B.6). With 15 sightings, only 42% NB-GAMs compared to 54% TW-GAMs and ZIP-GAMs predicted better than a homogeneous spatial distribution. The  $AUC_{mean}$  values for the MaxEnt model were very high ( $>0.85$ ) and slightly increased with a decreasing number of sightings. Overall, the  $AUC_{mean}$  values were higher than  $AUC_{ref}$  (Fig. B.5).

We noticed clear distinctions in averaged prediction maps between experimental models (Fig. B.8; Appendices B.10 and B.11). For NB-GAMs, the prediction patterns were maintained down to 116 sightings, but under this threshold, patterns gradually disappeared. Despite a decrease in predicted densities, the distribution patterns predicted by the TW-GAMs remained the same down to 15 sightings. The distribution patterns predicted by ZIP-GAMs progressively disappeared below 364 sightings. Higher densities predicted by NB-GAMs, TW-GAMs and ZIP-GAMs fitted to thinned datasets were associated with lower uncertainties (Appendices B.12 and B.13). Furthermore, uncertainties of TW-GAMs were lower than those of NB-GAMs and ZIP-GAMs. The MaxEnt models showed some homogenisation of the distribution patterns with a decreasing number of sightings, but the general pattern was maintained no matter the number of sightings (Fig. B.8).



**Fig. B.8.** Prediction maps of auks averaged over the 100 models fitted to thinned datasets for each type of model in the Bay of Biscay and the English Channel. The rows represent the different types of generic models, and the columns represent the number of sightings used to fit the models. The numbers in the right corner of each map represent the number of sightings used to fit the model. The scale is in individuals.km<sup>-2</sup> (Ind/km<sup>2</sup>) for the NB-GAM, the TW-GAM and the ZIP-GAM and in the probability of presence (Pr.prob) for the MaxEnt model. This figure only shows the results for which a change was observed compared with the other predictions. All maps are presented in Appendices B.10 and B.11. The dotted lines represent the bathymetric strata of the survey area.

## B.4. DISCUSSION

### B.4.1 General considerations

To determine the model that would best predict the distribution of a rare species, we compared different types of models, both presence-absence and presence-only models. We assessed the predictive performance of a known model using a reduced amount of available data. Our findings suggest that the habitats for species that are rare or seldom seen are best described using a GAM with a Tweedie distribution (if effort data are available). GAMs with a negative binomial or zero-inflated Poisson distribution and MaxEnt models became inadequate for dataset under 130 sightings while TW-GAMs kept performing well down to a sample size of 30 sightings.

### Biological systems

Dolphins, including common and striped dolphins, and auks, including common guillemot, razorbill and Atlantic puffin, were used as biological models for two reasons. First, sightings of these taxa were large enough to allow proper statistical analyses and thinning to be conducted (277 dolphin sightings and 1,455 auk sightings). Second, dolphins and auks in the Bay of Biscay show well-defined and distinct patterns of distribution [11], which allows evaluating the predictive accuracy of the models.

Species groups pooled different species because of the difficulty to distinguish individuals at a species level from air. Pooling species into groups probably create categories with a broader habitat than the habitat of any of the constituting species, resulting in slightly larger sample size recommendations. However, the auk taxon is mainly dominated by the common guillemot and distribution patterns obtained in the study would mainly represent the common guillemot winter distribution. Indeed, auks wintering in the Bay of Biscay mostly originate from colonies located in the British Isles, where breeding populations of razorbill amount to 187,000 individuals, Atlantic puffin to 580,000 individuals and common guillemot to 1,416,000 individuals [39]. Concerning dolphins, combining the two species resulted in a bimodal habitat. Indeed, shipboard surveys (CODA; partly SCANS-II and SCANS-III) have shown that if the two species are present in all offshore habitats, the common dolphin would predominate over the shelf and the shelf-break, whereas the striped dolphin would be more frequent in oceanic waters. Consequently, this species complex could be seen as artificial but in fact it reflects habitat characteristics found in some delphinids like the bottlenose dolphin *Tursiops truncatus* with its pelagic and coastal ecotypes [40].

Auks and dolphins differ widely in their habitat specificity, particularly regarding depth, an environmental variable of major importance to characterise marine habitats. Hence, the sighting thinning experiment conducted in both taxa simulated two different cases of rarity (Table B.1). Thinning small delphinid sightings simulate a rare non-specialist species living in a broad habitat (row 2, column 1 of Table B.1) while thinning auk sightings generate a rare specialist species living in a narrower habitat (row 2 and column 4 in Table B.1). Modelling the habitat of the species described in the first row of Table 1 is not challenged by the number of sightings as the species is locally abundant, but is challenged by the location of the survey (if the survey was outside the core distribution of a species, sighting data would be scarce). Consequently, only habitat modelling for rare species of the second row of Table 1 remain an issue. To provide a more complete answer regarding the sample sizes needed to characterise pelagic animal distributions, further analyses and meta-analyses with multiple and diversified datasets should be conducted to obtain robust recommendations. We are aware that determining the number

of data needed to model the rare species habitats is an important challenge, for example to inform field efforts, but that a single study cannot consider all possible cases. An alternative research avenue would be to use virtual species instead of real species [41], which would allow to control all the conditions of the procedure but would not reflect the complex reality of the ecosystems. A methodology can work with a virtual species but fail in a real case.

### Baseline models

To assess the effect of the number of sightings, we tested three presence-absence models, NB-GAM, TW-GAM and ZIP-GAM, and one presence-only model, the MaxEnt model. All models tested in this study can handle datasets with many zeros but in different ways. We also wanted to test a presence-only model because in the case of rare and elusive species, opportunistic data, which represent a common example of presence-only data, often represent the majority of available data [42]. The MaxEnt model is able to model complex interactions between the response and the predictor variables [17,33,43], has been reported to be appropriate for presence-only datasets [44] and is widely used in species conservation planning due to its simplicity of use [21-23].

Variable selection was only performed on the baseline models. As the performance of a model is largely controlled by its selected variables [24], the models that used thinned-out sightings might be biased and are suboptimal (because some sighting data are ignored). Indeed, variables selected by a model fitted to few data could differ from models fitted on much larger datasets. However, we did not attempt to find the best model fit but to test the robustness of model predictions to thinning; variable selection was, to a certain extent, secondary to our purposes. In an ideal situation, the habitat of the species is known a priori. In practice, this is rarely the case, but in realistic situations, a SDM is first developed and then used repeatedly until the need to update it becomes an imperative. Thus the same SDM specification may be used without undergoing rounds of variable selection each time a new datum is added to an existing dataset. In a similar fashion, while MSE give guarantee on the predictive performance on average (*i.e.* under repeated use of the same model with different data generated from the same process), more often than not a single dataset is available for a given area. Consequently, to approximate a real situation in which one needs to model rare species habitats from a single dataset, the predictive capacity of each experimental model has been assessed in order to determine the probability for a single experimental model to reproduce the baseline model predictions.

### Thinning-out sighting data

Thinning rates applied in this study were arbitrarily determined to obtain, in the most extreme scenario, as few as 15-20 sightings, which is a threshold commonly observed for very rare species, particularly in marine megafauna [45,46]. Overfitting can be an issue with small datasets, *i.e.* the selected model becomes too complex compared to the number of implemented sightings [47,48]. Particularly, overfitting could have occurred in the models with the highest thinning rates. Nonetheless, the NB-GAM, the TW-GAM and the ZIP-GAM performed differently with the same small number of sightings (14-15 sightings). The NB-GAM and the ZIP-GAM did not manage to predict distribution patterns consistent with the baseline models, whereas the TW-GAM did.

#### B.4.2 Predicting habitats of rare species

Our aims were to assess the robustness of predictions from different SDMs by assessing prediction invariance under increasing levels of thinning of sightings used in model fitting. Overall, predictive

robustness differed between SDMs. All distributions predicted by the MaxEnt model were better than a homogeneous spatial distribution. There was, however, a gradual homogenisation of predicted dolphin presence probabilities over the whole area with increasing thinning rates. With very few sightings (approximately 28), MaxEnt was no longer able to distinguish key areas of either high or low presence probabilities. In contrast, despite some homogenisation of the predicted probabilities, auks' distribution patterns were correctly predicted, even with as few as 15 sightings. Consequently, thinning affected model predictive performance differently whether the studied taxon was a generalist or specialist one. However, these results not be truly representative of the empirical performance of MaxEnt. Our data were collected with a standard protocol that ensured a balanced coverage over the Bay of Biscay, which conforms to the assumptions underlying the appropriate use of presence-only models [20]. This may not be the general case with presence-only data, where survey effort is often biased. Despite a balanced sampling effort in the field, MaxEnt did not provide satisfactory results for the highest thinning rates, calling into question its use for rare species.

Because thinning sightings emulate false absences (that is a zero observation due for example to imperfect detection in a nevertheless suitable habitat), we expected a better performance by the ZIP-GAM. However, the results were less reliable than those obtained with a TW-GAM. Below approximately 130 sightings, the predicted distributions of the ZIP-GAM were unreliable compared to the predictions of the baseline model, whereas this threshold was as low as approximately 30 sightings for the TW-GAM. This difference is likely due to the current parametrisation of the ZIP family in the 'mgcv' package [25,26]. In fact, the current parametrisation uses the linear predictors and linearly scales them on a logit scale to generate extra zero observations (see the help pages in mgcv v1.8-9; [26]). This parametrisation implicitly assumes that the areas with lower densities have a higher probability of non-detection. However, the parameterisation does not allow for incorporating detection-specific covariates, which may better explain the non-detection patterns. Similarly, the NB-GAM provided less convincing results and unreliable predicted distribution patterns compared to the baseline model below approximately 130 sightings.

Even if the TW-GAM provided good results with approximately 20-25 sightings, the results were based on the averages of 100 fitted models and hid substantial variations. In practice, habitat modellers have only one dataset. Therefore, we assessed the individual performance of each experimental model by computing the number of models in which the MSE was higher than the  $MSE_{ref}$  and by examining the explained deviances of each experimental model (results not shown). It appeared that with 20 sightings, approximately 50 of the 100 experimental TW-GAMs predicted better than a homogeneous spatial distribution of the two species groups whereas with 40 sightings, 90 of the 100 experimental models provided reliable results (Fig. B.6). Moreover, by examining the explained deviances for each experimental Tweedie model (results not shown), we found that explained deviances of the experimental models fitted to 28 and 69 sightings for dolphins were good (30-50%). For the smallest number of data (15 and 23 sightings), the explained deviances were very high (>50%) which suggested overfitting. Consequently, to obtain robust predictions, a number of 50 sightings would represent a conservative empirical measure. However, this number is only valid for the TW-GAM because with the NB-GAM and the ZIP-GAM, the threshold for which all experimental models provided good results (better than a homogeneous spatial distribution) was 100 sightings (Fig. B.6).

Finally, this study provided a first answer to the question commonly asked by habitat modellers: "What model should be used when studying rare species?" If modellers only have presence data, MaxEnt could be used but with great caution and preferably for specialist species with restricted



distributions. With effort data, we would recommend using a GAM with a Tweedie distribution and a minimum of 50 sightings, which is a conservative empirical measure.

### Acknowledgments

We thank H el ene Falchetto for processing the survey data that were used in the study. We are grateful to PREVIMER for providing us outputs from MARS-2D and MARS-3D models. We thank M elanie Racine for participating in the development of the baseline models. We thank Charlotte Lambert and the anonymous reviewers for their very helpful comments that led to a clearer and much improved manuscript.

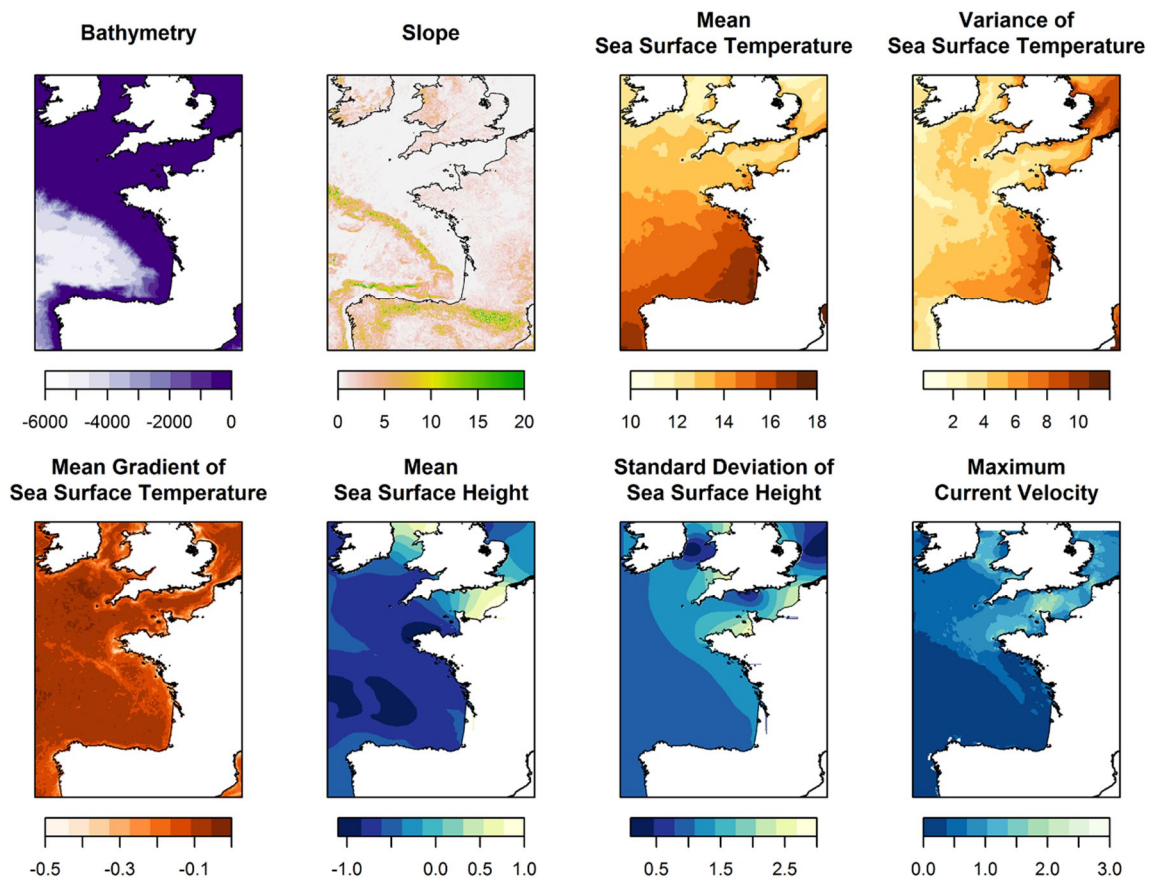
### References

- [1] Rabinowitz D (1981) Seven forms of rarity. The biological aspects of rare plant conservation.
- [2] Cunningham R.B and Lindenmayer D.B (2005) Modeling Count Data of Rare Species. *Ecology* 86(5): 1135–1142.
- [3] Lawler J.J, White D, Sifneos J.C, Master L.L (2003) Rare Species and the Use of Indicator Groups for Conservation Planning. *Conservation Biology* 17(3): 875–882.
- [4] Redfern J.V, Ferguson M.C, Becker E.A, Hyrenbach K.D, Good C, Barlow J et al. (2006) Techniques for cetacean – habitat modeling. *Mar. Ecol.Prog. Ser.* 310: 271–295.
- [5] 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(1–3): 297–308.
- [6] Barry S.C and Welsh A.H (2002) Generalized additive modelling and zero inflated count data. *Ecological Modelling* 157(2-3): 179–188.
- [7] Engler R, Guisan A, Rechsteiner L (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology* 41(2): 263–274.
- [8] Buckland S.T, Anderson D.R, Burnham H.P, Laake J.L, Borchers D.L, Thomas L (2001) *Introduction to Distance Sampling: Estimating Abundance of Biological Populations*. Oxford University Press, Oxford.
- [9] Certain G. and Bretagnolle V (2008) Monitoring seabirds population in marine ecosystem: The use of strip-transect aerial surveys. *Remote Sensing of Environment* 112: 3314–3322.
- [10] Laran S, Authier M, Blanck A, Dor emus G, Falchetto H, Monestiez P et al. (2017) Seasonal distribution and abundance of cetaceans within French waters. Part II: the Bay of Biscay and the English Channel. *Deep Sea Research Part II* 141: 31-40.
- [11] Lambert C, Pettex E, Dor emus G, Laran S, Stephan E, Van Canneyt O, Ridoux V (2017) How does ocean seasonality drive habitat preferences of highly mobile top predators? Part II: the eastern North-Atlantic. *Deep-Sea Research II* 141: 133-154.
- [12] Virgili A, Lambert C, Pettex E, Dor emus G, Van Canneyt O, Ridoux V (2017) Predicting seasonal variations in coastal seabird habitats in the English Channel and the Bay of Biscay. *Deep Sea Research II* 141: 212-223.
- [13] Previmer (2014) *Previmer - Observation et previsions c otieres*. Catalogue version 2.1.
- [14] Guisan A. and Zimmermann N.E (2000) Predictive habitat distribution models in ecology. *Ecological Modelling* 135(2-3): 147–186.

- [15] Brotons L, Thuiller W, Araujo M.B, Hirzel A.H (2004) Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography* 4: 437–448.
- [16] Gormley A.M, Forsyth D.M, Griffioen P, Lindeman M, Ramsey D.S.L, Scroggie M.P, Woodford L (2011) Using presence-only and presence-absence data to estimate the current and potential distributions of established invasive species. *Journal of Applied Ecology* 48(1): 25–34.
- [17] Elith J, Graham C.H, Anderson R.P, Dudík M, Ferrier S, Guisan A et al. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29: 129–151.
- [18] Tsoar A, Allouche O, Steinitz O, Rotem D, Kadmon R (2007) A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions* 13(4): 397–405.
- [19] Zaniwski A.E, Lehmann A, Overton J.M (2002) Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling* 157(2-3): 261–280.
- [20] Yackulic C.B, Chandler R, Zipkin E.F, Royle A, Nichols J.D, Campbell Grant E.H, Veran S (2013) Presence-only modelling using MAXENT: when can we trust the inferences? *Methods in Ecology and Evolution* 4: 236–243.
- [21] Elith, J. & Leathwick, J. (2009) Conservation prioritisation using species distribution modelling. *Spatial conservation prioritization: quantitative methods and computational tools*. Oxford University Press, Oxford, UK, 70-93.
- [22] Phillips S.J, Anderson R.P, Schapire R.E (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling* 190: 231–259.
- [23] McClellan C.M, Brereton T, Dell'Amico F, Johns D.G, Cucknell A.C, Patrick, S.C. et al. (2014) Understanding the distribution of marine megafauna in the English Channel region: identifying key habitats for conservation within the busiest seaway on earth. *PloS one* 9(2): e89720.
- [24] Syphard A.D and Franklin, J (2009) Differences in spatial predictions among species distribution modeling methods vary with species traits and environmental predictors. *Ecography* 32(6): 907–918.
- [25] Wood S.N (2006) On confidence intervals for generalized additive models based on penalized regression splines. *Australian and New Zealand Journal of Statistics* 48(4): 445–464.
- [26] Wood S. (2013) mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation. Retrieved 7 July 2014, from <http://cran.r-project.org/web/packages/mgcv/index.html>
- [27] Ferguson M.C, Barlow J, Reilly S.B, Gerrodette T (2006) Predicting Cuvier's (*Ziphius cavirostris*) and Mesoplodon beaked whale population density from habitat characteristics in the eastern tropical Pacific Ocean. *Journal of Cetacean Research and Management* 7(3): 287–299.
- [28] Hastie T. and Tibshirani R (1986) Generalized Additive Models. *Statistical Science* 3: 297-313.
- [29] Mannocci L, Catalogna M, Dorémus G, Laran S, Lehodey P, Massart W et al. (2014) Predicting cetacean and seabird habitats across a productivity gradient in the South Pacific gyre. *Progress in Oceanography* 120: 383–398.
- [30] 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.
- [31] Wood S (2006) *Generalized Additive models: An Introduction with R*. Chapman and Hall/CRC.
- [32] Clark, M. (2013). *Generalized additive models: getting started with additive models in R*. Center for Social Research, University of Notre Dame, 35.

- [33] Elith J, Phillips S.J, Hastie T, Dudík M, Yung En Chee, Yates C.J (2011) A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions* 17(1): 43–57.
- [34] Phillips S.J and Dudík M (2008) Modeling of species distribution with Maxent: new extensions and a comprehensive evaluation. *Ecography* 31: 161–175.
- [35] Merow C, Smith M.J, Silander J.A (2013) A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography* 36(10): 1058–1069.
- [36] R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from <http://www.R-project.org/all>
- [37] Wallach D and Goffinet B (1989) Mean squared error of prediction as a criterion for evaluating and comparing system models. *Ecological Modelling* 44(3-4): 299–306.
- [38] Harvey D, Leybourne S, Newbold P (1997) Testing the equality of prediction mean squared errors. *International Journal of Forecasting* 13(2): 281–291.
- [39] Mitchell, P.I., Newton, S.F., Ratcliffe, N. Dunn, T.E. (2004) *Seabird Populations of Britain and Ireland*. Poyser, London.
- [40] Shirihai, H. and Jarrett, B. (2006) *Whales Dolphins and Other Marine Mammals of the World*. Princeton: Princeton Univ. Press. pp. 155–158.
- [41] Zurell D, Berger U, Cabral J.S, Jeltsch F, Meynard C.N, Münkemüller T et al. (2010) The virtual ecologist approach: simulating data and observers. *Oikos* 119(4): 622-635.
- [42] Pearce J.L and Boyce M.S (2006) Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology* 43: 405–412.
- [43] Phillips S.J, Dudík M, Schapire R.E (2004) A Maximum Entropy Approach to Species Distribution Modeling. Twenty-first international conference on Machine learning - ICML '04, p.83
- [44] Wisz M.S, Hijmans R.J, Li J, Peterson A.T, Graham C.H, Guisan A (2008) Effects of sample size on the performance of species distribution models. *Diversity and Distributions* 14(5): 763–773.
- [45] Marcos E, Sierra J.M, De Stephanis R (2010). Cetacean diversity and distribution in the coast of Gipuzkoa and adjacent waters, south-eastern Bay of Biscay. *Munibe Ciencias Naturales. Natur zientziak* 58: 221-231.
- [46] Arcangeli A, Campana I, Marini L, MacLeod C.D (2015) Long-term presence and habitat use of Cuvier's beaked whale (*Ziphius cavirostris*) in the Central Tyrrhenian Sea. *Marine Ecology*: 1–14.
- [47] Hawkins D.M (2004) The problem of overfitting. *J. Chem. Inf. Comput. Sci* 44: 1-12.
- [48] Subramanian J and Simon R (2013) Overfitting in prediction models - Is it a problem only in high dimensions? *Contemporary Clinical Trials* 36(2): 636–641.

## Appendix B.1. Maps of averaged covariates over the entire survey.



**Appendix B.2.** Some key concepts about the models used in the study.

- Negative Binomial distribution

The Negative Binomial distribution is extended from the Poisson regression and is defined by two parameters, the arithmetic mean and an exponent  $k$ . By modulating this exponent  $k$ , this distribution can be adapted to over-dispersed data (Bliss & Fisher 2016). If the response variable  $Y$  obeys to a Negative Binomial distribution, the variance  $V(Y)$  and the mean  $E(Y)$  are related by the relationship  $V(Y) = E(Y) + kE(Y)^2$  (Ver Hoef & Boveng 2007).

In this study, we wanted to fit a GAM with a Negative Binomial distribution so we used the R package `mgcv` (Wood 2006; 2013) and the `gam` function specifying the “Negative Binomial” family. Besides, we used the `nb` function to estimate the parameter  $k$  during the fitting.

- Tweedie distribution

The Tweedie distribution is useful to model continuous positive data because, compared to the Poisson distribution, it includes an additional parameter  $p$  which defines the model distribution. Indeed, if  $p=0$ , it is a normal distribution, if  $p=1$ , it is a Poisson distribution and if  $p=2$ , it is a Gamma distribution. Tweedie models can handle zero-inflated data (*i.e.* data with many zeros), because when  $1 < p < 2$ , they are a Poisson mixture of Gammas distributions (Arcuti et al. 2013). Besides, for a response variable  $Y$  that obeys a Tweedie distribution, the variance  $V(Y)$  and the mean  $E(Y)$  are related by the relationship  $V(Y) = \varphi E(Y)^p$  where  $\varphi$  represents the dispersion parameter (Dunn & Smyth 2005).

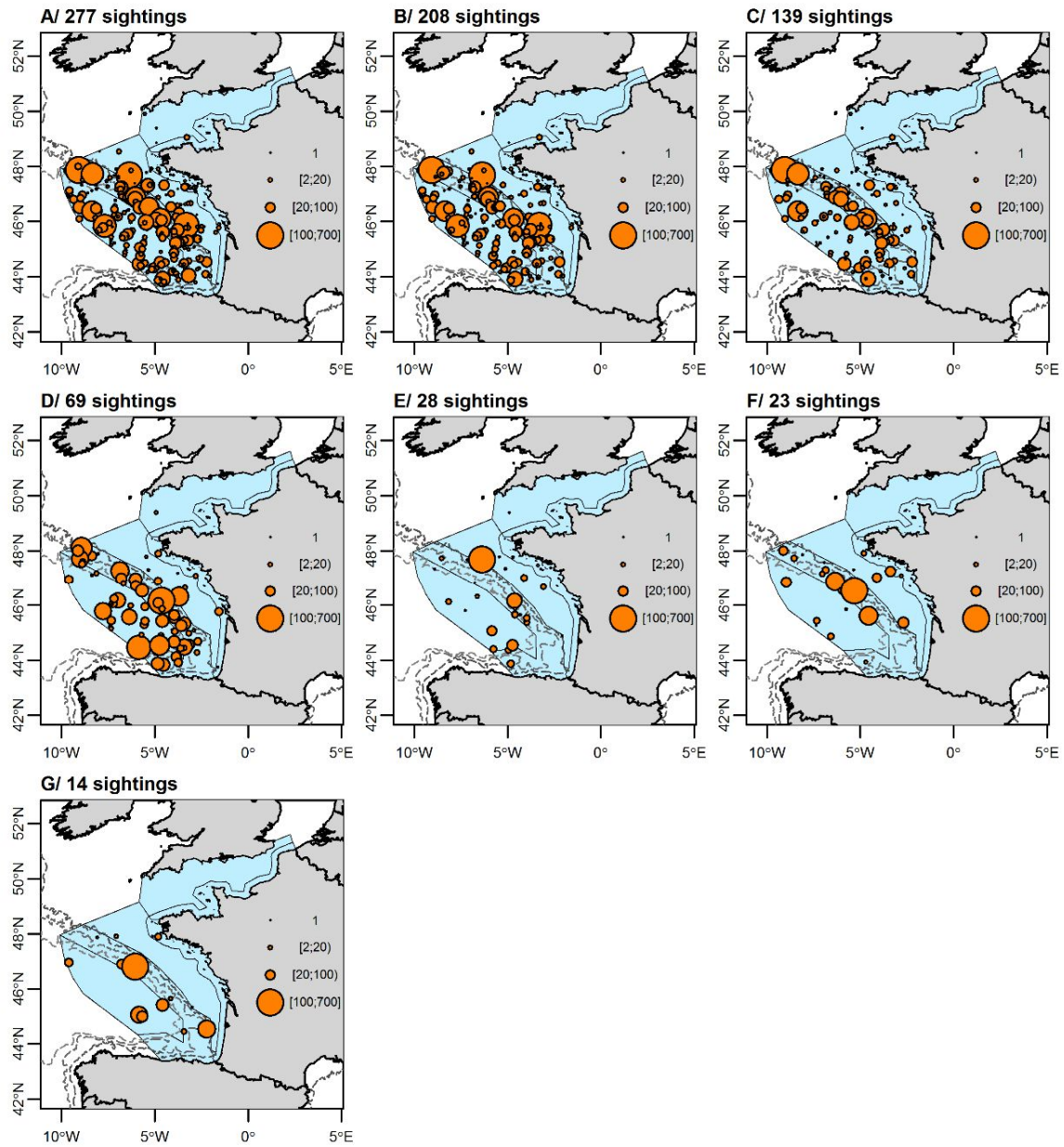
To fit a GAM with a Tweedie distribution, we used the R package `mgcv` (Wood 2006; 2013) and the `gam` function specifying the “Tweedie” family. Besides, as we ignored the value of the parameter  $p$ , we used the `tw` function which estimates this parameter during the fitting.

- Zero-inflated Poisson distribution

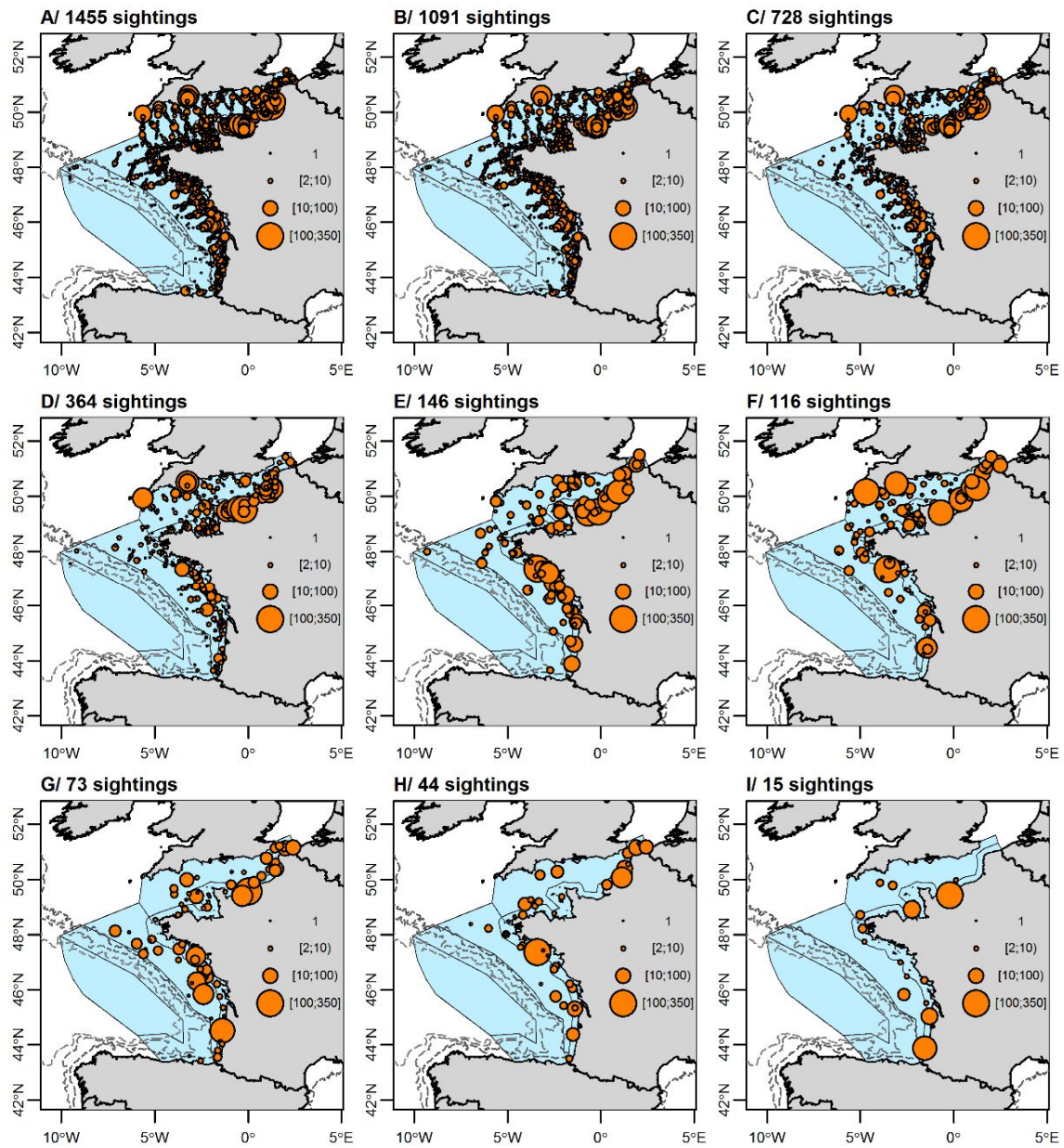
Zero-inflated Poisson distribution is used to model count data with extra zero counts by modelling independently the count values, with a Poisson distribution (Zeileis et al. 2007), and the excess of zeros (Lambert 1992). Thus, the ZIP regression is divided into two parts in which the species probability of presence and, given the presence, the species abundance are modelled sequentially (Ridout et al. 1998; Wenger & Freeman 2014).

In the study, we tested the ZIP distribution with a GAM which showed nonlinear relationships between the response variable and the environmental predictors. To fit the model, we used the `mgcv` package and the `gam` function but we specified the ZIP family and we used the `ziP` function to estimate the  $\vartheta$  parameter. This parameter includes two parameters which control the slope and the intercept of the zero model (Wood 2006; 2013). To fit smooth functions for the GAM we introduced a  $k$  parameter which worth 4, for 4 degrees of freedom.

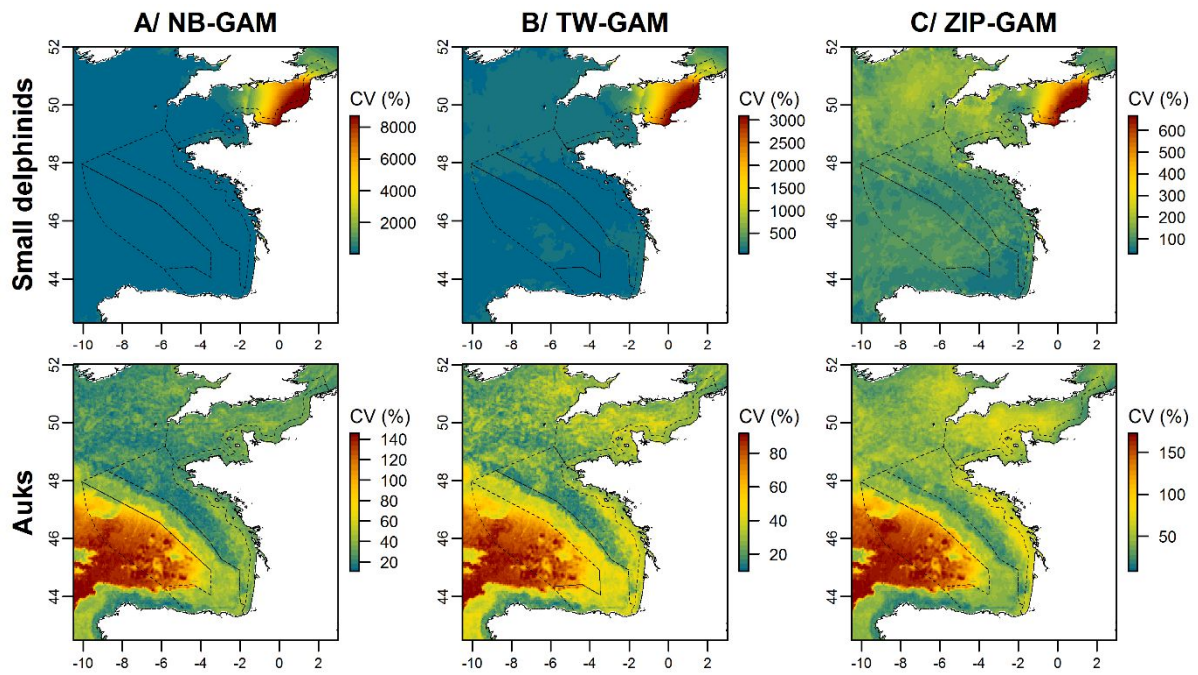
**Appendix B.3. Sighting thinning example for the dolphin dataset.** Sightings are classified by group sizes (1; 2-20; 20-100 and 100-700 individuals) with each point representing a group of individuals.



**Appendix B.4. Sighting thinning example for the auk dataset.** Sightings are classified by group sizes (1; 2-10; 10-100 and 100-350 individuals) with each point representing a group of individuals.

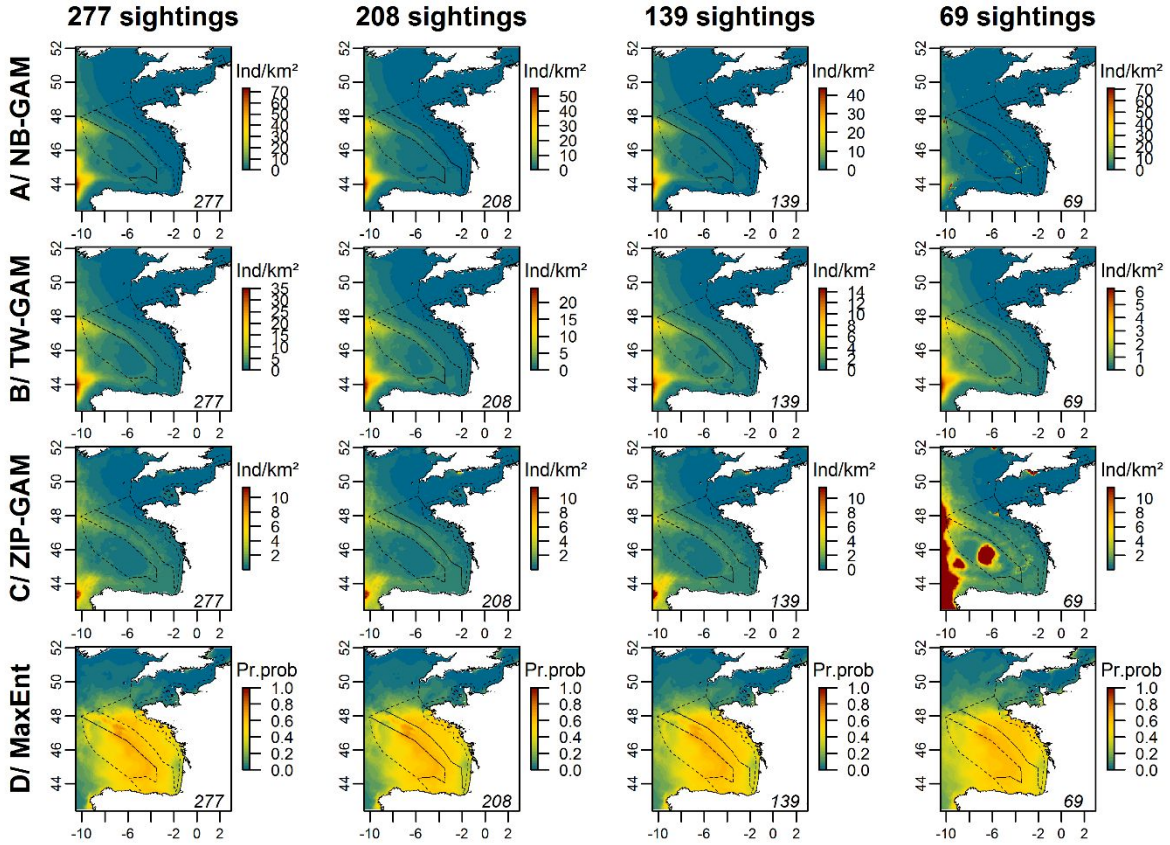


**Appendix B.5. Uncertainty maps of baseline models.** Uncertainty maps representing the coefficient of variation in % associated with the predictive relative density of dolphin and auk groups. Dotted lines represent the survey area.

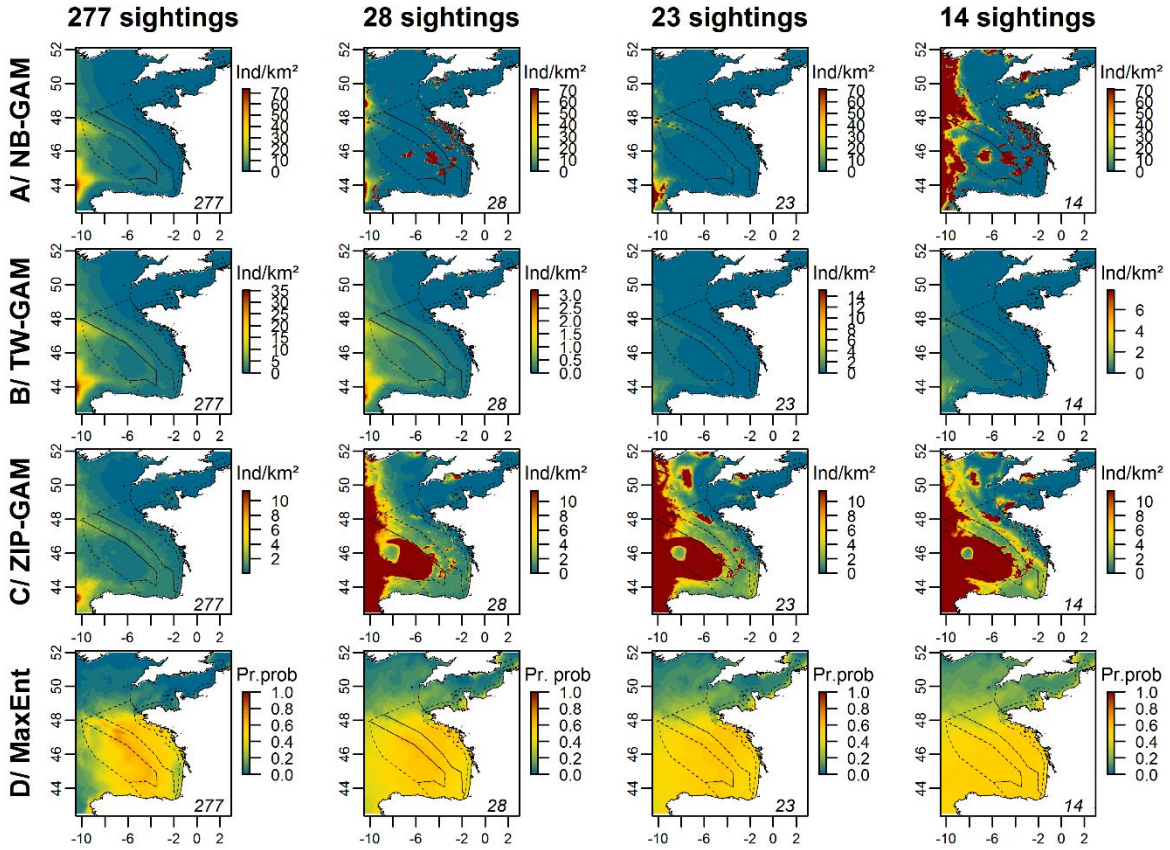




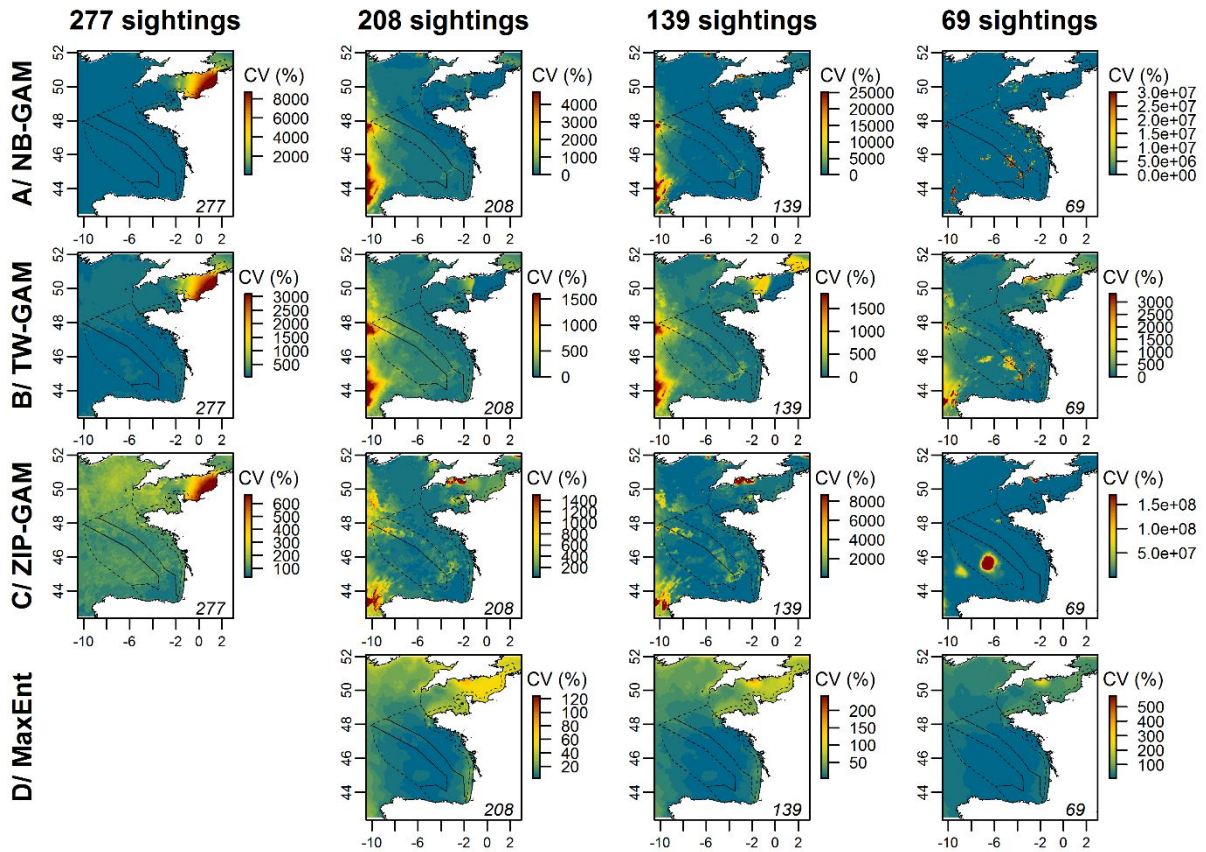
Appendix B.6. Prediction maps of dolphins averaged over 100 models fitted to thinned datasets for each type of model from 25 to 75% of sighting thinning. The rows represent the different types of generic models, and the columns represent the number of sightings used to fit the models. The numbers in the right corner of each map represent the number of sightings used to fit the model. The scale is in individuals.km-2 (Ind/km<sup>2</sup>) for the NB-GAM, the TW-GAM and the ZIP-GAM and in the probability of presence (Pr.prob) for MaxEnt.



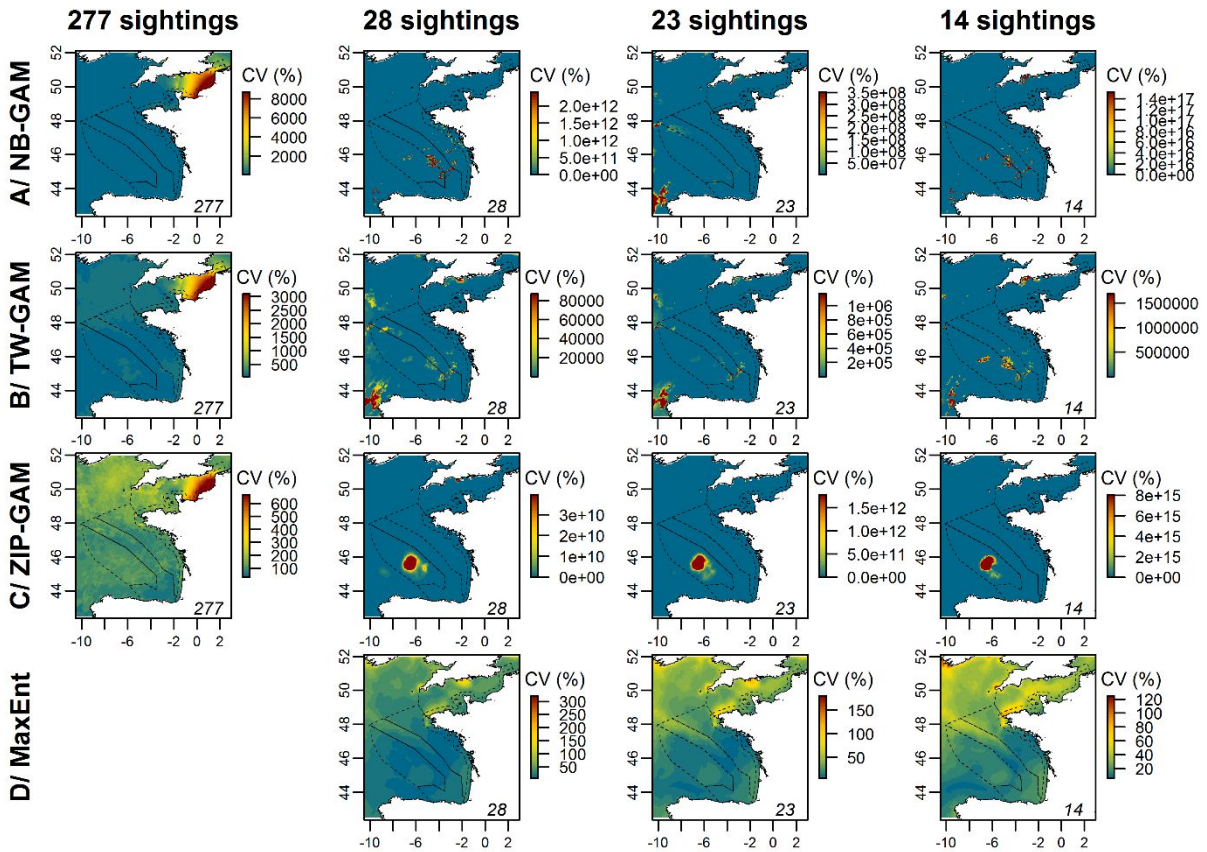
Appendix B.7. Prediction maps of dolphins averaged over 100 models fitted to thinned datasets for each type of model from 90 to 95% of sighting thinning. The rows represent the different types of generic models, and the columns represent the number of sightings used to fit the models. The numbers in the right corner of each map represent the number of sightings used to fit the model. The scale is in individuals.km-2 (Ind/km<sup>2</sup>) for the NB-GAM, the TW-GAM and the ZIP-GAM and in the probability of presence (Pr.prob) for MaxEnt.



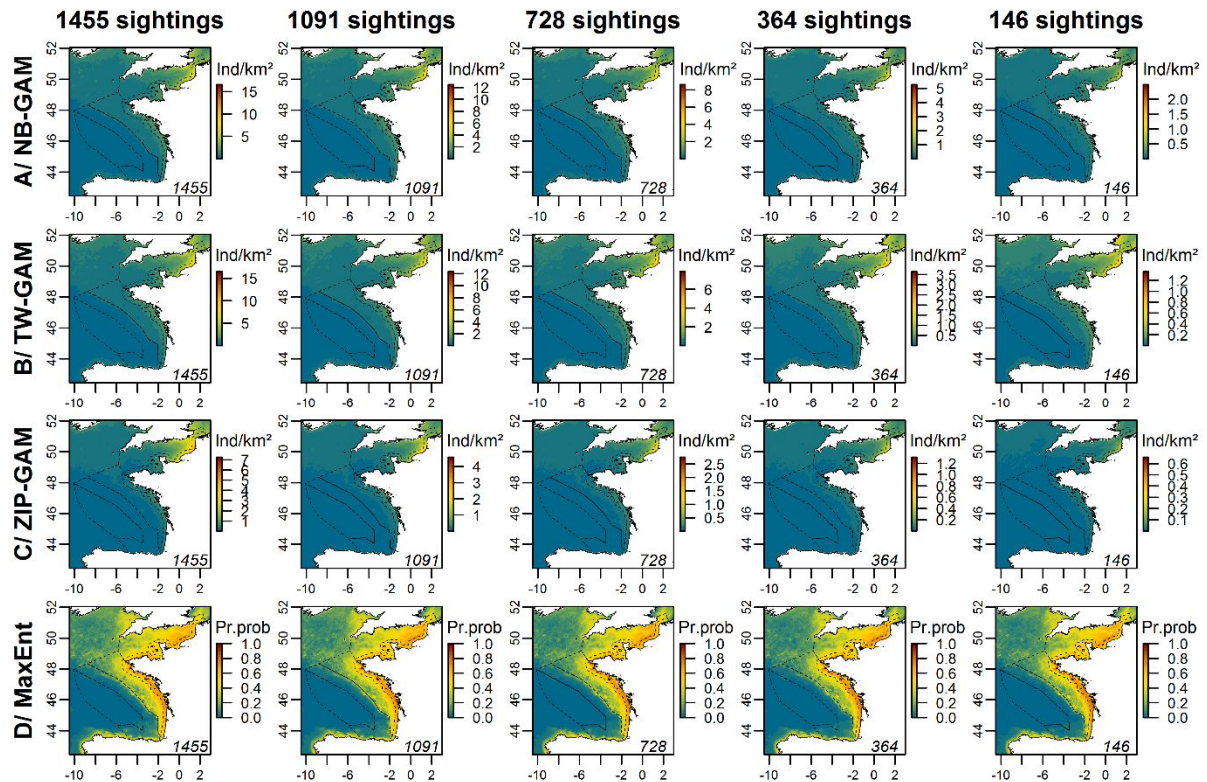
Appendix B.8. Averaged uncertainty maps representing the coefficient of variation of each thinning rate in % associated with the averaged predictive density of dolphin group, from 25 to 75% of sighting thinning. The rows represent the different types of generic models, and the columns represent the number of sightings used to fit the models. The numbers in the right corner of each map represent the number of sightings used to fit the model. Due to very high isolated values, the maps were not contrasted so each coefficient of variation value beyond the 99% quantile were truncated. Dotted lines represent the survey area.



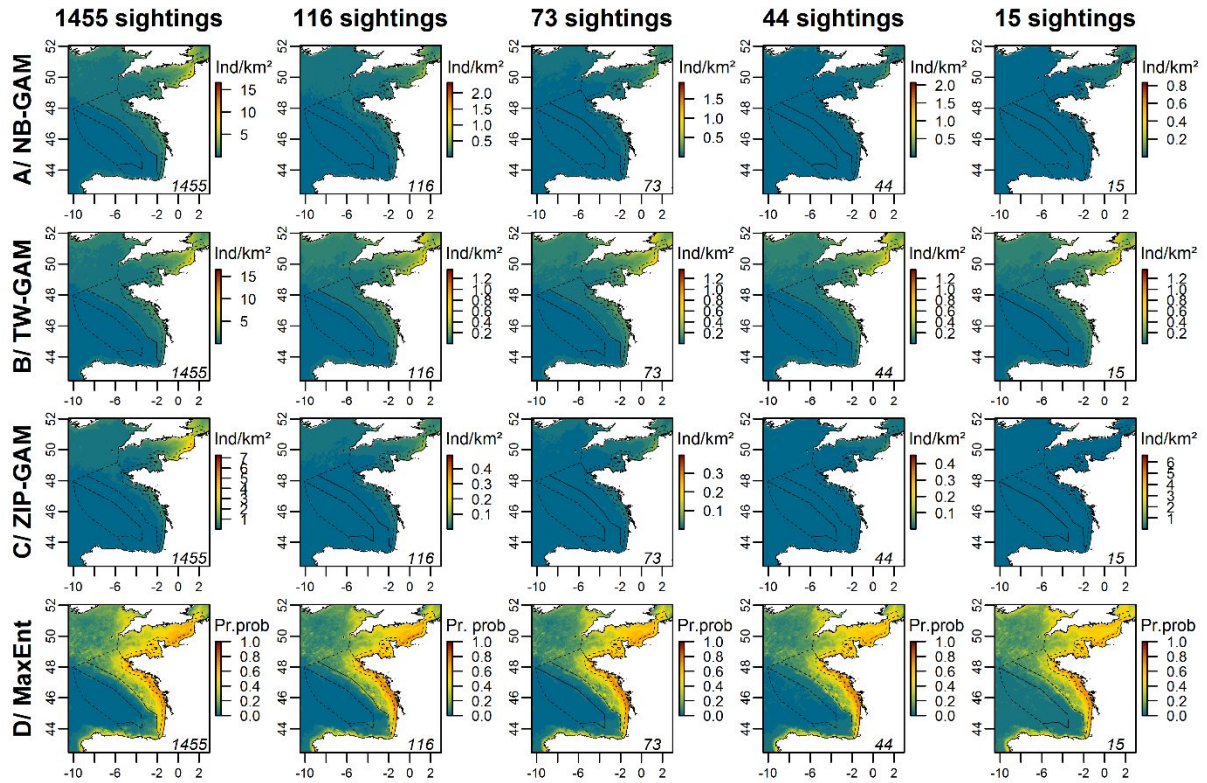
Appendix B.9. Averaged uncertainty maps representing the coefficient of variation of each thinning rate in % associated with the averaged predictive density of dolphin group, from 90 to 95% of sighting thinning. The rows represent the different types of generic models, and the columns represent the number of sightings used to fit the models. The numbers in the right corner of each map represent the number of sightings used to fit the model. Due to very high isolated values, the maps were not contrasted so each coefficient of variation value beyond the 99% quantile were truncated. Dotted lines represent the survey area.



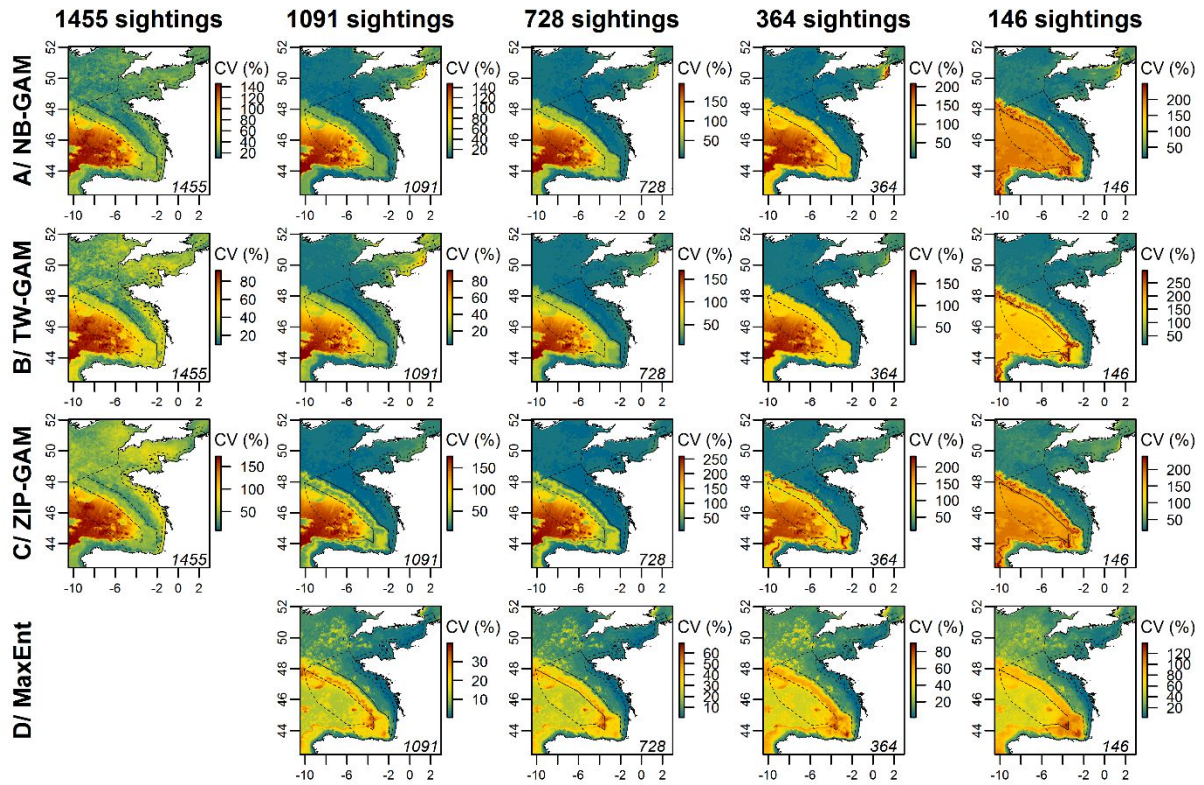
Appendix B.10. Prediction maps of auks averaged over 100 models fitted to thinned datasets for each type of model from 25 to 90% of sighting thinning. The rows represent the different types of generic models, and the columns represent the number of sightings used to fit the models. The numbers in the right corner of each map represent the number of sightings used to fit the model. The scale is in individuals.km-2 (Ind/km<sup>2</sup>) for the NB-GAM, the TW-GAM and the ZIP-GAM and in the probability of presence (Pr.prob) for MaxEnt.



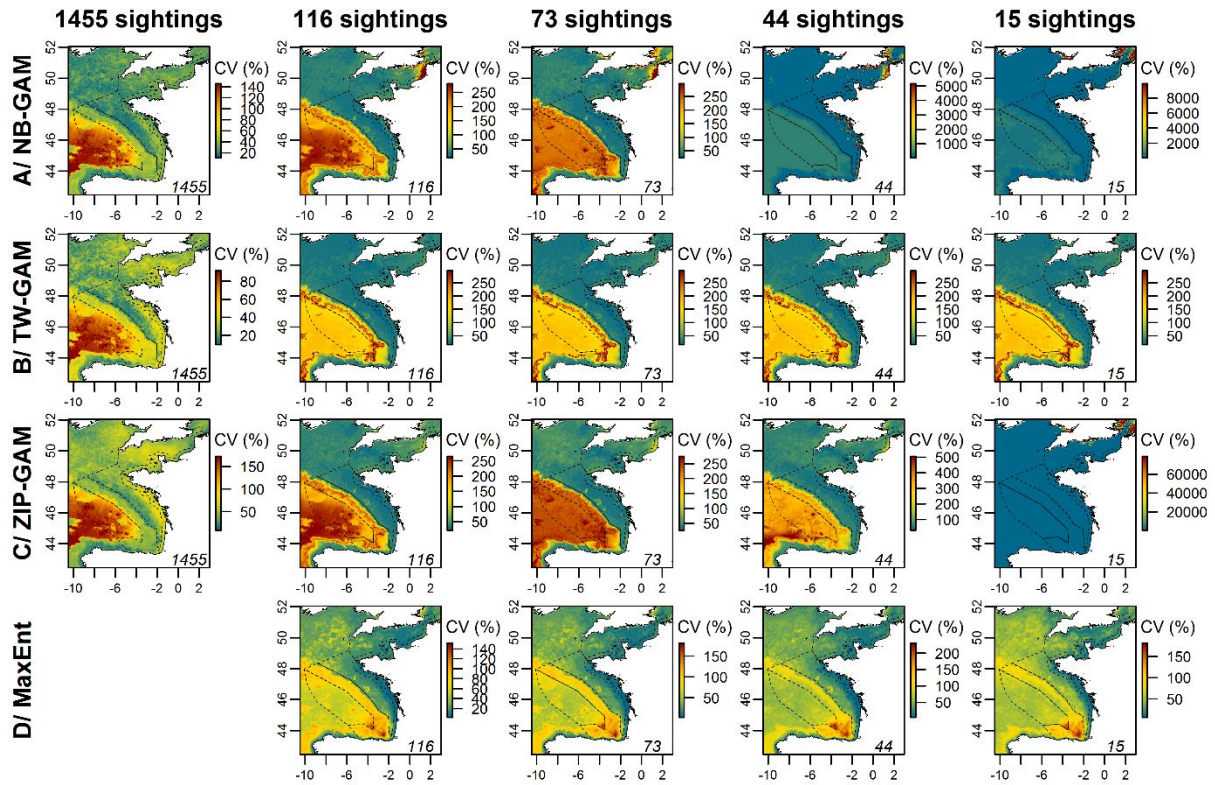
Appendix B.11. Prediction maps of auks averaged over 100 models fitted to thinned datasets for each type of model from 92 to 99% of sighting thinning. The rows represent the different types of generic models, and the columns represent the number of sightings used to fit the models. The numbers in the right corner of each map represent the number of sightings used to fit the model. The scale is in individuals.km-2 (Ind/km<sup>2</sup>) for the NB-GAM, the TW-GAM and the ZIP-GAM and in the probability of presence (Pr.prob) for MaxEnt.



Appendix B.12. Averaged uncertainty maps representing the coefficient of variation of each thinning rate in % associated with the averaged predictive density of auk group, from 25 to 90% of sighting thinning. The rows represent the different types of generic models, and the columns represent the number of sightings used to fit the models. The numbers in the right corner of each map represent the number of sightings used to fit the model. Due to very high isolated values, the maps were not contrasted so each coefficient of variation value beyond the 99% quantile were truncated. Dotted lines represent the survey area.



Appendix B.13. Averaged uncertainty maps representing the coefficient of variation of each thinning rate in % associated with the averaged predictive density of auk group, from 92 to 99% of sighting thinning. The rows represent the different types of generic models, and the columns represent the number of sightings used to fit the models. The numbers in the right corner of each map represent the number of sightings used to fit the model. Due to very high isolated values, the maps were not contrasted so each coefficient of variation value beyond the 99% quantile were truncated. Dotted lines represent the survey area.







# Annex C

---

## COMBINING VISUAL SURVEYS TO MODEL HABITAT OF DEEP-DIVING CETACEANS AT THE BASIN SCALE

---

Auriane Virgili, Matthieu Authier, Oliver Boisseau, Ana Cañadas, Diane Claridge,  
Tim Cole, Peter Corkeron, Ghislain Dorémus, Léa David, Nathalie Di-Méglio,  
Charlotte Dunn, Tim E. Dunn, Isabel García Barón, Sophie Laran, Giancarlo  
Lauriano, Mark Lewis, Maite Louzao, Laura Mannocci, José Martínez-Cedeira,  
Debra Palka, Simone Panigada, Emeline Pettex, Jason J. Roberts, Leire Ruiz,  
Camilo Saavedra, M. Begoña. Santos, Olivier Van Canneyt, José Antonio Vázquez  
Bonales, Pascal Monestiez, Vincent Ridoux

To be submitted

## Combining visual surveys to model habitat of deep-diving cetaceans at the basin scale.

A. Virgili <sup>1\*</sup>, M. Authier <sup>2</sup>, O. Boisseau <sup>3,4</sup>, A. Cañadas <sup>5</sup>, D. Claridge <sup>6</sup>, T. Cole <sup>7</sup>, P. Corkeron <sup>7</sup>, G. Dorémus <sup>2</sup>, L. David <sup>8</sup>, N. Di-Méglio <sup>8</sup>, C. Dunn <sup>6</sup>, T.E. Dunn <sup>9</sup>, I. García Barón <sup>10</sup>, S. Laran <sup>2</sup>, G. Lauriano <sup>11</sup>, M. Lewis <sup>9</sup>, M. Louzao <sup>10</sup>, L. Mannocci <sup>12,13</sup>, J. Martínez-Cedeira <sup>14</sup>, D. Palka <sup>7</sup>, S. Panigada <sup>15</sup>, E. Pettex <sup>2</sup>, J.J. Roberts <sup>12</sup>, L. Ruiz <sup>16</sup>, C. Saavedra <sup>17</sup>, M.B. Santos <sup>17</sup>, O. Van Canneyt <sup>2</sup>, J. A. Vázquez Bonales <sup>4</sup>, P. Monestiez <sup>1,18</sup>, V. Ridoux <sup>1,2</sup>

<sup>1</sup> Centre d'Etudes Biologiques de Chizé - La Rochelle, UMR 7372 CNRS - Université de La Rochelle, Institut du Littoral et de l'Environnement, 17000 La Rochelle, France; <sup>2</sup> Observatoire PELAGIS, UMS 3462 CNRS - Université de La Rochelle, Systèmes d'Observation pour la Conservation des Mammifères et des Oiseaux Marins, 17000 La Rochelle, France; <sup>3</sup> Marine Conservation Research, 94 High Street, Kelvedon, CO5 9AA, UK; <sup>4</sup> Song of the Whale research team, International Fund for Animal Welfare (IFAW), 87-90 Albert Embankment, London, SE1 7UD, UK; <sup>5</sup> Alnilam Research and Conservation, Pradillos 29, 28491-Navacerrada, Madrid, Spain; <sup>6</sup> Bahamas Marine Mammal Research Organisation; <sup>7</sup> Protected Species Branch, NOAA Fisheries Northeast Fisheries Science, 166 Water St Woods Hole Massachusetts, 02543 USA; <sup>8</sup> EcoOcéan Institut, 34090 Montpellier, France; <sup>9</sup> Joint Nature Conservation Committee, Inverdee House, Baxter Street, Aberdeen, AB11 9QA, UK; <sup>10</sup> AZTI, Herrera Kaia, Portualdea z/g, 20110 Pasaia, Spain; <sup>11</sup> Institute for Environmental Protection and Research – ISPRA, via V. Brancati 60, 00144 Roma, Italy; <sup>12</sup> Marine Geospatial Ecology Laboratory, Duke University, Durham, North Carolina, USA; <sup>13</sup> UMR MARBEC (IRD, Ifremer, Univ. Montpellier, CNRS), Institut Français de Recherche pour l'Exploitation de la Mer, Av. Jean Monnet, CS 30171, 34203 Sète, France; <sup>14</sup> CEMMA, Camiño do Ceán, n° 2, 36350, Nigrán, Pontevedra, Spain; <sup>15</sup> Tethys Research Institute, Acquario Civico, 20121 Milano, Italy; <sup>16</sup> AMBAR Elkartea organisation, Mungia Bidea 9, 3B, 48620 Plentzia, Bizkaia, Spain; <sup>17</sup> Instituto Español de Oceanografía, Centro Oceanográfico de Vigo, 36390 Vigo, Spain; <sup>18</sup> BioSP, INRA, 84914 Avignon, France.

### Abstract

Deep-diving cetaceans are oceanic species exposed to multiple anthropogenic pressures (e.g. high intensity underwater noise) and knowledge of their distributions and abundance is crucial to inform their conservation. However, due to their low densities, wide distribution ranges and limited presence at the water surface, visual surveys usually result in low sighting rates. To circumvent this limitation, gathering data from multiple visual surveys appeared as a key strategy but implied to take into account the various protocols, platforms and temporal heterogeneity between the surveys. This study aimed to describe the entire procedure that assemble data from different surveys in order to model the large scale habitats of deep-divers. About 1,240,000 km of effort performed in multiple visual surveys across the North Atlantic Ocean and the Mediterranean Sea provided 630 sightings of ziphiids, 836 of physeteriids and 106 of kogiids. We implemented a meta-analysis to determine the effective strip width for each species group and modelled species relative densities in a Generalised Additive Model framework. We produced the first basin-wide deep-diver density maps in the North Atlantic Ocean and the Mediterranean Sea. A gap analysis highlighted areas of environmental interpolation. Deeper areas of the North Atlantic gyre were mostly areas of environmental extrapolation and were not intensively sampled. For the three species groups, highest densities were predicted in deeper waters and close to thermal fronts. Predictions identified areas of concentration along the continental slopes, in particular in the western North Atlantic Ocean where the Gulf Stream runs.

## C.1. INTRODUCTION

Deep-diving cetaceans, defined here as beaked whales (family *Ziphiidae*; e.g. *Ziphius cavirostris*, *Hyperoodon* spp. and *Mesoplodon* spp.) and sperm whales (families *Physeteridae* and *Kogiidae*), are distributed worldwide. They are oceanic species groups frequently associated with steep slope habitats where they feed in deep waters during long dives (often more than an hour; Perrin et al. 2009). Due to their offshore habitat, short time at the surface and therefore low availability to sightings, little is known about their densities within their distributional range (especially for kogiids and ziphiids). These species are threatened by a variety of anthropogenic activities, such as bycatch, debris ingestion and ship collisions (Carrillo and Ritter 2010; Madsen et al. 2014; Unger et al. 2016) but the threat whose effect are best known are activities that produce high intensity signals (e.g. military sonars, seismic guns or techniques used on large maritime construction projects; Stone and Tasker 2006). Recent studies have demonstrated the sensitivity of deep-diving cetaceans, and particularly beaked whales, to underwater noise pollution. Certain sounds can cause death to these whales and several unusual stranding events have occurred in connection with the use of military sonars (Frantzis 1998; Balcomb and Claridge 2001; Brownell et al. 2004; Fernández et al. 2005; D'Amico et al. 2009). To mitigate the impact of these activities, good knowledge of the distribution and hotspots of concentration of deep-diving cetaceans is crucial to Marine Spatial Planning to guide management measures (Douve 2008).

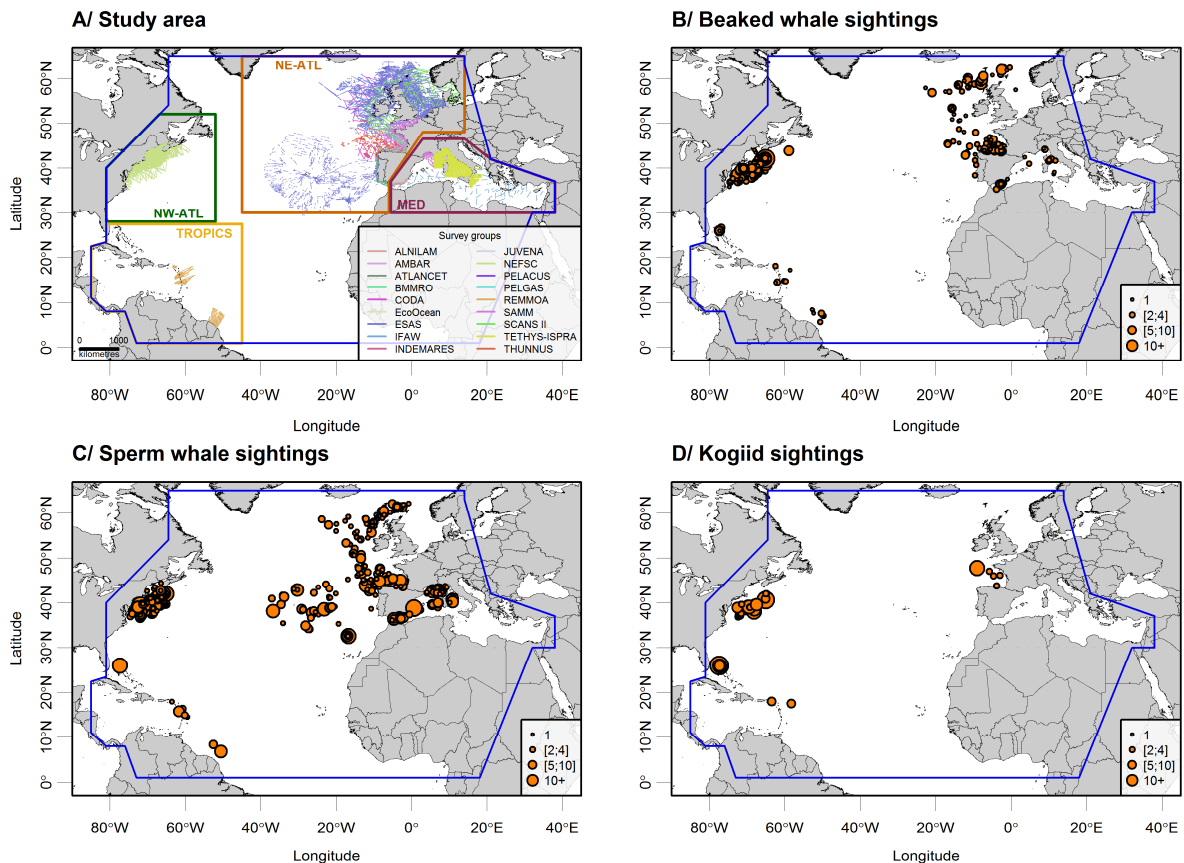
Due to the low sighting rates usually reported for these species, each individual survey can rarely provide sufficient sightings to model their habitat preferences (Waring et al. 2001; Barlow et al. 2006; Kiska et al. 2007), particularly at a large scale representing a major challenge for their conservation. To address this issue, we merged datasets from different visual surveys conducted in the North Atlantic Ocean and the Mediterranean Sea to increase the available number of sightings and model habitats used by deep-diving cetaceans and thus, understand the environmental processes that drive their basin-wide distribution.

Data-assembling is often necessary to successfully model habitat preferences of cetaceans (Roberts et al. 2016; Mannocci et al. 2017, Rogan et al. 2017) but requires methodological considerations. Due to the various protocols, platforms and observation heights, the species detection capacity and data quality vary depending on the survey. Each survey does not collect the same information, particularly regarding the observation conditions, some surveys record only the Beaufort seastate while other surveys also record other parameters that influence the species detection, such as the sun glare, the cloud coverage or the wave height. Consequently, the homogenisation of these different data may require levelling to the coarsest commonalities across datasets (i.e. data degradation). Moreover, because surveys are carried out in different years and seasons, spatial heterogeneity in the data could be an issue when the studied species have variable distributions over time. Our study aims to maximise the number of analogous datasets available from different surveys in order to model the habitats of deep-diving cetaceans at a large scale. In this work, we have aggregated cetacean visual survey datasets collected in the North Atlantic Ocean and the Mediterranean Sea. To take into account the various protocols we implemented a meta-analysis of the detection process across platforms and observation conditions and modelled densities of three groups of deep-diving cetaceans in a Generalised Additive Model framework. Finally, we performed a gap analysis (Jennings 2000) to assess the reliability of the predictions outside the surveyed area. We thus produced the first basin-wide density maps for deep-diving cetaceans in the North Atlantic Ocean and the Mediterranean Sea.

## C.2. METHODS

### C.2.1 Study area

We defined the study area as the North Atlantic and Mediterranean basins from the Guiana Plateau to Iceland, *i.e.* approximately from 1–65°N, excluding the Baltic, Red and Black Seas, the Gulf of Mexico and the Hudson Bay (Fig. C.1A).



**Fig. C.2.** The study area (A), and the beaked whale (B), sperm whale (C) and kogiid (D) sightings recorded during the surveys. The blue polygon delineates the study area. Surveys were carried out along transects (lines) following a line-transect methodology (details of the surveys in Appendix C.1). Sightings were classified by group sizes with each point representing one group of individuals and point size relating to the number of animals in a group.

In the North Atlantic Ocean, the global thermohaline circulation is characterised by the formation of deep saline and cold water masses flowing from the Labrador basin and Greenland Sea southward and warmer surface waters generally flowing northward but affected by a large gyre (Levin and Gooday 2003; Tomczak and Godfrey 2003). This subtropical gyre is delimited by the North Atlantic, Azores and Canary Currents in the east and the North Equatorial Current and Gulf Stream in the west. The latter is narrower and swifter than its eastern counterparts and follows the continental slope off Florida to the Grand Banks of Newfoundland. Large seasonal variations of the wind field (except in the subtropical area), high salinity and a general gradient of temperatures from west to east (about 8°C difference) are characteristics of the North Atlantic Ocean (Tomczak and Godfrey 2003). Within this ocean, primary production is quite low in the tropical zone, and varies seasonally with maximum productivity in winter

in the subtropical zone, a spring bloom and oligotrophic summer conditions at mid-latitude and maximum productivity in summer in the subpolar zone (Campbell and Aarup 1997).

The Atlantic Ocean is connected to the Mediterranean Sea through the Strait of Gibraltar. Fresh Atlantic waters flows into the Mediterranean forming the upper layer and generates gyres and eddies along the Mediterranean coasts, while a deeper layer of dense Mediterranean waters outflows into the Atlantic Ocean (Pinardi and Masetti 2000; Tomczak and Godfrey 2003). The Mediterranean Sea is a generally oligotrophic sea characterised by longitudinal gradients of density, salinity and temperature resulting in a gradually decreasing primary production from west to east (Bethoux et al. 1999; Longhurst 2007).

### C.2.2 Data origin

We aggregated visual shipboard and aerial surveys performed by 13 independent organisations in the North Atlantic Ocean and the Mediterranean Sea between 1998 and 2015 (details of the surveys in Appendix C.1). Cetacean sightings were recorded following line-transect methodologies that allow Effective Strip Width (ESW) to be estimated from the measurement of the perpendicular distances to the sightings (Buckland et al. 2015).

To account for the difficulty to identify individuals at species level (*e.g.* genera *Mesoplodon*, *Kogia*), we pooled species into three groups: (1) the beaked whales, consisting of Cuvier's beaked whales (*Ziphius cavirostris*), the mesoplodonts (*Mesoplodon* spp.) and the northern bottlenose whale (*Hyperoodon ampullatus*), (2) the sperm whales (*Physeter macrocephalus*), and (3) the kogiids, including the pygmy (*Kogia breviceps*) and dwarf sperm whales (*K. sima*).

A total of 630 sightings of beaked whales, 836 sightings of sperm whales and 106 sightings of kogiids, mainly distributed in the northeast and northwest Atlantic Ocean (north of the 35°N latitude), were assembled for the present study (Fig. C.1B-D). Aggregated effort data represented about 1,240,300 km of on-effort transects (*i.e.* following a transect at specified speed/altitude with a specified level of visual effort) of which 58% was carried out by plane (Fig. C.1A, Table C.1). Only 9% of the effort was conducted under Beaufort seastate > 4 and these data were removed from the analyses. Even if it is difficult to detect beaked whales and kogiids with a Beaufort seastate equal to 4, it was a trade-off between keeping a maximum number of data and limiting biases related to detection.

To account for differences between surveyed regions, four sub-regions were defined (Table C.1): the northeast Atlantic Ocean (NE-ATL; from 40°W-10°E and 36°N-65°N), the northwest Atlantic Ocean (NW-ATL; from 80°W-55°W and 30°N-50°N), the tropics (from 80°W-45°W and 1°N-28°N) and the Mediterranean sea (MED; from 5°W-40°E and 30°N-46°N). Most of sampling effort was performed in the northeast (37 %) and northwest (45 %) Atlantic Ocean. Mediterranean surveys represented only 16 % of the total sampling effort and surveys near the tropics represented only 2 %. Surveys were mostly carried by plane (58 %) than by boat (42 %).

Encounter rates were calculated in each sub-region as:

$$(\text{number of encounters} / \text{total distance travelled}) * 100.$$

**Table C.1. Effort performed by platform type or Beaufort seastate for all the surveys.** For the analyses, all segments with Beaufort seastate > 4 were excluded. ‘NE-ATL’ means northeast Atlantic Ocean; ‘NW-ATL’ means northwest Atlantic Ocean and ‘MED’ means Mediterranean Sea.

Sectors	Total survey effort (km and %)	Total aerial effort (km and %)	Total shipboard effort (km and %)	Total effort by Beaufort seastate class (km and %)				
				[0-1]	]1-2]	]2-3]	]3-4]	]4-7]
NE-ATL	468,892	70,358	398,533	76,705	118,456	135,699	84,812	53,220
	37 %	15 %	85 %	16 %	25 %	30 %	18 %	11 %
NW-ATL	556,963	545,677	11,286	42,737	121,184	199,317	131,947	61,777
	45 %	98%	2 %	8 %	22 %	36 %	23 %	11 %
MED	195,440	86,930	108,510	92,126	69,882	26,649	5,984	799
	16 %	44 %	56 %	47 %	36 %	14 %	3 %	0.4 %
TROPICS	19,041	15,356	3,685	10,590	2,495	3,681	1,897	378
	2 %	81 %	19 %	56 %	13 %	19 %	10 %	2 %
<b>STUDY AREA</b>	<b>1,240,336</b>	<b>718,321</b>	<b>522,014</b>	<b>222,158</b>	<b>312,017</b>	<b>365,346</b>	<b>224,640</b>	<b>116,174</b>
		<b>58 %</b>	<b>42 %</b>	<b>18 %</b>	<b>25 %</b>	<b>30 %</b>	<b>18 %</b>	<b>9%</b>

### C.2.3 Data processing

#### Data-assembling

All survey datasets were standardised for units and formats (*e.g.* date, time and coordinates) and aggregated into a single common dataset. A specific coordinate projection encompassing the entire survey area was defined (Albers equal-area conic from <http://projectionwizard.org>). Effort data were linearized and discretised into 5 km segments using ArcGIS 10.3 (ESRI 2016) and the Marine Geospatial Ecology Tools software (Roberts et al. 2010). Because of the large disparity between aerial and shipboard surveys, aerial surveys had transect lengths of up to several tens of kilometres, while transects in shipboard surveys could be much shorter, the 5 km segment length was the value that best homogenised the various transect lengths of the different surveys. Finally, sightings were linked to the segments for each species group.

#### Environmental variables

We used static and dynamic variables that are believed to influence the distributions of deep-divers (Table C.2). All variables were resampled at a 0.25° resolution (about 27-28 km depending on the latitude) instead of a 5 km resolution that would match the 5 km effort segments, because of the very large size of the study area and the spatial resolution of the variables (Table C.2). This implies that same values of environmental variables are attributed to neighbouring segments. However, considering the total extent of the sampling effort, it is a considerable undertaking to scan a large range of environmental variables values.

Depth, slope and the surface of canyon and seamount habitats within each 0.25° cell are physiographic variables. Sea Surface Temperature (SST; mean, standard error and spatial gradients, calculated as the difference between the minimum and maximum SST values found in the eight pixels surrounding any given pixel of the grid), Sea Surface Height (SSH; mean and standard error) and Eddy Kinetic Energy (EKE; mean and standard error) are dynamic oceanographic variables related to the movements of water masses. Net Primary Production (NPP) is a biological variable used as a proxy of

prey availability (Appendix C.2 shows the maps of the averaged situation of the variables over the 18 years of surveys). Dynamic variables were computed at a monthly resolution, *i.e.* averaged over the 29 days prior to each sampled day to avoid gaps in remote sensing oceanographic variables and to take into account the time-lag between an environmental condition and its effect on intermediate trophic levels (Jaquet 1996; Austin et al. 2006; Redfern et al. 2006; Cotté et al. 2009).

**Table C.2. Candidate environmental predictors used for the habitat modelling.** All variables were resampled at a 0.25° resolution. A: Depth and slope were derived from GEBCO-08 30 arc-second database (<http://www.gebco.net/>); 30 arc-second is approximately equal to 0.008°. B: Surface per cell was calculated in ArcGIS 10.3 from the shapefile of canyons and seamounts provided by Harris et al. (2014). C: The mean, standard error and gradient of Sea Surface Temperature (SST) were calculated from the GHRSSST Level 4 CMC SST v.2.0 (Canada Meteorological Centre, <https://podaac.jpl.nasa.gov/dataset/CMC0.2deg-CMC-L4-GLOB-v2.0>). D: The Aviso ¼° DT-MADT geostrophic currents dataset was used to compute mean and standard deviation of Sea Surface Height (SSH) and Eddy Kinetic Energy (EKE; <https://www.aviso.altimetry.fr/en/data/products/sea-surface-height-products/global/madt-h-uv.html>). E: Net primary production (NPP) was derived from SeaWiFS and Aqua using the Vertically Generalised Production Model (VGPM; <http://orca.science.oregonstate.edu/1080.by.2160.8day.hdf.vgpm.m.chl.m.sst.php>).

Environmental variables and units	Original Resolution	Sources	Effects on pelagic ecosystems of potential interest to deep-divers
<b>Physiographic</b>			
Depth (m)	30 arc sec	A	Deep-divers feed on squids and fish in the deep water column
Slope (°)	30 sec arc	A	Associated with currents, high slope induce prey aggregation or enhanced primary production
Surface of canyons and seamounts in a 0.25° cell (km <sup>2</sup> )	30 sec arc	B	Deep-divers are often associated with canyons and seamounts structures; the variable indicates the proportion of this habitat in each cell
<b>Oceanographic</b>			
Mean of SST (°C)	0.2°, daily	C	Variability over time and horizontal gradients of SST reveal front locations, potentially associated with prey aggregations or enhanced primary production
Standard error of SST (°C)	0.2°, daily	C	
Mean gradient of SST (°C)	0.2°, daily	C	
Mean of SSH (m)	0.25°, daily	D	High SSH is associated with high mesoscale activity and enhanced prey aggregation or primary production
Standard deviation of SSH (m)	0.25°, daily	D	
Mean of EKE (m <sup>2</sup> .s <sup>-2</sup> )	0.25°, daily	D	High EKE relates to the development of eddies and sediment resuspension induce prey aggregation
Standard error of EKE (m <sup>2</sup> .s <sup>-2</sup> )	0.25°, daily	D	
Mean of NPP (mgC.m <sup>-2</sup> .day <sup>-1</sup> )	9 km, 8 days	E	Net primary production as a proxy of prey availability

### Effective Strip Width estimation

From sighting and effort data, we fitted a detection function to determine the ESW for each species group (Thomas et al. 2010; Buckland et al. 2015). The estimation of the ESWs was a key step in the data-assembling process to take into account heterogeneity in effort per segment in the models (Hedley and Buckland 2004). Even if we only considered line-transect survey data, protocols differed to some extent



and datasets did not always provide the same information, in particular regarding the observation conditions. Some surveys recorded Beaufort seastate, cloud coverage, sun glare and subjective observation conditions while others only provided Beaufort seastate. Hence, Beaufort seastate was the only descriptor of observation conditions shared by all datasets. Consequently, the platform type, the observation heights and the Beaufort seastate were used as covariates following the conventional distance sampling methodology (Marques and Buckland 2003; Buckland et al. 2015). In addition, we had not enough sightings to fit detection functions for each survey. Consequently, to take into account the various protocols, we performed a meta-analysis (Gurevitch et al. 2001; Higgins et al. 2009). Firstly, for each species group, truncation distance  $w$  was determined as the 95<sup>th</sup> percentile of the set of perpendicular distances: the 5% most distant sightings were discarded from the analysis. Then, we created classes to pool the different surveys: classes of platform type (plane or boat), observation heights (e.g. 0-5 m; 5-10 m...) and Beaufort seastate (0-1; 1-2; 2-3 and 3-4). The meta-analysis was performed in R-3.3.1 (R Core Team 2016) in a Bayesian framework using JAGS version 4-6 and package rjags (jags model available in Appendix C.3; Plummer 2016). First, for each species group, perpendicular distances of all sightings were used to estimate a detection function with a hazard key.

For a sighting  $i$  made from survey  $s$  at height  $j$  in class of Beaufort seastate  $k$ , let  $d_{jks}^i$  denotes the perpendicular distance. The detection probability of sighting  $i$  is:

$$\begin{cases} p_{ijk}^s = g_s(d_{ijk}) = 1 - \exp\left(-\left(\frac{d_{ijk}}{\sigma_{jks}}\right)^{-\nu_s}\right) \\ \log(\sigma_{jks}) = \beta_{j0} + \beta_{j1} \times k + \alpha_s \end{cases}$$

where  $\beta_{j0}$  and  $\beta_{j1}$  are respectively random intercept and slope parameters for the effect of platform height; and  $\alpha_s$  and  $\nu_s$  are survey random effects. Bivariate random effects were specified with a Cholesky decomposition and using the priors for the Cholesky factors as Kinney and Dunson (2008). We used half Student-t distributions with 3 degrees of freedom and scale set to 1.5 as priors for dispersion parameters, and standard normal priors for all other parameters. Four chains were run with a warmup of 10,000 iterations, followed by another 10,000 iterations (with a thinning factor of 10). Parameter convergence was assessed with the Gelman-Rubin  $\hat{R}$  statistics. Posterior inferences are based on the pooled sample of 4,000 values (1,000 per chain).

The advantage of setting a hierarchical model to estimate detection functions is to borrow strength across the different datasets to increase the precision of estimates. For each combination of survey – platform type – observation height – Beaufort seastate, estimated detection functions are shrunk towards a common detection function (itself estimated from the data) according to the available data corresponding to this particular combination of survey – platform type – observation height – Beaufort seastate. If, for a given combination of parameters, there were few sighting data, the estimated detection function was very close to the common detection function, whereas if there were enough data, the estimated detection function could deviate from this common function. Upon model fitting and successful parameter estimation, the ESW for each combination of survey – platform type – observation height – Beaufort seastate was computed:

$$ESW_{jks} = \int_0^w g_s(x) dx = \int_0^w \left[ 1 - \exp\left(-\left(\frac{x}{e^{\beta_{j0} + \beta_{j1} \times k + \alpha_s}}\right)^{-\nu_s}\right) \right] dx$$

The posterior mean value of estimated ESW was then allocated to each segment with respect to species group, survey, platform type, seastate and observation height class.

### C.2.4 Habitat modelling

To model habitat preferences of deep-divers, we fitted Generalised Additive Models (GAMs; Hastie and Tibshirani 1986; Wood 2006) with a Tweedie distribution to account for over-dispersion (Foster and Bravington 2013) with the 'mgcv' R-package (R-3.3.1. version; Wood 2013). GAMs extend Generalised Linear Models to allow for smooth, nonlinear functions of predictor variables to be determined by observed data rather than by strict parametric relationships (Hastie and Tibshirani 1986; Wood 2006). With its additional parameter  $p$  ( $1 \leq p \leq 2$ ), which determines a Poisson mixture of Gamma distributions, the Tweedie distribution can handle datasets with a large proportion of zeros and model continuous positive data (Dunn and Smyth 2005; Arcuti et al. 2013). This parameter was directly estimated by the 'mgcv' function. We modelled the mean number of individuals per segment  $\mu = E(Y|X_1, \dots, X_n)$  as:

$$\log(\mu) = \alpha + \sum_p f(X_p)$$

where  $f(X_p)$  are non-parametric smooth functions (splines) of the covariates and  $\alpha$  is the intercept (Hastie and Tibshirani, 1986). The response variable was linked to the additive predictors with a log-function. By testing different degree of freedom, a number of four appeared consistent to link the response variable to the environmental variables. An offset equal to segment length multiplied by twice ESW was included (Hedley and Buckland 2004). We removed combinations of variables with Spearman partial correlation coefficients higher than  $|0.7|$ , tested all models with combinations of one to four variables and selected the best model with the lowest mean prediction error determined by a leave-one-out cross validation process (minimum Generalised Cross-Validation score; Wood 2006; Clark 2013). The GCV score results from an internal procedure used for smooth parameter estimation within 'mgcv'. A maximum of four covariates per model was used to avoid excessive complexity of models and difficulty in their interpretation (Mannocci et al. 2014; Virgili et al. 2017).

We performed "year-round" models as the studied species groups showed little or no seasonal variation in their habitats (Hooker et al. 2000; Wimmer and Whitehead 2004; McSweeney et al. 2007) and because we did not have enough data to fit seasonal models. We did not include the year as a factor in the models because a preliminary analysis (not shown) failed to detect a statistically significant (at the 5% level) effect of year on geographical sighting distribution.

Monthly predictions at  $0.25^\circ$  resolution were averaged over the entire time period (1998-2015) to produce maps of mean predicted densities which represent the average expected long-term distributional patterns of the beaked whales, sperm whales and kogiids. We did not attempt to correct predicted densities for availability bias thus predicted densities are relative densities. To lighten the reading, the relative densities will be hereafter labelled as densities. Finally, we provided uncertainty maps by computing the variance around the predictions as the sum of the variance around the mean prediction and the mean of the monthly variances. Then, the coefficient of variation was calculated as:

$$CV = (\sqrt{\text{variance over the survey period}} / \text{mean over the survey period}) * 100.$$

### C.2.5 Environmental space coverage gap analysis

To model the habitats of deep-divers in the North Atlantic and Mediterranean basins, we gathered data from a large region collected over a long period. The cumulative effort was not homogeneous and showed extensive geographical gaps. Therefore, we conducted gap analysis on environmental space coverage to identify areas where habitat models could produce reliable predictions outside the survey

blocks, *i.e.* geographical extrapolation, whilst remaining within the ranges of surveyed conditions for the combinations of covariates selected by the models, *i.e.* areas of environmental interpolation (Jennings 2000).

To do this, we determined the extent of the environmental interpolation (versus extrapolation) obtained by combining the four variables selected by the best models (one analysis was done for each species group). This was calculated by using the convex hull methodology defining effort data as the calibration dataset and climatological predictors (*i.e.* the average situation of each predictor over the 18 years period) for the entire study area as the prediction dataset (King and Zeng 2007; Authier et al. 2016). Here we used climatological predictors instead of monthly predictors to limit computation time (the analysis would have been done for each month and then averaged over 18 years). The convex hull of a set of points is the smallest convex envelope that contains these points, *i.e.* all effort data points described by the selected covariates. If prediction data fall inside the convex hull, they are interpolations while if they fall outside the convex hull, they are extrapolations; prediction made at any interpolation point within study area being considered as a more reliable (less model-dependent) than predictions made at extrapolation points (King and Zeng 2007; Authier et al. 2016).

Due to the large number of data (more than 280,000 points in the calibration dataset), convex hulls were estimated by random sub-sampling with the 'WhatIf' R-package (Stoll et al. 2014). We randomly extracted a fraction of the calibration dataset (10,000 points) to estimate a convex hull and assess environmental extrapolation in the prediction dataset. A combination of climatological predictor values that falls inside the convex hull corresponds to interpolation. The combinations of climatological predictor values that were classified as interpolations were set aside but the other combinations were retained and further tested against another random sample of 10,000 points from the calibration data. This procedure was carried out until the full calibration dataset was examined.

The full procedure was conducted twice. Firstly, what we called 'simple interpolation', considered the full range of sampled variables to identify all points of the whole study area where the actual combinations of environmental variables had been sampled in survey blocks. Secondly, in the 'precautionary interpolation', we arbitrarily applied a 5% precautionary approach, *i.e.* 5% of the extreme values of the sampled variables were removed to include in the interpolation areas only the points whose associated combinations of covariates fell within the 95% core ranges sampled. This allowed the definition of two levels of confidence in the predictions.

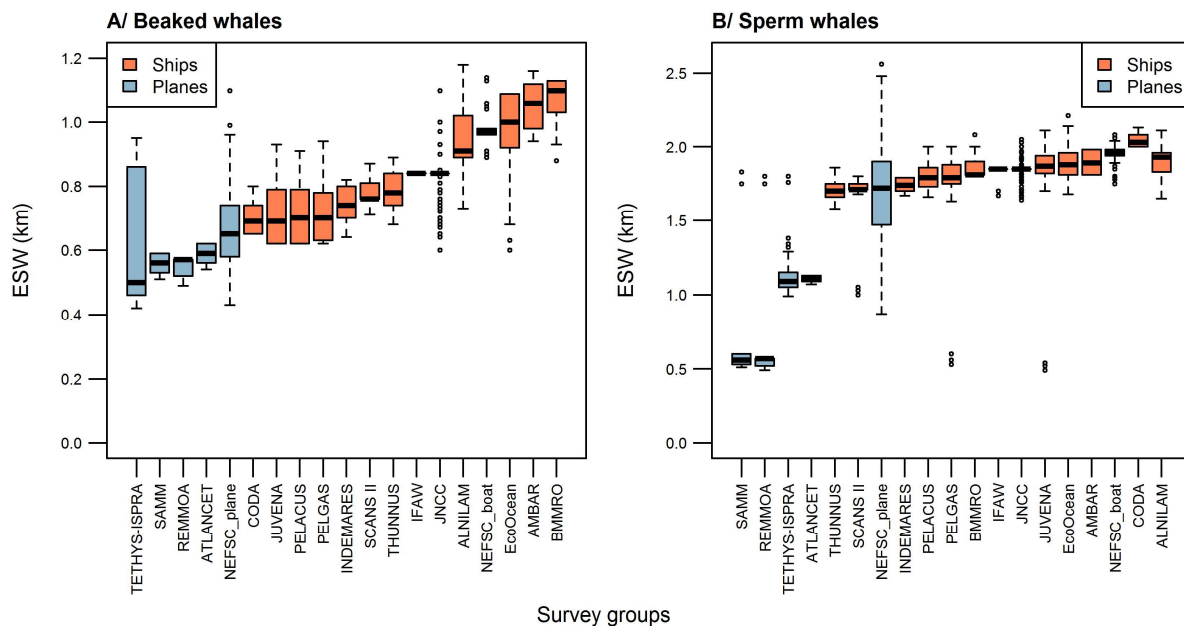
Finally, we produced maps delineating the extent of the simple and the precautionary interpolation areas, and overlaid them to the density prediction maps to highlight areas with a greater reliability.

### C.3. RESULTS

#### Effective strip width

The ESWs estimated with the meta-analysis varied with the surveys and the platform type; they were on average narrower from aerial than shipboard surveys (Fig. C.2). This is probably because aerial observers usually record animals below the plane while shipboard observers look further afield. ESWs were generally larger and more consistent, between surveys using the same platform type, for sperm whales than for beaked whales. There were not enough kogiid sightings to estimate an ESW for each survey particularly in shipboard surveys; consequently we pooled all aerial surveys and estimated an ESW of 1.1 km that we applied to all surveys (shipboard and aerial surveys). The outcomes of this

analysis were consistent with expectations, with a decrease in Beaufort seastate resulting in an increase in ESW estimations (Appendix C.4). Compared to ESW obtained by using the more conventional Distance approach (Appendix C.4; Thomas et al. 2010; Buckland et al. 2015), ESW estimated in the present meta-analysis were shorter and their confidence intervals smaller.



**Fig. C.2. Beaked whale and sperm whale averaged ESWs estimated with the meta-analysis for each survey group and each platform type.** For each survey group, the boxplot represents the extent of estimated ESWs depending on Beaufort seastates and observation heights recorded within the group.

Predictions of the three species groups can be considered as the summer habitats as most sightings were recorded from June–October (84% beaked whale, 76% sperm whales and 77% kogiids). Although effort was almost evenly distributed between the two seasons (53% in the hot season – June to October – and 47% in the cold season – November to May), there were not enough data to fit a model in winter maybe because of the poorer sighting conditions (mean Beaufort seastate was equal to 2.6 in summer and 3.1 in winter).

Overall, encounter rates were very low with 0.05 sightings·100 km<sup>-1</sup> for beaked whales, 0.07 sightings·100 km<sup>-1</sup> for sperm whales and <0.01 sightings·100 km<sup>-1</sup> for kogiids (Table C.3). Highest encounter rates were recorded in the tropics for the three species groups, particularly for the kogiids. There was no sighting of kogiids in the Mediterranean Sea.

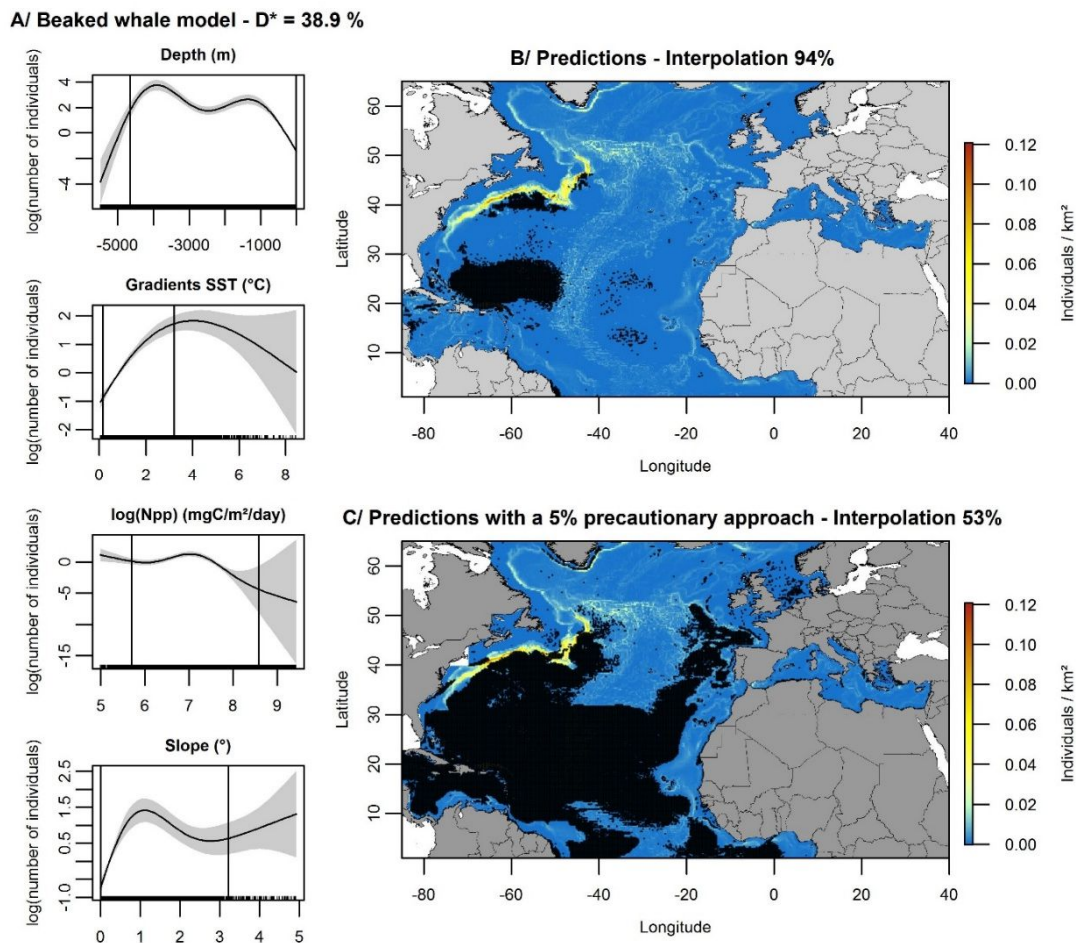
**Table C.3. Encounter rates in sightings·100 km<sup>-1</sup> calculated for the entire study area and each sub-region.** ‘NE-ATL’ means northeast Atlantic Ocean; ‘NW-ATL’ means northwest Atlantic Ocean and ‘MED’ means Mediterranean Sea.

	NE-ATL	NW-ATL	MED	TROPICS	STUDY AREA
Beaked whales	4.2 × 10 <sup>-2</sup>	5.8 × 10 <sup>-2</sup>	3.5 × 10 <sup>-2</sup>	2.2 × 10 <sup>-1</sup>	5.1 × 10 <sup>-2</sup>
Sperm whales	5.7 × 10 <sup>-2</sup>	6.7 × 10 <sup>-2</sup>	9.0 × 10 <sup>-2</sup>	9.5 × 10 <sup>-2</sup>	6.7 × 10 <sup>-2</sup>
Kogiids	1.3 × 10 <sup>-3</sup>	1.0 × 10 <sup>-2</sup>	0.0	2.3 × 10 <sup>-1</sup>	8.5 × 10 <sup>-3</sup>

### Beaked whales

The beaked whale model accounted for 38.9% of the deviance (Fig. C.3A). Depth, spatial gradients of SST, NPP and slope were the variables that most influenced the habitats of beaked whales. Highest densities were predicted for two depth ranges (*ca.* 1,500 m and *ca.* 4,000 m), high slopes (*ca.* 1°), high

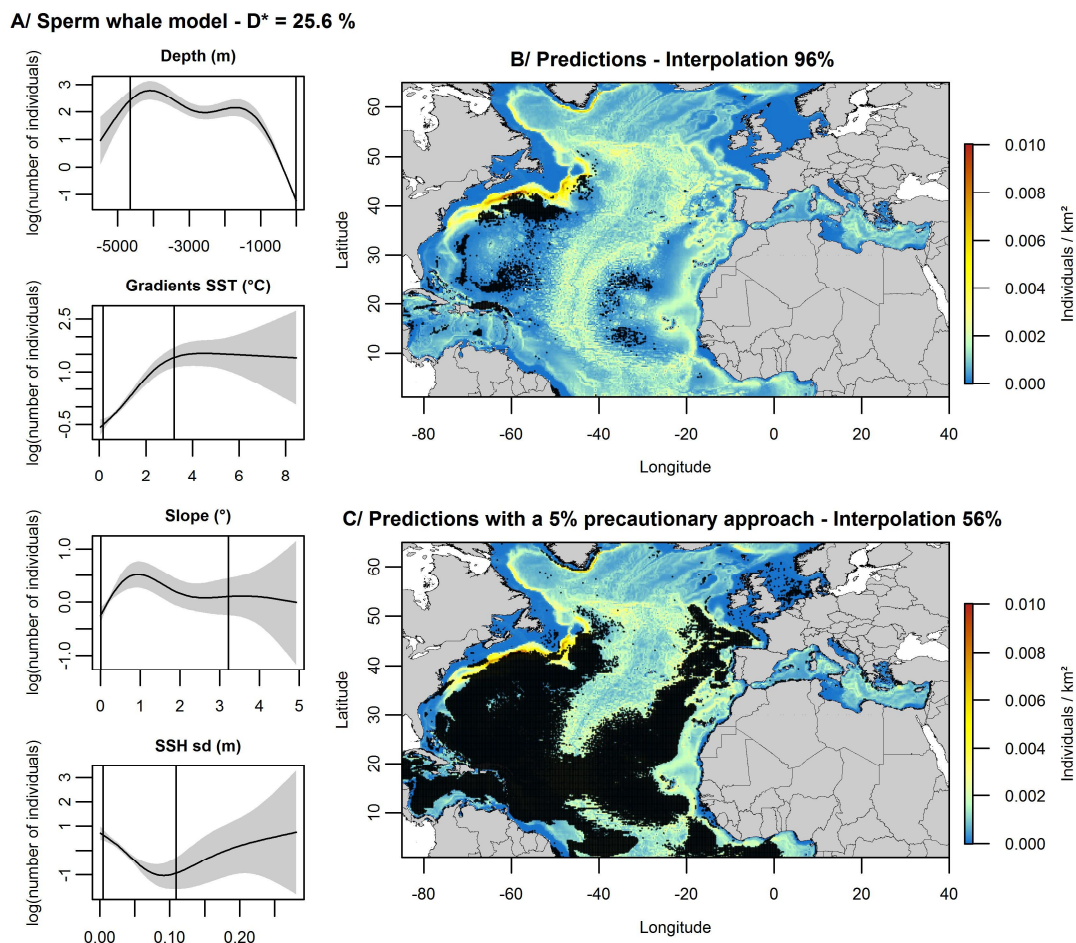
gradients of SST (*ca.* 3°C) and medium productivity (*ca.* 1,100 mgC.m<sup>-2</sup>.day<sup>-1</sup>). This resulted in a concentration of individuals along steep slope areas associated with high depths, with highest densities predicted on the western side of the Atlantic Ocean (Fig. C.3B). In the Mediterranean Sea, predicted densities were lower than in the Atlantic Ocean with highest densities predicted in the Alboran Sea, near the Gibraltar Strait, in the north of the Levantine basin, between Cyprus and Crete, and along the continental slopes. No individuals were predicted near the Tunisia or in the north of the Adriatic Sea (Fig. C.3B). The gap analysis identified areas where the combination of the four variables selected by the best model had not been sampled, resulting in an absence of prediction in 6% of the sampled area, *i.e.* 94% of the sampled area was available for interpolation (Fig. C.3B). However, the precautionary interpolation area obtained by retaining the 95% core distribution of the environmental variables represented only 53% of the study area (Fig. C.3C), mostly because sampling effort in the open oceanic was insufficient to predict with confidence densities in the entire study area, particularly in the centre of the Atlantic Ocean. Coefficients of variation were higher in shallow waters associated with high gradients of SST, where beaked whales have not been reported by any surveys (Appendix C.5A).



**Fig. C.3. Functional relationships for the selected variable (A) and the predicted relative densities of beaked whales in individuals.km<sup>-2</sup> (B and C).** A: Solid lines are the estimated smooth functions, and the shaded regions represent the approximate 95% confidence intervals. The y-axes indicate the number of individuals on a log scale, where zero indicates no effect of the covariate. The vertical lines indicate the 2.5<sup>th</sup> and 97.5<sup>th</sup> quantiles of the data. Black areas on prediction maps (B: without precautionary approach and C: with a 5% precautionary approach) represent zones where we did not extrapolate the predictions. Percentages represent the proportion of the study area defined as interpolation with the gap analysis.

## Sperm whales

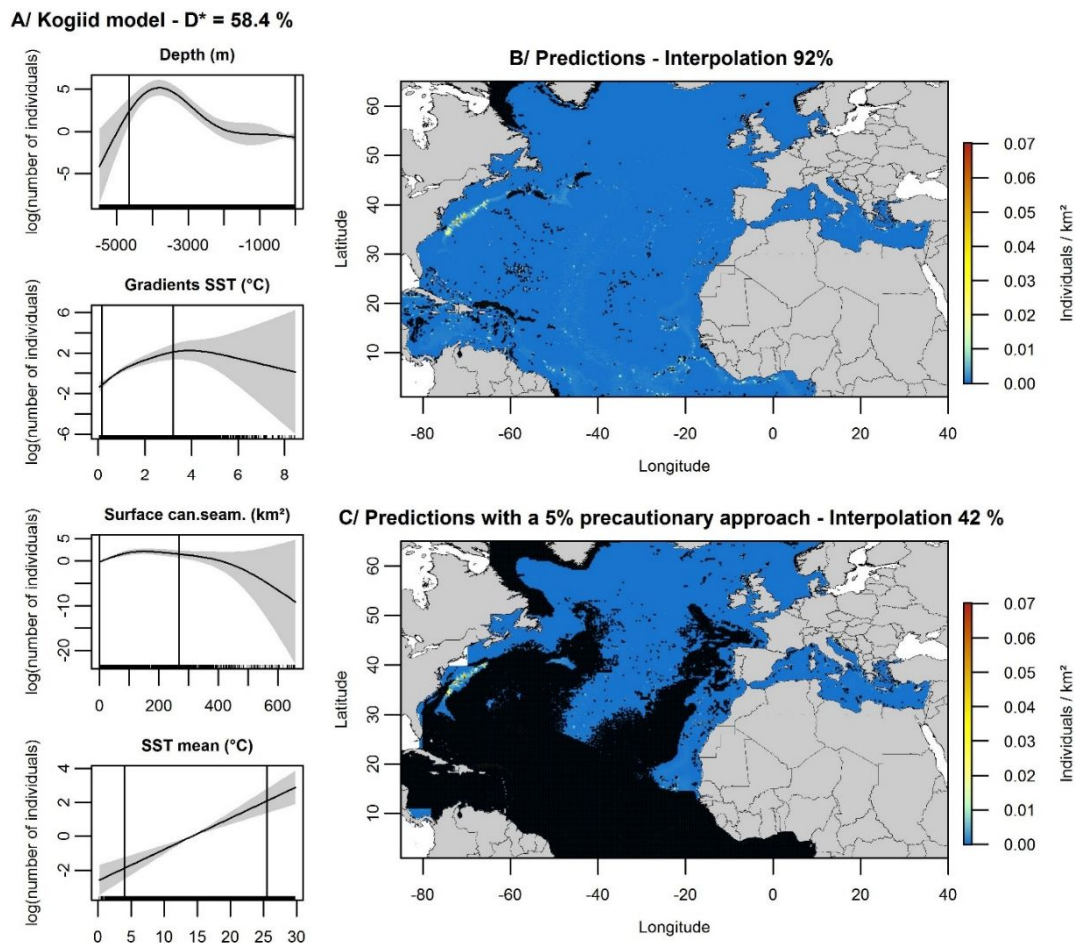
The explained deviance of the sperm whale model was 25.6% (Fig. C.4A). As for beaked whales, depth, spatial gradients of SST and slope were the variables that most influenced the habitats of the species group, complemented by the standard deviation of SSH. Densities of sperm whales were predicted to increase in deep waters associated with steep slopes and high gradients of SST. The predicted habitats of sperm whales were more homogenous than for beaked whales, since the former appeared less restricted to slope areas (Fig. C.4B). Highest densities were also predicted on the western side of the Atlantic basin, along the Gulf Stream. As for beaked whales, predicted densities of sperm whales were lower in the Mediterranean Sea than in the Atlantic Ocean. Highest densities were predicted in the north of the Levantine basin, between Cyprus and Crete and fairly evenly predicted between the continental slopes and the oceanic waters, except near the Tunisia and in the north of Adriatic Sea (Fig. C.4B). Only 4% of the study area (Fig. C.4B) corresponded to values of the selected covariates that had not been sampled during the surveys, but predictions within the core range of covariates only covered 44% of the study area. In fact, the highest predicted densities were partly outside this confidence zone (Fig. C.4C). Coefficients of variation were highest in non-sampled areas where uncertainty was therefore greatest (Appendix C.5B).



**Fig. C.4. Functional relationships for the selected variable (A) and the predicted relative densities of sperm whales in individuals.km<sup>-2</sup> (B and C).** A: Solid lines are the estimated smooth functions, and the shaded regions represent the approximate 95% confidence intervals. The y-axes indicate the number of individuals on a log scale, where zero indicates no effect of the covariate. The vertical lines indicate the 2.5<sup>th</sup> and 97.5<sup>th</sup> quantiles of the data. Black areas on prediction maps (B: without precautionary approach and C: with a 5% precautionary approach) represent zones where we did not extrapolate the predictions. Percentages represent the proportion of the study area defined as interpolation with the gap analysis.

## Kogiids

The kogiid model accounted for 58.4% of the explained deviance (Fig. C.5A). Depth, spatial gradients of SST, surface of canyon and seamount habitat per cell and mean SST were the variables that most influence the habitats of kogiids. Highest densities were predicted in warm and deep waters associated with fronts and canyons or seamounts habitats. Consequently, individuals were not predicted in the northern part of the study area but mainly along the Gulf Stream where fronts and canyons are abundant (Fig. C.5B). Low densities were predicted in the Mediterranean Sea although no individuals were sighted (Fig. C.5B). Because SST was among the selected covariates, 7% of the study area was classified as extrapolation zone (Fig. C.5B). As there was little sampling effort in tropical and sub-polar regions, extreme temperature values were less sampled, resulting in a smaller prediction confidence zone for kogiids than for other species groups. The precautionary interpolation area, based on the 95% core distribution of the covariates' ranges, was reduced to 42% of the study area (Fig. C.5C). Coefficients of variation were the highest in shallow waters and Mediterranean Sea, where kogiids have not been reported by any surveys (Appendix C.5C).



**Fig. C.5. Functional relationships for the selected variable (A) and the predicted relative densities of kogiids in individuals.km<sup>-2</sup> (B and C).** A: Solid lines are the estimated smooth functions, and the shaded regions represent the approximate 95% confidence intervals. The y-axes indicate the number of individuals on a log scale, where zero indicates no effect of the covariate. The vertical lines indicate the 2.5<sup>th</sup> and 97.5<sup>th</sup> quantiles of the data. Black areas on prediction maps (B: without precautionary approach and C: with a 5% precautionary approach) represent zones where we did not extrapolate the predictions. Percentages represent the proportion of the study area defined as interpolation with the gap analysis.

## C.4. DISCUSSION

Deep-divers are species characterised by low sightings rates; and modelling their habitats is particularly challenging. Our study merged different surveys to capitalise on more than 1,240,000 km of sighting effort deployed over the North Atlantic Ocean and the Mediterranean Sea in the past two decades. This data-assembling endeavour required taking into account the different protocols or platform types and therefore the different species detection capacity depending on the surveys. We then investigated the habitats of deep-divers using state-of-the-art statistical methods with a focus on how much confidence can be given to predictions. The habitats of deep-diving cetaceans were mainly influenced by static environmental variables such as depth or slope as well as spatial gradients of temperatures, revealing density hotspots in the western North Atlantic Ocean.

### C.4.1 Methodological considerations

Over the past few years, data-assembling has been increasingly used for the study of top marine predators (Winiarski et al. 2014; Roberts et al. 2016; Mannocci et al. 2017; Rogan et al. 2017). Due to the very low sighting rates of deep-diving cetaceans, taken separately each survey could not provide sufficient data to model the habitats of rare species, thus data-assembling was necessary. However, in contrast to Rogan et al. (2017), we did not assemble data collected with similar protocols but data collected with different protocols which implied to homogenise and somehow to degrade the data before developing a single spatial model. At a time when shared databases are becoming increasingly important (*e.g.* OBIS SEAMAP - <http://seamap.env.duke.edu/>, EMODnet - <http://www.emodnet.eu/>), a sharing of standardised observation protocols would be useful to facilitate data-assembling, would allow to less degrade the data and help describing the large-scale habitats of the species.

Winiarski et al. (2014) warned how data collected in different surveys must be checked for compatibility, especially with respect to segment sizes. In this study, the large disparity between transects of aerial and shipboard surveys (from about two km to hundreds of kilometres) required to format the data in small segments of 5 km for modelling. The 5 km data format could induce a mismatch with covariate resolution, which was coarser due to the vast extent of the study area. However, this mismatch turned to be limited.

Regarding used environmental data, most of the oceanographic variables used here were related to phenomena in the euphotic zone (upper layer). This is because most of environmental variables are based on satellite data and variables that describe deep water column are difficult to obtain. As deep-diving cetacean spend most of their time in depth (Perrin et al. 2009), the use of surface variables might lead to a mis-interpretation of species habitats. Indeed, using proxies related to surface waters, we may not have identified the true causal relationships that explain the habitats but indirect relationships (Austin 2002). However, explained deviances of the models were high (from 25.6% to 58.4%) and coefficients of variation were the highest on the continental shelf (Appendix C.5), where deep-divers are known to be mostly absent (Waring et al. 2001; Cañadas and Vázquez 2014; Arcangeli et al. 2015; Roberts et al. 2016). This indicated a good effectiveness of the models to make coherent predictions despite indirect variables. Nevertheless, the very high deviance of the kogiid model (58.4%) might indicate some level of model over-fitting due to the small number of data, even if predictions were consistent with the known ecology of the species group (McAlpine 2009).



By assembling data collected in different regions (*e.g.* Mediterranean Sea and Atlantic Ocean or northwest and northeast Atlantic Ocean) we assumed similar relationships of deep-divers with their habitat throughout the multiple ecosystems. However these ecosystems are very different with an active frontal system associated with the Gulf Stream in the western Atlantic Ocean (Tomczak and Godfrey 2003) or an oligotrophic Mediterranean Sea (Bethoux et al. 1999). Consequently, deep-diver habitats may be different between regions with for example, a possible greater influence of thermal fronts on species habitat in the western Atlantic Ocean than in the eastern Atlantic Ocean or the Mediterranean Sea. Indeed, Roberts et al. (2016) evidenced in the western Atlantic Ocean an influence of the depth, the distance to fronts and to eddies on the habitats of the three species groups. On the other hand, both Cañadas and Vázquez (2014) and Rogan et al. (2017) found depth one of the most important predictor of deep-diver habitat in the Mediterranean Sea and in the northeast Atlantic Ocean respectively. This suggests a consistency in the habitats of these species groups, which are highly associated with topographic features. Consequently, a data-assembling at such a large scale seem consistent. However, bimodal response to depth for beaked whales and sperm whale with peaks of densities predicted at 1,500 m and 4,000 m might reveal different habitats, the species groups probably use different habitats to forage. The 1,500 m peak was essentially made up of sightings from the Mediterranean Sea while the 4,000 m peak was essentially made up of sightings from the northwest and northeast Atlantic Ocean (Fig. C.1; Appendix C.2). A model for each ecoregion or the inclusion of an interaction with ecoregion in the model could help determining whether the variables selected by the different models would be identical.

#### C.4.2 Large-scale deep-diver habitats

Physiographic variables were highly predictive of deep-diver habitats. In the three models, depth was one of the most influential variable. The surface of canyon and seamount habitats per cell was a significant variable only for the kogiids. This is consistent with the influence of topographic features noticed in smaller regions (Fergusson et al. 2006; MacLeod et al. 2011; Whitehead 2013; Wong and Whitehead 2014). Oceanographic variables were also important. For each species group, spatial gradients of SST significantly contributed to the models. Deep-divers seemed to concentrate in areas of strong gradients such as thermal fronts in which prey aggregate (Brandt 1993; Bost et al. 2009; Woodson and Litvin 2015). Hence, the Gulf Stream, which is the most active frontal zone in the study area compared to the eastern boundary currents that are broader and much slower, may explain the high densities of deep-divers on the western side of the North Atlantic Ocean (Griffin 1999; Waring et al. 2001; Hamazaki 2002; Roberts et al. 2016).

In our study, we geographically extrapolated the deep-diver habitats to the entire North Atlantic Ocean and Mediterranean Sea while keeping within sampled environmental conditions (environmental interpolation). At a local scale, predictions were consistent with known distributions. As Cañadas and Vázquez (2014), we identified a beaked whale density hotspot in deep waters of the Alboran Sea but our predicted densities were lower and more extended towards the Gibraltar Strait. In our predictions, the Tyrrhenian and Ligurian Seas also appeared as suitable habitats for beaked whales, consistent with the results of Arcangeli et al. (2015) and Lanfredi et al.'s (2016). In addition, recorded strandings of Cuvier's beaked whale along the coasts of the Ligurian and Ionian Seas and the eastern coasts of the Mediterranean Sea (Podestà et al. 2006) revealed the presence of the species group close to these coasts, as suggested by our predictions. For sperm whales, our predictions agreed with Praca and

Gannier's (2008) results with potential habitats predicted on the continental slope off France and off islands of the western Mediterranean Sea. Sperm whales codas recorded in the Ligurian, Tyrrhenian and Ionian Seas (Pavan et al. 2000) reveal the presence of the species in these areas, as suggested by our predictions. In the Bay of Biscay, highest densities of beaked whales and sperm whales were predicted along the slope, consistent with encounter rates estimated from platforms of opportunity (Kiszka et al. 2007) and abundances predicted from shipboard and aerial surveys (Rogan et al. 2017). On the west Atlantic Ocean, our models predicted highest densities of beaked whales and sperm whales along the continental slope consistently with Roberts et al. (2016) but predicted densities were lower (about two times lower). Concerning the kogiids, there is little published literature allowing predictions to be compared. In the North West Atlantic Ocean, kogiids were predicted in warm deep waters, which was consistent with their known ecology (McAlpine 2009) and the patterns of distribution predicted by Mannocci et al. (2017) except that no individual was predicted off the coast of Florida. However this area was an extrapolation on the precautionary approach.

In the present study from 92 to 96% of the study area, with no precautionary approach, and from 44 to 58% of the study area, with a 5% precautionary approach, were classified as confident predictions. Large gaps in environmental space coverage were revealed, especially in deeper waters of the central north Atlantic gyre and in tropical waters. It can be noted that areas of interest for deep-divers were predicted at the margin of the precautionary interpolation zone in particular because deeper waters and steeper slopes were within the upper 2.5% quantiles of aggregated survey coverage for these two physiographic covariates. This suggested that sampling effort was not sufficient in deeper and steeper areas and more intensive sampling effort performed in these areas could help to better describe habitats used by deep-divers. Meanwhile, the predicted habitats provided in this study could be included in a marine spatial planning. This consists in analysing and allocating the spatial and temporal distribution of human activities in marine areas, here for example anthropogenic sound, to achieve ecological objectives, such as species conservation (Douve 2008). Thus, with this methodology, the conservation of deep-diving cetaceans could be improved.

## C.5. CONCLUSION

Modelling rare species habitats is particularly challenging because habitat models require large datasets yet rare species typically yield low numbers of sightings. As a result, assembling datasets appeared to be an appropriate strategy to model the large scale habitats of deep-divers, the beaked whales, sperm whales and kogiids, across the North Atlantic and the Mediterranean basins.

At a local scale, predicted habitats were consistent with previous studies. Predictions at a larger scale highlighted a gradient of predicted densities (with highest densities predicted on the western side of the study area) which would not have been evident at a local scale and showed pronounced influence of active frontal zones, such as the Gulf Stream, on the habitats of these species groups. Even though gaps remain at such a large scale, we were able to predict the habitats of these species groups throughout the Atlantic basin and thus identify potential habitats, even in non-sampled areas. In addition, due to the large extent of the study area, a prediction relevance assessment was needed. Through an environmental space coverage gap analysis, we identified areas in tropical and deep oceanic waters where sampling effort was insufficient and need to be intensified to increase prediction reliability. Finally, by developing a data-assembling procedure that could be applied to any species and to any local or extended study area, we helped to improve the knowledge of deep-diver distribution.

## Acknowledgments

We are grateful to the many observers who participated in the surveys and collected all the data but also the ships' captains, crews and pilots. We thank Phil Hammond and his team for providing SCANS and CODA survey data. THUNNUS survey was carried out thanks to the collaboration of the General Directorate of Fisheries and Maritime Affairs, Government of Galicia. We thank the *Direction Générale de l'Armement* (DGA), including Odile Gérard and Carole Nahum, for funding Auriane Virgili's doctoral research grant. ML was funded by a Ramón y Cajal (RYC-2012-09897) postdoctoral contract of the Spanish Ministry of Economy, Industry and Competitiveness, whereas IGB was supported by a PhD fellowship (FPI, BES-2014-070597) of the Spanish Ministry of Economy, Industry and Competitiveness. This study is a contribution to the CHALLENGES (CTM2013-47032-R) project of the Spanish Ministry of Economy, Industry and Competitiveness. EcoOcéan Institut acknowledge its partners for the participation to the collection of data at sea: École Pratique des Hautes Études, WWF-France, Swiss Cetacean Society, Cybelle Planète, Participe Futur and Fondation Nicolas Hulot.

## References

- Arcangeli, A., Campana, I., Marini, L., MacLeod, C.D. (2015). Long-term presence and habitat use of Cuvier's beaked whale (*Ziphius cavirostris*) in the Central Tyrrhenian Sea. *Marine Ecology* 37: 269–282.
- Arcuti, S., Calulli, C., Pollice, A., D'Onghia, G. (2013). Spatio-temporal modelling of zero-inflated deep-sea shrimp data by Tweedie generalized additive. *Statistica* 73: 87–101.
- Austin, D., Bowen, W.D., McMillan, J.I., Iverson, S.J. (2006). Linking Movement, Diving, and Habitat to Foraging Success in a Large Marine Predator. *Ecology* 87: 3095–3108.
- Authier, M., Saraux, C., Péron, C. (2016). Variable selection and accurate predictions in habitat modelling: a shrinkage approach. *Ecography* 40: 549–560.
- Balcomb, K.C., and Claridge, D.E. (2001). A mass stranding of cetaceans caused by naval sonar in the Bahamas. *Bahamas Journal of Science* 8: 2–12.
- Barlow, J., Ferguson, M., Perrin, W., Gerrodette, T., Joyce, G., Macleod, C., Mullin, K., Palka, D., Waring, G. (2006). Abundance and densities of beaked and bottlenose whales (family *Ziphiidae*). *Journal of Cetacean Research and Management* 7: 263–270.
- Bethoux, J.P., Gentili, B., Morin, P., Nicolas, E., Pierre, C., Ruiz-Pino, D. (1999). The Mediterranean Sea: A miniature ocean for climatic and environmental studies and a key for the climatic functioning of the North Atlantic. *Progress in Oceanography* 44: 131–146.
- Bost, C.A., Cotté, C., Bailleul, F., Cherel, Y., Charrassin, J.B., Guinet, C., Ainley, D.G., Weimerskirch, H. (2009). The importance of oceanographic fronts to marine birds and mammals of the southern oceans. *Journal of Marine Systems* 78: 363–376.
- Brandt, S.B. (1993). The effect of thermal fronts on fish growth: a bioenergetics evaluation of food and temperature. *Estuaries* 16: 142–159.
- Brownell, Robert L., Jr., Yamada, T., Mead, James G. Helden, A. L. (2004). Mass stranding of Cuvier's beaked whales in Japan: U.S. Naval acoustic link? *Journal of Cetacean Research and Management* 7: 1-10.
- Buckland, S. T., Rexstad, E. A., Marques, T. A., Oedekoven, C. S. (2015). *Distance sampling: methods and applications*. Springer.

- Campbell, J.W., and Aarup, T. (1992). New production in the North Atlantic derived from seasonal patterns of surface chlorophyll. *Deep Sea Research Part A. Oceanographic Research Papers* 39: 1669–1694.
- Canada Meteorological Center. (2012). GHRST Level 4 CMC0.2deg Global Foundation Sea Surface Temperature Analysis (GDS version 2). Ver. 2.0. PO.DAAC, CA, USA.
- Cañadas, A., and Vázquez, J.A. (2014). Conserving Cuvier's beaked whales in the Alboran Sea (SW Mediterranean): Identification of high density areas to be avoided by intense man-made sound. *Biological Conservation* 178: 155–162.
- Carrillo, M., and Ritter, F. (2010). Increasing numbers of ship strikes in the Canary Islands: proposals for immediate action to reduce risk of vessel-whale collisions. *Journal of Cetacean Research and Management* 11(2): 131–138.
- Clark, M. (2013). Generalized additive models: getting started with additive models in R. Center for Social Research, University of Notre Dame, 35.
- Cotté, C., Guinet, C., Taupier-Letage, I., Mate, B., Petiau, E. (2009). Scale-dependent habitat use by a large free-ranging predator, the Mediterranean fin whale. *Deep Sea Research Part I: Oceanographic Research Papers* 56: 801–811.
- D'Amico, A., Gisiner, R.C., Ketten, D.R., Hammock, J.A., Johnson, C., Tyack, P.L., Mead, J. (2009). Beaked whale strandings and naval exercises. *Aquatic Mammals* 35: 452–472.
- Douve, F. (2008). The importance of marine spatial planning in advancing ecosystem-based sea use management. *Marine policy* 32(5): 762–771.
- Dunn, P.K., and Smyth, G.K. (2005). Series evaluation of Tweedie exponential dispersion model densities. *Statistics and Computing*, 15(4) 267–280.
- ESRI, 2008. ArcGIS - A Complete Integrated System Environmental Systems Research Institute, Inc., Redlands, California. <<http://esri.com/arcgis>>.
- Ferguson, M.C., Barlow, J., Reilly, S.B., Gerrodette, T. (2006). Predicting Cuvier's (*Ziphius cavirostris*) and Mesoplodon beaked whale population density from habitat characteristics in the eastern tropical Pacific Ocean. *Journal of Cetacean Research and Management* 7: 287–299.
- Fernández, A., Edwards, J.F., Rodríguez, F., Espinosa de los Monteros, A., Herráez, P., Castro, P., Jaber, J.R., Martín, V., Arbelo, M. (2005). "Gas and Fat Embolic Syndrome" Involving a Mass Stranding of Beaked Whales (Family Ziphiidae) Exposed to Anthropogenic Sonar Signals. *Veterinary Pathology* 42: 446–457.
- Foster, S.D., and Bravington, M. V. (2013). A Poisson-Gamma model for analysis of ecological non-negative continuous data. *Environmental and ecological statistics* 20: 533–552.
- Frantz, A. (1998). Does acoustic testing strand whales? *Nature* 392: 29.
- Griffin, R.B. (1999). Sperm whale distributions and community ecology associated with a warm-core ring off Georges Bank. *Marine Mammal Science* 15: 33–51.
- Gurevitch, J., Curtis, P.S., Jones, M.H. (2001). Meta-analysis in ecology. *Advances in ecological research* 32: 199–247.
- Hamazaki, T. (2002). Spatiotemporal prediction models of cetacean habitats in the mid-western north Atlantic ocean (from Cape Hatteras, North Carolina, U.S.A. to Nova Scotia, Canada). *Marine Mammal Science* 18: 920–939.

- Harris, P.T., Macmillan-Lawler, M., Rupp, J., Baker, E.K. (2014). Geomorphology of the oceans. *Marine Geology* 352: 4–24.
- Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science* 3: 297-313.
- Hedley, S.L., and Buckland, S.T. (2004). Spatial models for line transect sampling. *Journal of Agricultural, Biological, and Environmental Statistics* 9(2): 181-199.
- Higgins, J., Thompson, S.G., Spiegelhalter, D.J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 172:137–159.
- Hooker, S.B., Rees, N.W., Aiken, J. (2000). An objective methodology for identifying oceanic provinces. *Progress in Oceanography* 45: 313–338.
- Jaquet, N. (1996). How spatial and temporal scales influence understanding of sperm whale distribution: a review. *Mammal Review* 26(1): 51-65.
- Jennings, M.D. (2000). Gap analysis : concepts, methods, and recent results. *Landscape Ecology* 15: 5–20.
- King, G., and Zeng, L. (2007). When Can History be Our Guide ? The Pitfalls of Counterfactual Inference. *International Studies Quarterly* 51: 183–210.
- Kinney, S.K., and Dunson, D.B. (2008). Bayesian model uncertainty in mixed effects models. In *Random effect and latent variable model selection*. Springer New York. pp. 37-62.
- Kiszka, J., Macleod, K., Van Canneyt, O., Walker, D., Ridoux, V. (2007). Distribution, encounter rates, and habitat characteristics of toothed cetaceans in the Bay of Biscay and adjacent waters from platform-of-opportunity Data. *ICES Journal of Marine Science* 64: 1033–1043.
- Lanfredi, C., Azzellino, A., D’Amico, A., Centurioni, L., Ampolo Rella, M., Pavan, G., Podestà, M. (2016). Key Oceanographic Characteristics of Cuvier’s Beaked Whale (*Ziphius cavirostris*) Habitat in the Gulf of Genoa (Ligurian Sea, NW Mediterranean). *Journal of Oceanography and Marine Research* 4: 145.
- Laran, S., Authier, M., Blanck, A., Dorémus, G., Falchetto, H., Monestiez, P., Pettex, E., Stephan, E., Van Canneyt, O., Ridoux, V. (2017). Seasonal distribution and abundance of cetaceans within French waters- Part II: The Bay of Biscay and the English Channel. *Deep Sea Research Part II: Topical Studies in Oceanography* 141: 31–40.
- Levin, L.A., and Gooday A. (2003). *The Deep Atlantic Ocean. Ecosystems of the deep oceans.* (Tyler PA, Ed.). pp. 111-178. Amsterdam; New York: Elsevier.
- Longhurst, A. (2007). *Ecological geography of the sea.* Academic Press, Oxford.
- Macleod, K., Brereton, T., Evans, P.G., Swift, R., Vázquez, J.A. (2011). Distribution and abundance of Cuvier’s beaked whales in the canyons of southern Biscay. SC/63/SM7). In: 63st Annual Meeting of the International Whaling Commission, 1–13 June 2011, Tromsø, Norway.
- Madsen, P.T., de Soto, N.A., Tyack, P.L., Johnson, M. (2014). Beaked whales. *Current Biology* 24(16): 728-730.
- Mannocci, L., Catalogna, M., Dorémus, G., Laran, S., Lehodey, P., Massart, W., Monestiez, P., Van Canneyt, O., Watremez, P., Ridoux, V. (2014). Predicting cetacean and seabird habitats across a productivity gradient in the South Pacific gyre. *Progress in Oceanography* 120: 383–398.
- Mannocci, L., Monestiez, P., Spitz, J., Ridoux, V. (2015). Extrapolating cetacean densities beyond surveyed regions: habitat-based predictions in the circumtropical belt. *Journal of Biogeography* 42: 1267–1280.

- Mannocci, L., Roberts, J.J., Miller, D.L., Halpin, P.N. (2017). Extrapolating cetacean densities to quantitatively assess human impacts on populations in the high seas. *Conservation Biology*. 31: 601–614.
- Marques, F.F., and Buckland, S.T. (2003). Incorporating covariates into standard line transect analyses. *Biometrics* 59(4): 924-935.
- McAlpine D.F. (2009). Pygmy and dwarf sperm whales. *Encyclopedia of marine mammals 2nd Edition*. pp 936–938. Academic Press.
- McSweeney, D.J., Baird, R.W., Mahaffy, S.D. (2007). Site fidelity, associations, and movements of Cuvier's (*Ziphius cavirostris*) and Blainville's (*Mesoplodon densirostris*) beaked whales off the island of Hawai'i. *Marine Mammal Science* 23: 666–687.
- Perrin, W.F., Würsig, B., Thewissen, J.G.M. (Eds.). (2009). *Encyclopedia of marine mammals*. Academic Press.
- Pinardi, N., and Masetti, E. (2000). Variability of the large scale general circulation of the Mediterranean Sea from observations and modelling: A review. *Palaeogeography, Palaeoclimatology, Palaeoecology* 158: 153–174.
- 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.
- Plummer, M. (2016). rjags: Bayesian Graphical Models using MCMC. R package version 4-6. <https://CRAN.R-project.org/package=rjags>
- Praca, E., and Gannier, A. (2007). Ecological niche of three teuthophageous odontocetes in the northwestern Mediterranean Sea. *Ocean Science Discussions* 4: 49-59.
- R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Redfern, J.V., Moore, T.J., Fiedler, P.C., de Vos, A., Brownell, R.L., Forney, K.A., Becker, E.A., Ballance, L.T. (2017). Predicting cetacean distributions in data-poor marine ecosystems. *Diversity and Distributions* 23: 394–408.
- Roberts, J.J., Best, B.D., Dunn, D.C., Treml, E.A., Halpin, P.N. (2010). Marine Geospatial Ecology Tools: An integrated framework for ecological geoprocessing with ArcGIS, Python, R, MATLAB, and C++. *Environmental Modelling & Software* 25: 1197–1207.
- Roberts, J.J., Best, B.D., Mannocci, L., Fujioka, E., Halpin, P.N., Palka, D.L., Garrison, L.P., Mullin, K.D., Cole, T.V.N., Khan, C.B. et al. (2016). Habitat-based cetacean density models for the U.S. Atlantic and Gulf of Mexico. *Scientific Report* 6.
- Rogan, E., Cañadas, A., Macleod, K., Santos, M.B., Mikkelsen, B., Uriarte, A., Van Canneyt, O., Antonio Vázquez, J., Hammond, P.S. (2017). Distribution abundance and habitat use of deep diving cetaceans in the North-East Atlantic. *Deep Sea Research Part II: Topical Studies in Oceanography* 141: 8-19.
- Stoll, H., King, G., Zeng, L. (2014). WhatIf: Software for Evaluating Counterfactuals. R package version 1.5-6. <https://cran.r-project.org/web/packages/WhatIf/index.html>.
- Stone, C.J., and Tasker, M.L. (2006). The effects of seismic airguns on cetaceans in UK waters. *Journal of Cetacean Research and Management* 8: 255–263.
- Thomas, L., Buckland, S.T., Rexstad, E.A., Laake, J.L., Strindberg, S., Hedley, S.L., Bishop, J.R.B., Marques, T.A., Burnham, K.P. (2010). Distance software: Design and analysis of distance sampling surveys for estimating population size. *Journal of Applied Ecology* 47: 5–14.
- Tomczak, M., and Godfrey, J.S. (2003). *Regional oceanography: an introduction*. Elsevier.

- Unger, B., Rebolledo, E.L.B., Deaville, R., Gröne, A., IJsseldijk, L.L., Leopold, M.F. et al. (2016). Large amounts of marine debris found in sperm whales stranded along the North Sea coast in early 2016. *Marine pollution bulletin* 112(1): 134-141.
- Virgili, A., Racine, M., Authier, M., Monestiez, P., Ridoux, V. (2017). Comparison of habitat models for scarcely detected species. *Ecological Modelling* 346: 88–98.
- Waring, G.T., Hamazaki, T., Sheehan, D., Wood, G., Baker, S. (2001). Characterization of beaked whale (*Ziphiidae*) and sperm whale (*Physeter macrocephalus*) summer habitat in shelf-edge and deeper waters off the Northeast U.S. *Marine Mammal Science* 17: 703–717.
- Whitehead, H. (2013). Trends in cetacean abundance in the Gully submarine canyon, 1988–2011, highlight a 21% per year increase in Sowerby's beaked whales (*Mesoplodon bidens*). *Canadian Journal of Zoology* 148: 141–148.
- Wimmer, T., and Whitehead, H. (2004). Movements and distribution of northern bottlenose whales, *Hyperoodon ampullatus*, on the Scotian Slope and in adjacent waters. *Canadian Journal of Zoology* 82: 1782–1794.
- Winiarski, K.J., Burt, M.L., Rexstad, E., Miller, D.L., Trocki, C.L., Paton, P.W.C., McWilliams, S.R. (2014). Integrating aerial and ship surveys of marine birds into a combined density surface model: a case study of wintering Common Loons. *Condor* 116: 149–161.
- Wong, S.N.P., and Whitehead, H. (2014). Seasonal occurrence of sperm whales (*Physeter macrocephalus*) around Kelvin Seamount in the Sargasso Sea in relation to oceanographic processes. *Deep Sea Research Part I: Oceanographic Research Papers* 91: 10–16.
- Wood, S.N. (2006). On confidence intervals for generalized additive models based on penalized regression splines. *Australian & New Zealand Journal of Statistics* 48: 445–464.
- Wood, S. (2013). mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation. Retrieved 7 July 2014, from <http://cran.r-project.org/web/packages/mgcv/index.html>.
- Woodson, C.B., and Litvin, S.Y. (2015). Ocean fronts drive marine fishery production and biogeochemical cycling. *Proceedings of the National Academy of Sciences* 112(6): 1710-1715.

**Appendix C.1.** Details of surveys used in the analyses. Total effort represents the total length of transects of each survey (without removing the transects with a Beaufort seastate >4).

Survey name (Fig. C.1)	Organisation	Platform type	Surveyed years	Surveyed region and Sector	Total effort (km)	References
ALNILAM	ALNILAM	Ship	1998-2006	Alboran Sea – Strait of Gibraltar; MED	45,631	Cañadas and Vázquez 2014
AMBAR	AMBAR	Ship	2004-2005	South east of the Bay of Biscay; NE-ATL	5,073	Vázquez et al. 2004
ATLANCET	PELAGIS	Plane	2002	Bay of Biscay; NE-ATL	3,815	Certain et al. 2008
BMMRO	BMMRO	Ship	2000-2005	Bahamas; TROPICS	3,685	Shick et al. 2011
CODA	SMRU	Ship	2007	Northeast Atlantic; NE-ATL	9,645	Rogan et al. 2017
EcoOcean	EcoOcean institute and partners*	Ship	1998-2002, 2005-2015	Algero-Provencal basin; MED	53,294	David et al. 2011
ESAS	European Seabirds at Sea data providers	Ship	1998-2000, 2002, 2005, 2008-2010	Northeast Atlantic; NE-ATL	292,363	Reid et al. 2003
IFAW	IFAW	Ship	2003, 2004, 2007	Mediterranean Sea; MED	9,584	Lewis et al. 2007; Boisseau et al. 2010; Ryan et al. 2014; Lewis et al. 2017
INDEMARES	CEMMA	Ship	2009-2011	West of the Spanish coasts; NE-ATL	6,488	López and Martínez-Cedeira 2012
JUVENA	AZTI	Ship	2012-2015	Bay of Biscay; NE-ATL	8,862	
NEFSC	NEFSC	Plane and ship	1998-1999, 2010-2014	Continental shelf of the USA; NW-ATL	556,963	Roberts et al. 2016
PELACUS	Instituto Español de Oceanografía (IEO)	Ship	2007-2012	North and NW Spanish shelf waters; NE-ATL	9,585	Santos et al. 2013
PELGAS	PELAGIS	Ship	2007-2013	Bay of Biscay; NE-ATL	34,997	Certain et al. 2008
REMMOA	PELAGIS	Plane	2008	French West Indies and Guyana; TROPICS	15,356	Ridou et al. 2010; Mannocci et al. 2013



SAMM	PELAGIS	Plane	2011-2012	Bay of Biscay, English Channel and western Mediterranean Sea; NE-ATL and MED	98,799	Laran et al. 2017; Lambert et al. 2017
SCANS II	SMRU	Ship	2005	Northeast Atlantic; NE-ATL	19,827	Rogan et al. 2017
TETHYS - ISPRA	TETHYS - ISPRA	Plane	2009-2011, 2013-2014	Algero-Provencal basin, Tyrrhenian Sea, Ligurian Sea; MED	54,675	Panigada et al. 2011; Lauriano et al. 2014; Panigada et al. 2017
THUNNUS	CEMMA	Ship	2007-2010	Bay of Biscay; NE-ATL	11,693	Martínez-Cedeira and López 2010

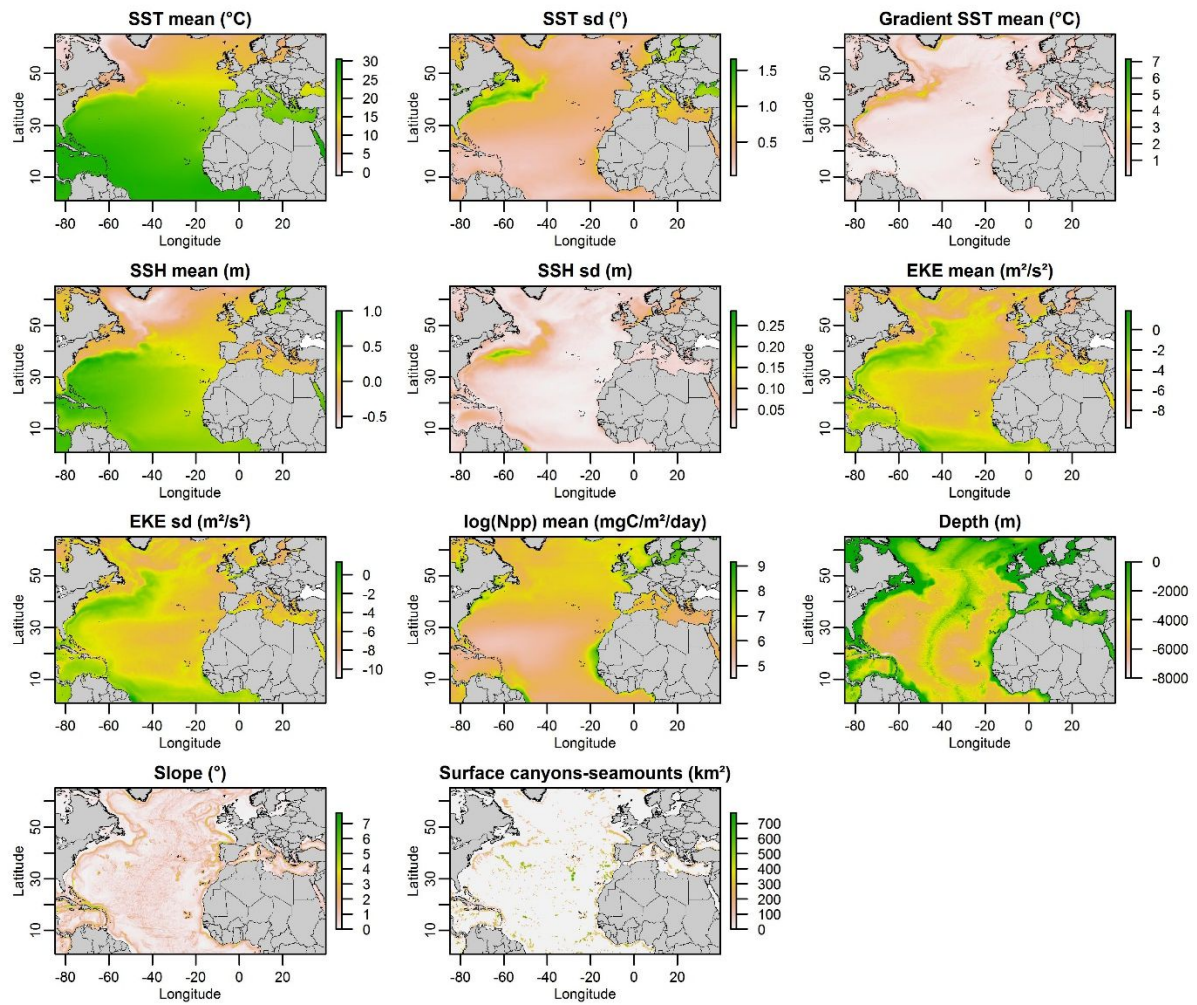
\* Partners: École Pratique des Hautes Études, WWF-France, Swiss Cetacean Society, Cybelle Planète, Participe Futur and Fondation Nicolas Hulot.

#### References of the table

- Boisseau, O., Lacey, C., Lewis, T., Moscrop, A., Danbolt, M., McLanaghan, R. (2010). Encounter rates of cetaceans in the Mediterranean Sea and contiguous Atlantic area. *Journal of the Marine Biological Association* 90: 1589-1599.
- Cañadas, A., and Vázquez, J.A. (2014). Conserving Cuvier's beaked whales in the Alboran Sea (SW Mediterranean): Identification of high density areas to be avoided by intense man-made sound. *Biol. Conserv.* 178: 155–162.
- Certain, G., Ridoux, V., Van Canneyt, O., Bretagnolle, V. (2008). Delphinid spatial distribution and abundance estimates over the shelf of the Bay of Biscay. *ICES Journal of Marine Science* 65(4): 656-666.
- David, L., Alleaume, S., Guinet, C. (2011). High risk areas of collision between fin whales and ferries in the North-western Mediterranean Sea. *Journal of Marine Animals and Their Environment* 4:17–28.
- Lambert, C., Pettex, E., Dorémus, G., Laran, S., Stephan, E., Van Canneyt, O., Ridoux, V. (2017). How does ocean seasonality drive habitat preferences of highly mobile top predators? Part II: the eastern North-Atlantic. *Deep-Sea Research II* 141: 133-154.
- Laran, S., Pettex, E., Authier, M., Blanck, A., David, L., Dorémus, G. et al. (2017). Seasonal distribution and abundance of cetaceans within French waters-Part I: The North-Western Mediterranean, including the Pelagos sanctuary. *Deep Sea Research Part II: Topical Studies in Oceanography* 141: 20-30.
- Lauriano, G., Pierantonio, N., Donovan, G., Panigada, S. (2014). Abundance and distribution of *Tursiops truncatus* in the Western Mediterranean Sea: an assessment towards the Marine Strategy Framework Directive requirements. *Marine Environmental Research* 100: 86-93.
- Lewis, T., Gillespie, D., Lacey, C., Matthews, J., Danbolt, M., Leaper, R., McLanaghan, R. Moscrop, A. (2007). Sperm whale abundance estimates from acoustic surveys of the Ionian Sea and Straits of Sicily in 2003. *Journal of the Marine Biological Association of the UK* 87: 353-357.
- Lewis, T., Boisseau, O., Danbolt, M., Gillespie, D., Lacey, C., Leaper, L., Matthews, J. N., McLanaghan, R., Moscrop, A. (2017). Abundance estimates for sperm whales in the south western and eastern Mediterranean Sea from acoustic line-transect surveys. *J. Cetacean Res. Manage.* in press.
- López, A., and Martínez-Cedeira, J. (2012). Final report of the project "LIFE 07/NAT/E/000732 INDEMARES". Unpublished technical report. CEMMA. 305 pp.

- Mannocci, L., Monestiez, P., Bolaños-Jiménez, J., Dorémus, G., Jeremie, S., Laran, S. et al. (2013). Megavertebrate communities from two contrasting ecosystems in the western tropical Atlantic. *Journal of Marine Systems* 111: 208-222.
- Martínez-Cedeira, J., and López, A. (2010). Final Report Thunnus 2007-2010 Surveys. Unpublished technical report. CEMMA. 87 pp.
- Panigada, S., Lauriano, G., Burt, L., Pierantonio, N., Donovan, G. (2011). Monitoring winter and summer abundance of cetaceans in the Pelagos Sanctuary (Northwestern Mediterranean Sea) through aerial surveys. *PLoS ONE* 6(7): e22878.
- Panigada, S., Lauriano, G., Donovan, G., Pierantonio, N., Cañadas, A., Vázquez, J. A., Burt, L. (2017). Estimating cetacean density and abundance in the Central and Western Mediterranean Sea through aerial surveys: implications for management. *Deep Sea Research Part II: Topical Studies in Oceanography*.
- Reid, J.B., Evans, P.G., Northridge, S.P. (2003). Atlas of cetacean distribution in north-west European waters. Joint Nature Conservation Committee.
- Ridoux, V., Certain, G., Doremus, G., Laran, S., van Canneyt, O., Watremez, P. (2010). Mapping diversity and relative density of cetaceans and other pelagic megafauna across the tropics: general design and progress of the REMMOA aerial surveys conducted in the French EEZ and adjacent waters (Vol. 14). SC/62.
- Roberts, J.J., Best, B.D., Dunn, D.C., Trembl, E.A., Halpin, P.N. (2010). Marine Geospatial Ecology Tools: An integrated framework for ecological geoprocessing with ArcGIS, Python, R, MATLAB, and C++. *Environ. Model. Softw.* 25: 1197–1207.
- Rogan, E., Cañadas, A., Macleod, K., Santos, M.B., Mikkelsen, B., Uriarte, A., Van Canneyt, O., Antonio Vázquez, J., Hammond, P.S., (2017). Distribution abundance and habitat use of deep diving cetaceans in the North-East Atlantic. *Deep Sea Research Part II: Topical Studies in Oceanography* 141: 8-19.
- Ryan, C., Cucknell, A. C., Romagosa, M., Boisseau, O., Moscrop, A., Frantzis, A., McLanaghan, R. (2014). A visual and acoustic survey for marine mammals in the eastern Mediterranean Sea during summer 2013. Unpublished report to the International Fund for Animal Welfare, Marine Conservation Research International, Kelvedon, UK.
- Santos, M.B., González-Quirós, R., Riveiro, I., Iglesias, M. Louzao, M., Pierce, G.J. (2013). Characterization of the pelagic fish community of the North Western and Northern Spanish shelf waters. *Journal of Fish Biology* 83(4): 716-738.
- Schick, R.S., Halpin, P.N., Read, A.J., et al. (2011). Community structure in pelagic marine mammals at large spatial scales. *Marine Ecology Progress Series* 434: 165-181.
- Vázquez, A., Ruiz, L., Maestre, Z., Ruiz-Gondra, J., Ruiz-Guijarro, J., Benedicto, L., Anza, M., Etxezarreta, A., García, O., Caballero, A., Amonarraiz, X., Goenaga I. (2003). Land-based sightings from the Basque Country coast (northeast Spain). 17th Annual Conference of the European Cetacean Society, Las Palmas de Gran Canaria, Canary Islands (Spain).
- Vázquez, J.A., Cermeño, P., Williams, A., Martin, C., Lazkano, O., Ruiz, L., Basáñez, A., Guzman, I. (2004). Identifying areas of special interest for Cuvier's beaked whale (*Ziphius cavirostris*) in the southern part of the Bay of Biscay. In Abstracts, 18th Annual Conference of the European Cetacean Society, Kolmårdon, Sweden.
- Vázquez, J.A., Guzmán, I. Lazkano, O., Olondo, M. (2005). Encounter rates of small cetaceans, pilot whales and *Ziphiidae* in coastal waters of Basque Country (Southern Bay of Biscay). 19th Annual Conference of the European Cetacean Society, La Rochelle, France.

## Appendix C.2. Monthly Environmental conditions averaged over the study period (from 1998 to 2015).



## Appendix C.3. Jags model used in the meta-analysis (based on Doyle and Dorazio's (2008) script).

```

model{
#####
### Meta analysis ###
#####

#####
# Parameters #
#####
# xia, a, tau_a: random effect for surveys
# A, L, xi, tau: random effects for height
# intercept, slope: fixed effects
# nu: shape parameters
# psi: occurrence proba

#####
# DATA #
#####
# n_obs, n_miss, n_survey, n_height
# SURVEY
# DISTANCE
# HEIGHT
# BEAUFORT
# DETECTED
# PRESENT

#####
# PRIORS #
#####

# random effect with PX-Cholesky decomposition for survey
for (j in 1:2) {
  A_a[j, j] ~ dnorm(0.0, 0.4444444)T(0.0,)
  Delta_a[j, j] <- 1/tau_a[j] ; tau_a[j] ~ dgamma(1.5, 1.5) ;
  L_a[j, j] <- 1.0;
}
L_a[1, 2] <- 0.0; A_a[1, 2] <- 0.0; Delta_a[1, 2] <- 0.0;
L_a[2, 1] ~ dnorm(0.0, 4.0); A_a[2, 1] <- 0.0; Delta_a[2, 1] <- 0.0;
# covariance matrix
Omega_a <- A_a%%L_a%%Delta_a%%t(L_a)%%A_a;
# random effects: bivariate normal
for (k in 1:n_survey) {
  alpha[k, 1] <- A_a[1, 1]*(L_a[1, 1]*xia[k, 1]);
  alpha[k, 2] <- A_a[2, 2]*(L_a[2, 1]*xia[k, 1] + L_a[2, 2]*xia[k, 2]);
  nu[k] <- exp(log_nu + alpha[k, 2]);
  for(j in 1:2){
    xia[k, j] ~ dnorm(0.0, tau_a[j]);
  }
}
sigma_alpha[1] <- sqrt(Omega_a[1, 1]); sigma_alpha[2] <- sqrt(Omega_a[2,
2]);
rho_alpha[1] <- Omega_a[1, 2]/sqrt(Omega_a[1, 1]*Omega_a[2, 2]);

# random effects with PX-Cholesky decomposition for height
# covariance matrix
Omega_b <- A_b%%L_b%%Delta_b%%t(L_b)%%A_b;
sigma_beta[1] <- sqrt(Omega_b[1, 1]); sigma_beta[2] <- sqrt(Omega_b[2,
2]);
rho_beta <- Omega_b[1, 2]/sqrt(Omega_b[1, 1]*Omega_b[2, 2]);

```

```

for (l in 1:2) {
  A_b[l, 1] ~ dnorm(0.0, 0.4444444)T(0.0,)
  Delta_b[l, 1] <- 1/tau_b[l]; tau_b[l] ~ dgamma(1.5, 1.5);
  L_b[l, 1] <- 1.0;
}
L_b[1, 2] <- 0.0; A_b[1, 2] <- 0.0; Delta_b[1, 2] <- 0.0;
L_b[2, 1] ~ dnorm(0.0, 4.0); A_b[2, 1] <- 0.0; Delta_b[2, 1] <- 0.0;

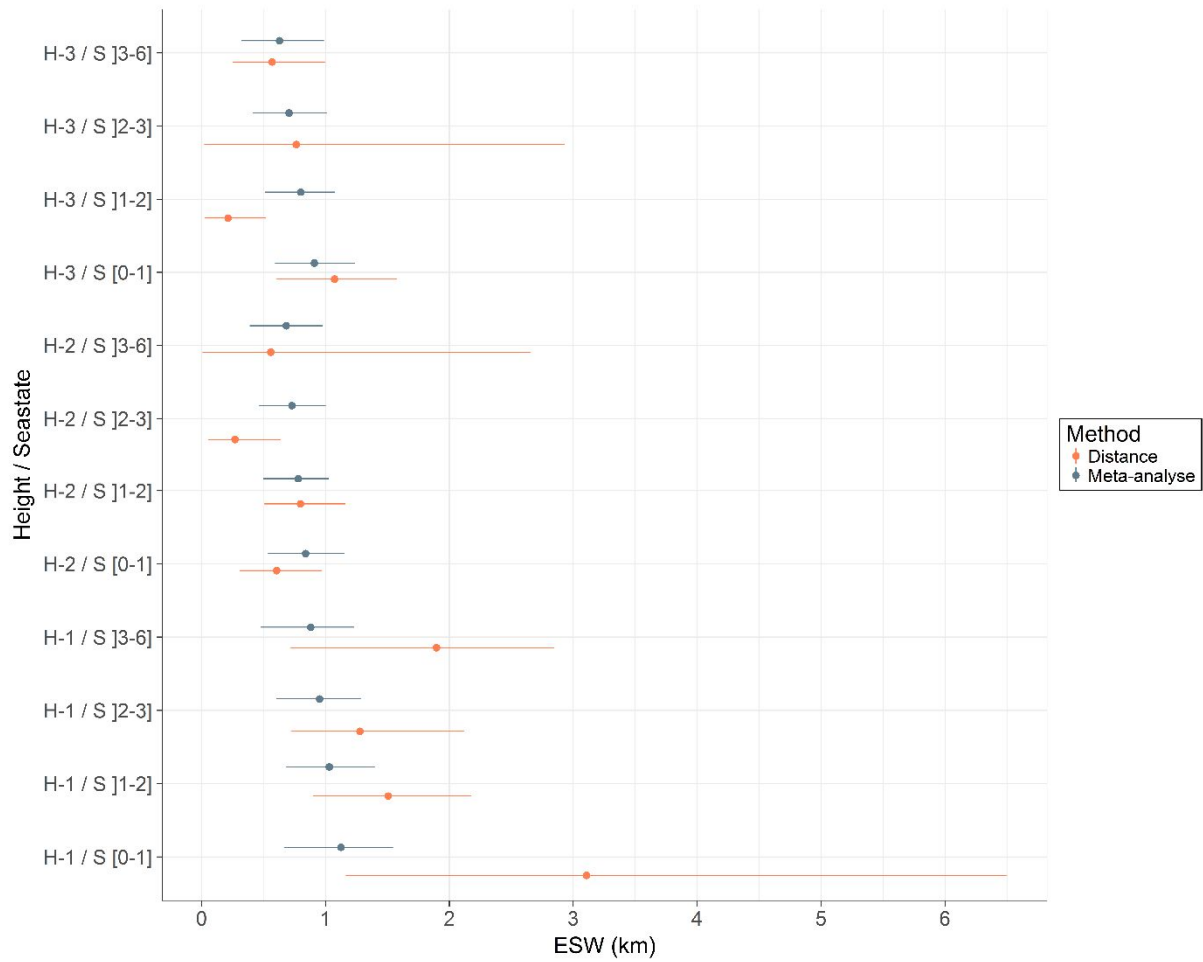
for(i in 1:n_height){
  beta[i, 1] <- intercept[1] + slope[1]*(i-2) + A_b[1, 1]*(L_b[1,
1]*xib[i, 1]);
  beta[i, 2] <- intercept[2] + slope[2]*(i-2) + A_b[2, 2]*(L_b[2,
1]*xib[i, 1] + L_b[2, 2]*xib[i, 2]);
  for(j in 1:2) {
    xib[i, j] ~ dnorm(0.0, tau_b[j]);
  }
}

# fixed effects
for (l in 1:2) {
  intercept[l] ~ dnorm(0.0, 1.0);
  slope[l] ~ dnorm(0.0, 1.0);
}
log_nu ~ dnorm(0.0, 0.64);
psi ~ dunif(0.0, 1.0);

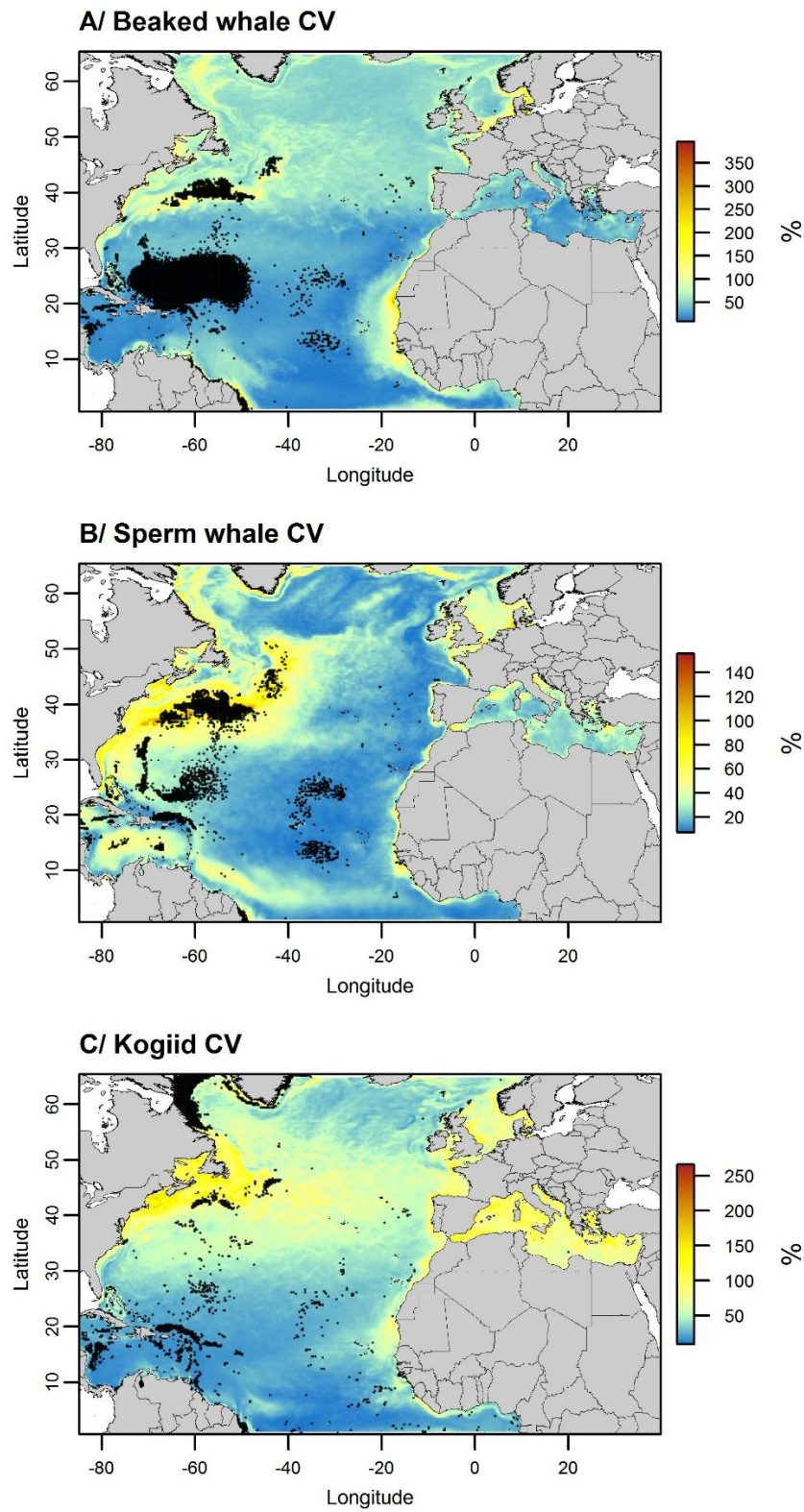
#####
# likelihood #
#####
for (j in 1:(n_obs+n_miss)){
  DISTANCE[j] ~ dunif(0.0, TRUNC);
  PRESENT[j] ~ dbern(psi);
  sigma[j] <- exp(beta[HEIGHT[j], 1] + beta[HEIGHT[j], 2]*BEAUFORT[j] +
alpha[SURVEY[j], 1]);
  z[j] <- max(DISTANCE[j]/sigma[j], 0.001);
  x[j] <- (1 - (1 - equals(DISTANCE[j], 0))*exp(-pow(z[j], -
nu[SURVEY[j]])))*PRESENT[j];
  prob[j] <- max(0.0001, min(x[j], 0.9999));
  DETECTED[j] ~ dbern(prob[j]);
}
}

```

**Appendix C.4.** Comparison of beaked whale ESWs estimated with the meta-analysis in blue and Distance software in orange for shipboard surveys. On the y-axis, all combinations of observation height (H) and Beaufort seastate (S) are represented. Class H-1 corresponds to observation heights between 0-5 m, H-2 between 5-10 m and H-3 between 10-40 m. Class S[0-1] corresponds to Beaufort seastate between 0-1, S[1-2] corresponds to Beaufort seastate between 1-2, S[2-3] corresponds to Beaufort seastate between 2-3 and S[3-6] corresponds to Beaufort seastate between 3-6.



**Appendix C.5.** Uncertainty maps representing the coefficient of variation (CV) in % associated with the predicted relative density of beaked whale, sperm whale and kogiid groups. Black areas represent extrapolation where we did not extrapolate the predictions.



# Annex D

---

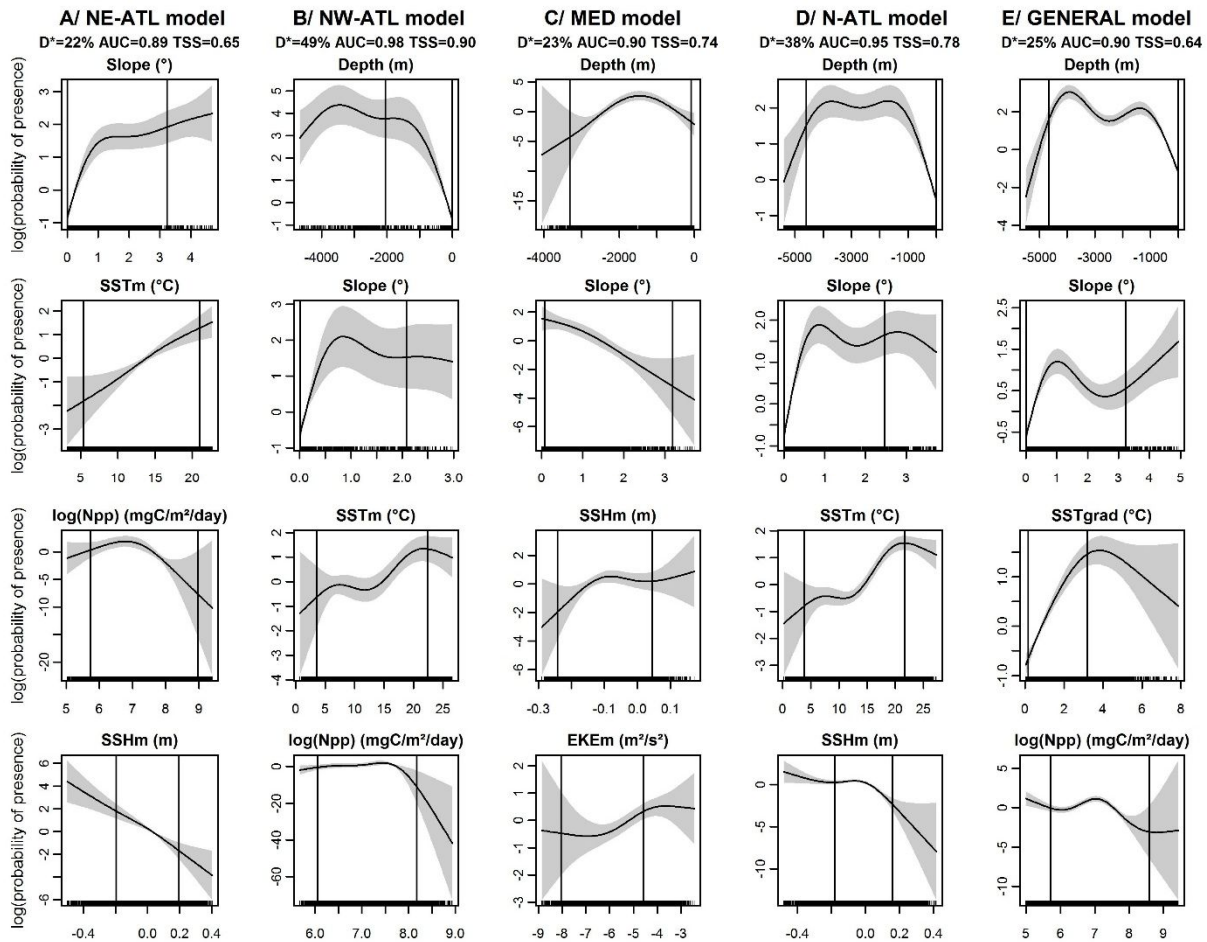
## DATA-ASSEMBLING: A MATTER OF ECOSYSTEMS SIMILARITY

---

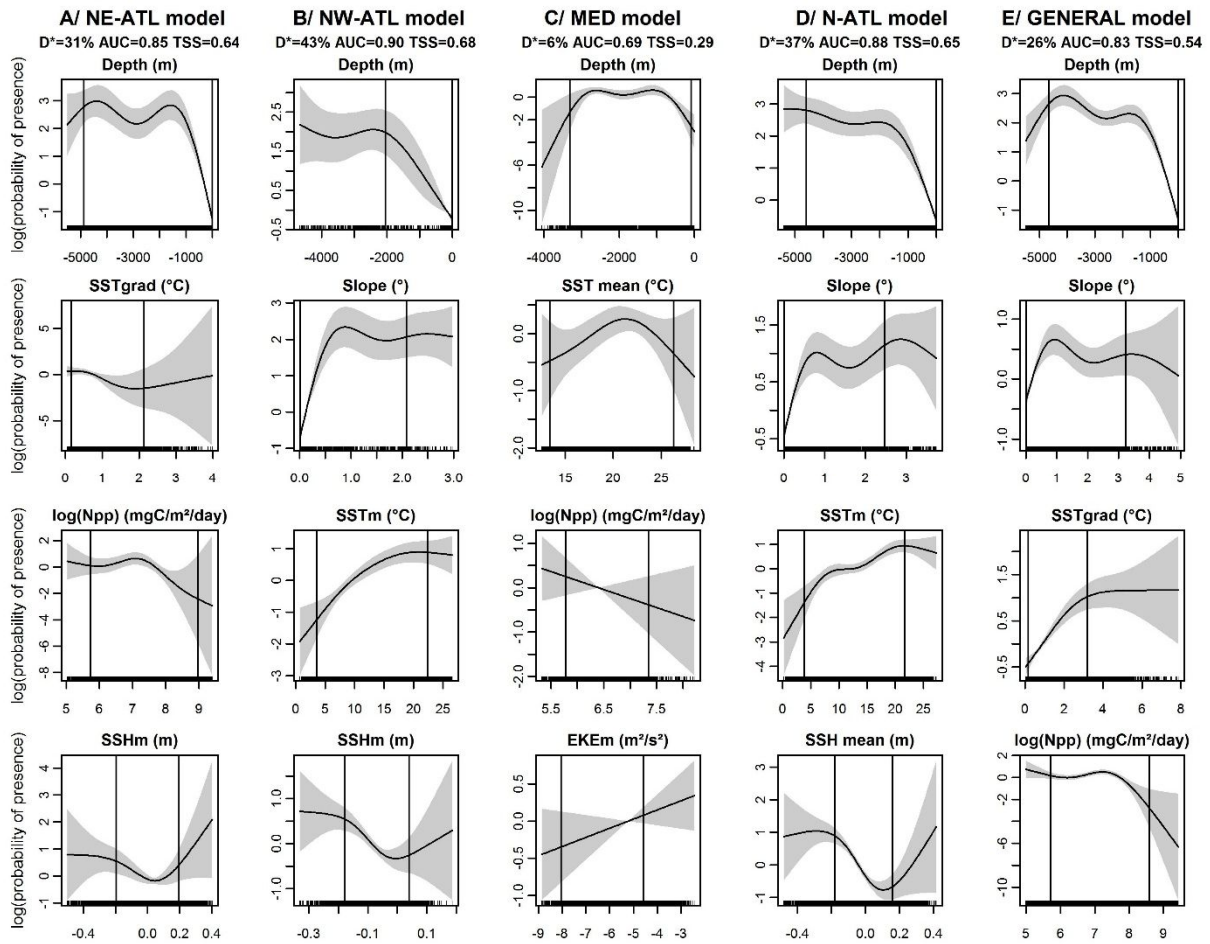
### Chapter 5 – Supporting information



**Appendix D.1. Functional relationships for the selected covariates of each beaked whale model fitted to the data of the corresponding region.** For example, 'NE-ATL model' refers to the model fitted to the data of the NE-ATL region. NE-ATL: north-east Atlantic region; NW-ATL: north-west Atlantic region; MED: Mediterranean region; N-ATL: north Atlantic region and pooled north-west and north-east Atlantic regions; GENERAL: study area and pooled the three regions; D\*: explained deviance; AUC: Area Under the receiving Curve; TSS: True Skill Statistics. The solid line in each plot is the smooth function estimate and shaded regions represent approximate 95% confidence intervals. The y-axis indicates the logarithm of the probability of presence. The x-axis indicates the values of the covariates and zero on the x-axis indicates no effect of the covariate. Best model fits are between the vertical lines indicating the 2.5<sup>th</sup> and 97.5<sup>th</sup> quantiles of the data.



**Appendix D.2. Functional relationships for the selected covariates of each beaked whale model fitted to the data of the corresponding region.** For example, ‘NE-ATL model’ refers to the model fitted to the data of the NE-ATL region. NE-ATL: north-east Atlantic region; NW-ATL: north-west Atlantic region; MED: Mediterranean region; N-ATL: north Atlantic region and pooled north-west and north-east Atlantic regions; GENERAL: study area and pooled the three regions; D\*: explained deviance; AUC: Area Under the receiving Curve; TSS: True Skill Statistics. The solid line in each plot is the smooth function estimate and shaded regions represent approximate 95% confidence intervals. The y-axis indicates the logarithm of the probability of presence. The x-axis indicates the values of the covariates and zero on the x-axis indicates no effect of the covariate. Best model fits are between the vertical lines indicating the 2.5<sup>th</sup> and 97.5<sup>th</sup> quantiles of the data.





# Annex E

---

## WOULD MODELS BE IMPROVED IF PREY DISTRIBUTIONS WERE INCLUDED?

---

### CONTENTS

---

E.1 CONTEXT AND OBJECTIVES .....	222
E.2 METHODOLOGY .....	222
E.2.1 Data origin.....	222
E.2.2 Model fitting, predictions and assessment.....	226
E.3 PRELIMINARY RESULTS .....	227
E.3.1 Beaked whales.....	227
E.3.2 Sperm whales.....	228

**T**HIS appendix describes a work in progress which attempts to explore how models would be improved if the distributions of deep-diver preys were included as explanatory variables. To do that, models with environmental variables, with variables that describe the prey distribution and with a combination of these two types of variables were compared. This annex is planned to be published as a separate stand-alone paper.

## E.1 CONTEXT AND OBJECTIVES

Environmental variables, such as depth, slope or sea surface temperature are supposed to be good indicators of the distribution of lower trophic levels and thus good proxies of the distribution of top predators (Ferguson et al. 2006; Redfern et al. 2006; Mannocei et al. 2014a). However, there is a time-lag between a change in an environmental condition and its effects on upper trophic levels (Jaquet 1996; Austin et al. 2006; Redfern et al. 2006; Cotté et al. 2009). Also, the relationships between these distal predictors and the actual quality of the habitat for a predator can vary with the underlying ecological processes (see above). The use of more proximal variables, such as prey distribution, could reduce these lags because marine top predators are supposed to be mostly sensitive to prey abundance (Österblom et al. 2008). Nevertheless, field data on prey distributions are not available at the scale of the Atlantic and Mediterranean basins and will not be so in a foreseeable future. To cope with this gap, a numerical model, the Spatial Ecosystem And POPulation DYNamics Model (SEAPODYM), provides simulation of distributions of zooplankton and six functional groups of the micronekton at the global scale. It has been initially used to model tuna populations (Lehodey et al. 2008) but its usage was recently extended to predict turtle and cetacean habitats (Abecassis et al. 2013; Lambert et al. 2014). Consequently, in a work in progress (Annex E), I aimed to explore if models fitted by using data of prey distributions predicted better the deep-diver distribution than models fitted by using more conventional environmental data. I also aimed to explore if the combination of environmental and prey distribution data would further improve model results. To do that, for each species group of deep-diving cetaceans (beaked whales, sperm whales and kogiids), I compared the performance of three models that used environmental variables, SEAPODYM variables and a combination of environmental and SEAPODYM variables.

## E.2 METHODOLOGY

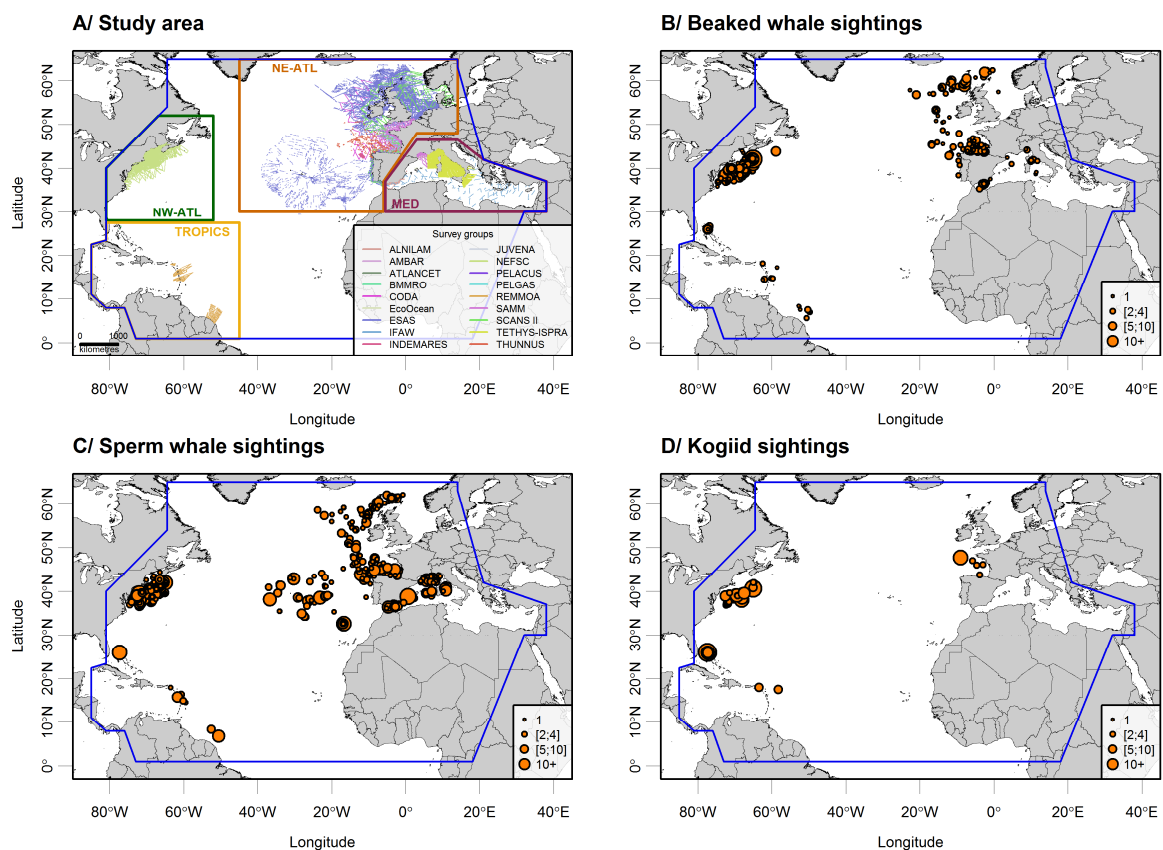
### E.2.1 Data origin

This study was based on the same effort and sighting data as in Chapter 4 (as a reminder, Fig. E.1), *i.e.* data assembled from surveys carried out in the Mediterranean Sea and the Atlantic Ocean. I used the three datasets of beaked whales, sperm whales and kogiids to perform the analyses. I also used the same environmental variables extracted for the entire time period (1998-2015) and resampled at a 0.25° resolution (Table E.1).

The SEAPODYM model is driven by water temperature, currents, net primary production and chlorophyll-a concentrations. It simulates the development in time and space of micronekton and, more recently, zooplankton (Lehodey et al. 1998; 2008; 2010; Conchon 2016). In the SEAPODYM model, the zooplankton is defined as all non-migratory phytoplanktonous organisms with a size between 200 µm and 2 mm that live in the epipelagic layer (Conchon 2016). The micronekton encompasses active swimming organisms in the range of 1-20 g and 2-20 cm and includes fishes, crustaceans and cephalopods (Lehodey et al. 2014; Conchon 2016). In the SEAPODYM model, depending on the vertical distribution of the organism biomass, three layers are defined according to the euphotic depth (*i.e.* the layer of sea water that receives enough sunlight for photosynthesis to occur; Fig. E.2; Lehodey et al. 2014). The epipelagic layer which extends from the surface to 1.5\* the euphotic depth, the mesopelagic layer which extends between 1.5 and 4.5\*euphotic depth and the bathypelagic layer which extends from 4.5 to 10.5\*euphotic depth with a maximum set at 1000 m (Lehodey et al. 2014). Organisms of

the micronekton can undertake nycthemeral vertical migrations between these three layers. According to their migration patterns, they are divided into six functional groups: epipelagic, non-migrant mesopelagic, migrant mesopelagic, non-migrant bathypelagic, migrant bathypelagic and highly migrant bathypelagic organisms (Fig. E.2; Lehodey et al. 2008; 2010; 2014). These migrations are induced by daylight variations and may be due to a strategy of predator avoidance; during daytime, they dive in deeper waters to avoid the predators (Hays 2003). Migrant mesopelagic and highly migrant bathypelagic organisms migrate between the epipelagic layer at night and respectively the mesopelagic and bathypelagic layers during daytime while the migrant bathypelagic organisms migrate between the mesopelagic and the bathypelagic layers. For each functional group (zooplankton and micronekton) biomass and production are simulated at a 0.25° resolution. Production is defined as the recruitment of a new cohort of organisms into a micronekton functional group when they reach 1 g body mass (fixed value). This variable would indicate a preference for smaller and more abundant preys of the functional group than for larger preys, better described by the biomass variable.

Due to the available parameterisation of the SEAPODYM model and the absence of the bathypelagic layer on the continental shelf, each time a combination of variables that contained a variable of the bathypelagic layer was tested in the model selection procedure, all effort data associated with the continental shelf were removed because no GAM can be fitted where the variable has no value. Therefore, to ensure that all tested models included the same number of effort data, and were thus comparable, all segments recorded on the continental shelf were removed for the three species groups.



**Fig. E.3.** The study area (A), and the beaked whale (B), sperm whale (C) and kogiid (D) sightings recorded during the surveys. The blue polygon delineates the study area. Surveys were carried out along transects (lines) following a line-transect methodology. Sightings were classified by group sizes with each point representing one group of individuals and point size relating to the number of animals in a group.

**Table E.1. Candidate environmental predictors used for the habitat modelling.** All variables were resampled at a 0.25° resolution. A: Depth and slope were derived from GEBCO-08 30 arc-second database (GEBCO; 30 arc-second is approximately equal to 0.008°. B: Surface per cell was calculated in ArcGIS 10.3 from the shapefile of canyons and seamounts provided by Harris et al. (2014). C: The mean, standard error and gradient of Sea Surface Temperature (SST) were calculated from the GHRSSST Level 4 CMC SST v.2.0 (Canada Meteorological Centre). D: The Aviso ¼° DT-MADT geostrophic currents dataset was used to compute mean and standard deviation of Sea Surface Height (SSH) and Eddy Kinetic Energy (EKE; AVISO). E: Net primary production (NPP) was derived from SeaWiFS and Aqua using the Vertically Generalised Production Model (OREGONSTATE). F: Euphotic depth (ZEU) and variables of prey distribution were obtained with the SEAPODYM model (Lehodey et al. 2010).

Variables used in the study with abbreviations and units	Original Resolution	Sources	Effects on pelagic ecosystems of potential interest to deep-divers
<b>Physiographic</b>			
Depth (m)	30 arc sec	A	Deep-divers feed on squids and fish in the deep water column
Slope (°)	30 sec arc	A	Associated with currents, high slope induce prey aggregation or enhanced primary production
Surface of canyons and seamounts in a 0.25° cell – Surface.can.seam (km <sup>2</sup> )	30 sec arc	B	Deep-divers are often associated with canyons and seamounts structures; the variable indicates the proportion of this habitat in each cell
<b>Oceanographic</b>			
Mean of SST – SSTm (°C)	0.2°, daily	C	Variability over time and horizontal gradients of SST reveal front locations, potentially associated with prey aggregations or enhanced primary production
Standard error of SST – SSTsd (°C)	0.2°, daily	C	
Mean gradient of SST – SSTgrad (°C)	0.2°, daily	C	
Mean of SSH – SSHm (m)	0.25°, daily	D	High SSH is associated with high mesoscale activity and enhanced prey aggregation or primary production
Standard deviation of SSH – SSHsd (m)	0.25°, daily	D	
Mean of EKE – EKEm (m <sup>2</sup> .s <sup>-2</sup> )	0.25°, daily	D	High EKE relates to the development of eddies and sediment resuspension induce prey aggregation
Standard error of EKE – EKESd (m <sup>2</sup> .s <sup>-2</sup> )	0.25°, daily	D	
Mean of NPP – Npp (mgC.m <sup>-2</sup> .day <sup>-1</sup> )	9 km, 8 days	E	Net primary production as a proxy of prey availability
Mean of ZEU – ZEUm (m)	0.25°, weekly	F	Depth of the euphotic zone as proxy of prey availability
Srandard error of ZEU – ZEUsd (m)	0.25°, weekly	F	

Table E.1. (Continued)

Variables used in the study with abbreviations and units	Original Resolution	Sources	Effects on pelagic ecosystems of potential interest to deep-divers
<b>Prey distribution</b>			
Epipelagic biomass and production – Epi_b ( $\text{g.m}^{-2}$ ) and Epi_p ( $\text{g.m}^{-2}.\text{day}^{-1}$ )	0.25°, weekly	F	All these variables relate to the distribution of potential direct or indirect prey of deep-divers.
Mesopelagic biomass and production – Meso_b ( $\text{g.m}^{-2}$ ) and Meso_p ( $\text{g.m}^{-2}.\text{day}^{-1}$ )	0.25°, weekly	F	
Migrant mesopelagic biomass and production – MMeso_b ( $\text{g.m}^{-2}$ ) and MMeso_p ( $\text{g.m}^{-2}.\text{day}^{-1}$ )	0.25°, weekly	F	
Bathypelagic biomass and production – Bathy_b ( $\text{g.m}^{-2}$ ) and Bathy_p ( $\text{g.m}^{-2}.\text{day}^{-1}$ )	0.25°, weekly	F	
Migrant Bathypelagic biomass and production – MBathy_b ( $\text{g.m}^{-2}$ ) and MBathy_p ( $\text{g.m}^{-2}.\text{day}^{-1}$ )	0.25°, weekly	F	
Highly migrant bathypelagic biomass and production – HMBathy_b ( $\text{g.m}^{-2}$ ) and HMBathy_p ( $\text{g.m}^{-2}.\text{day}^{-1}$ )	0.25°, weekly	F	
Zooplankton biomass and production – PK_b ( $\text{g.m}^{-2}$ ) and PK_p ( $\text{g.m}^{-2}.\text{day}^{-1}$ )	0.25°, weekly	F	

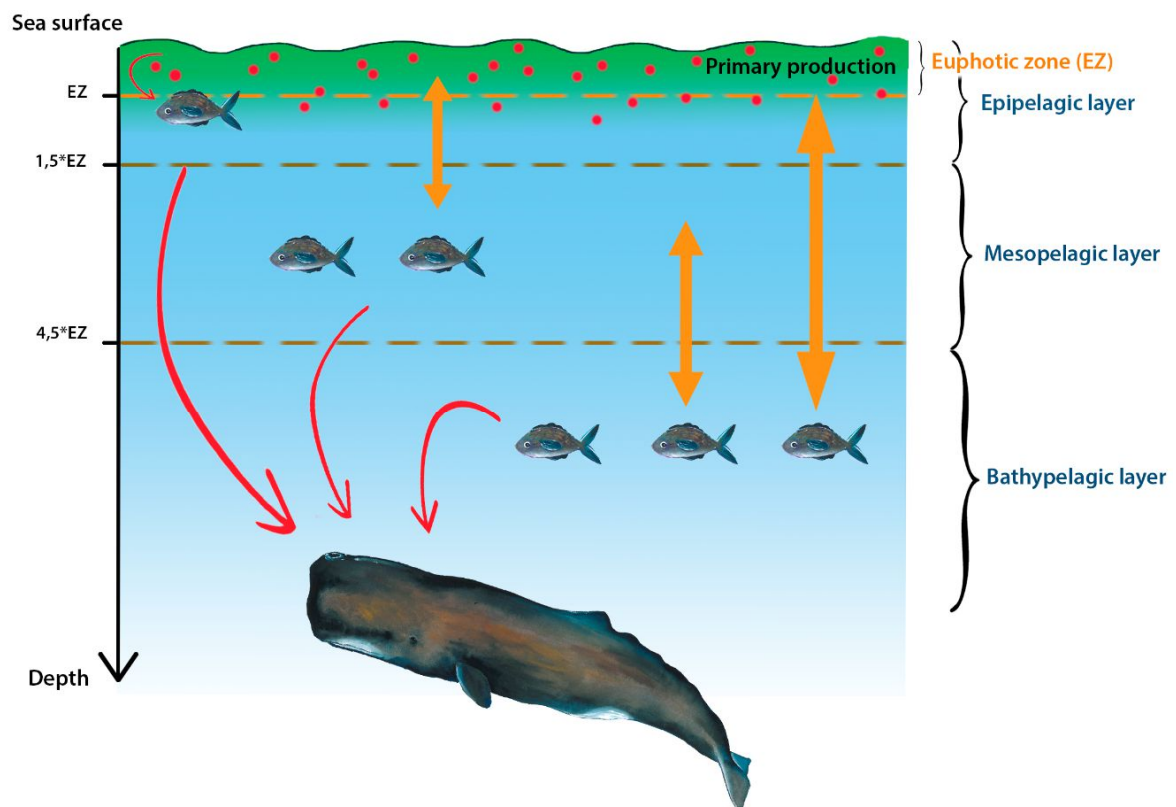


Fig. E.2. Vertical repartition of the functional groups of micronekton of the SEAPODYM model in the water column. The three layers are the layers defined in SEAPODYM, fishes represent the micronekton functional groups, the sperm whale represents the predators, the red dots represent the zooplankton, the red arrows represent the trophic relationships and the yellow arrows represent the daily vertical migrations of the functional groups between layers. Illustration from Laura Hedon.



### E.2.2 Model fitting, predictions and assessment

In a first step, I built, for each species group, two habitat models, one with environmental variables (hereafter 'ENVIRONMENTAL model') and one with SEAPODYM variables (hereafter 'SEAPODYM model'). I fitted Generalised Additive Models (GAMs) with a Tweedie distribution (Hastie and Tibshirani 1986; Wood 2006) by using the 'mgcv' R-package (R-3.3.1 version; Wood 2013). The parameter of the Tweedie distribution was directly estimated by the 'mgcv' function. I used the same variable selection procedure as in the previous chapters by removing combinations of variables with Spearman partial correlation coefficient higher than  $|0.7|$ , by testing all models with combinations of four variables (Mannocci et al. 2014; Virgili et al. 2017) and selecting the best model with the lowest generalised cross validation score (Wood 2006; Clark 2013). A maximum of four covariates per model was used to avoid excessive complexity of models and difficulty in their interpretation (Mannocci et al. 2014; Virgili et al. 2017). As in Chapter 5, to facilitate comparisons of models and predictions, count data were transformed into presence-absence data, *i.e.* any sighting, regardless of group size, was considered as a single observation ('1') and the '0' remain absences.

In a second step, I built, for each species group, a model which combined environmental and SEAPODYM variables (hereafter 'MIXED model'). As variables were at the same temporal and spatial resolutions, combination was possible. To limit computation times, I reduced the number of variables implemented in the variable selection procedure of the MIXED model by selecting only the variables that were selected in the five best ENVIRONMENTAL and SEAPODYM models, which represented six variables per model (combinations of variables selected by the best models were generally similar). All models with combinations of four variables (among the twelve available variables) were tested and the best MIXED model with the lowest GCV was selected.

Following the methodology of the previous chapters, monthly predictions at  $0.25^\circ$  resolution were averaged over the entire time period (1998-2015) to produce maps of mean predicted probabilities of presence and represent expected long-term patterns of the beaked whales, sperm whales and kogiids in the study area.

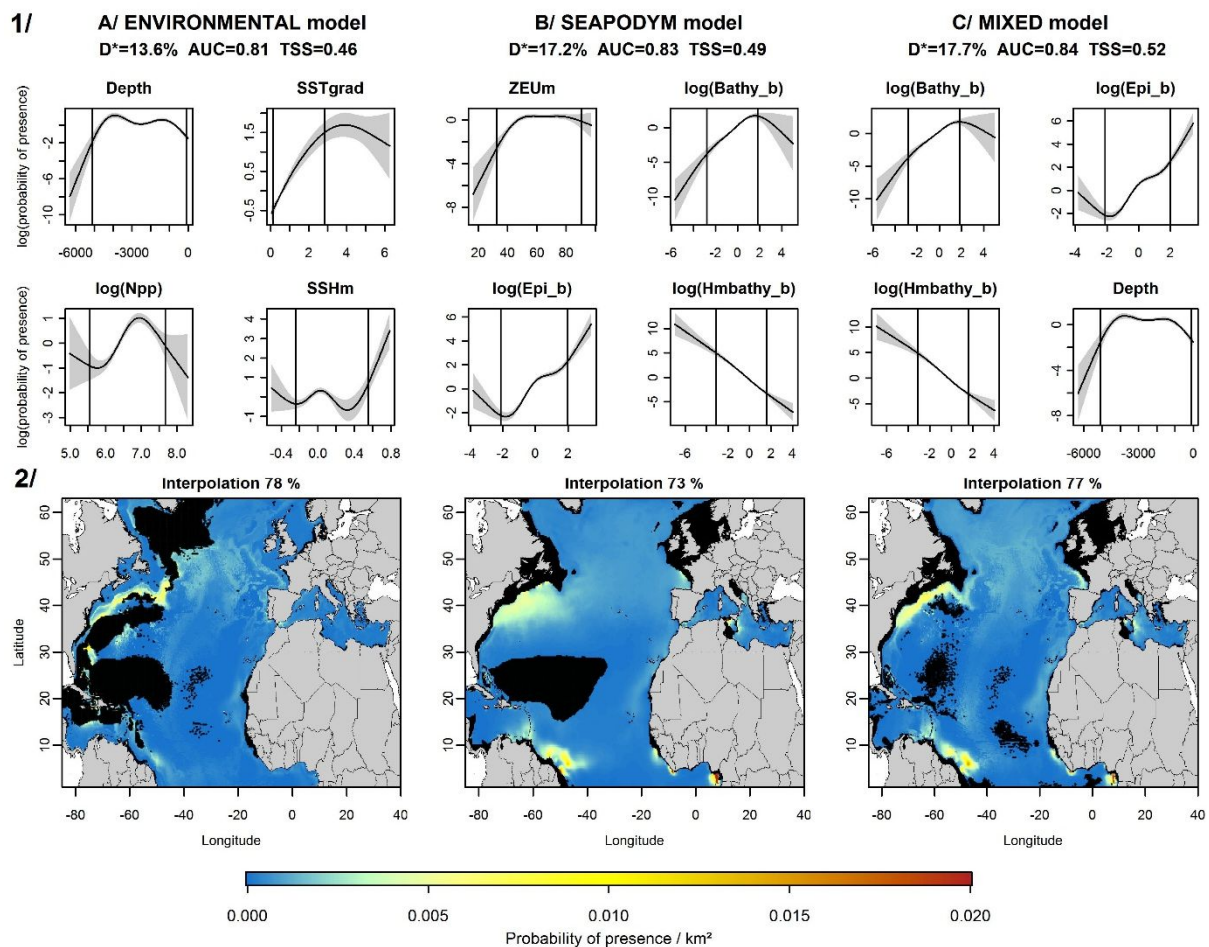
As in chapter 4, I also performed an environmental space coverage gap analysis to determine if the use of prey distribution variables could increase the extent of the interpolation areas, where predictions are considered as reliable (Jennings 2000). Here, I only delineated 'simple interpolation' areas (*i.e.* I considered the full range of sampled variables to identify all points of the whole study area where the actual combinations of environmental variables had been sampled in survey blocks) because it was sufficient to assess the effect of SEAPODYM variables on species distribution predictions.

The predicting performance of each model was assessed thanks to the explained deviances (Wood 2006), the AUC (Elith et al. 2011) and the TSS (Allouche et al. 2006), as in Chapter 5. This allowed to compare the three models in order to assess which variables better explained the distribution of deep-diving cetaceans.

## E.3 PRELIMINARY RESULTS

### E.3.1 Beaked whales

With the ENVIRONMENTAL model, highest beaked whale probabilities of presence were predicted at depths *ca.* 1,500 m and 4,000 m and high spatial gradients of SST, net primary production and SSH (Fig. E.3A). This indicated a preference for habitats associated with high depths and thermal fronts. Highest probabilities of presence were predicted on the western side of the Atlantic Ocean near the Gulf Stream and along continental slopes and the Mid-Atlantic Ridge. In the Mediterranean Sea, predicted probabilities of presence were lower than in the Atlantic Ocean with highest probabilities of presence predicted in the Alboran Sea.



**Fig. E.3. Functional relationships for the selected variables (1/) and the predicted probabilities of presence of beaked whales (2/) for the ENVIRONMENTAL model (A), the SEAPODYM model (B) and the MIXED model (C). 1/:** Solid lines are the estimated smooth functions, and the shaded regions represent the approximate 95% confidence intervals. The y-axes indicate the probabilities of presence on a log scale, where zero indicates no effect of the covariate. The vertical lines indicate the 2.5<sup>th</sup> and 97.5<sup>th</sup> quantiles of the data. D\*: explained deviance; AUC: Area Under the receiving Curve; TSS: True Skill Statistics. 2/:

Black areas on prediction maps represent zones where we did not extrapolate the predictions. Percentages represent the proportion of the study area defined as interpolation with the gap analysis.

With the SEAPODYM model, highest beaked whale probabilities of presence were predicted at euphotic depth *ca.* 80 m, high biomass of bathypelagic and epipelagic organisms and low biomass of highly migrant bathypelagic organisms (Fig. E.3B). This indicated a preference for high productive habitats and resulted in a prediction of highest probabilities of presence along the Gulf Stream and near the Ecuador in the Atlantic Ocean and in the western and the centre of the Mediterranean Sea.

The MIXED model combined variables of the ENVIRONMENTAL and the SEAPODYM models and predicted highest probabilities of presence predicted for a high biomass of bathypelagic and epipelagic organisms, a low biomass of highly migrant bathypelagic organisms and at depths *ca.* 1,500 m and 4,000 m (Fig. E.3C). This indicated a preference for high productive habitats associated with high depths. This resulted in similar patterns as for the ENVIRONMENTAL model but with highest probabilities of presence near the Ecuador.

Interpolation areas varied between the three models with the highest percentage of interpolation observed for the ENVIRONMENTAL model (78%) and the lowest for the SEAPODYM model (73%; Fig. E.3.2). In SEAPODYM and MIXED models, the continental shelf was an extrapolation because there were no values of biomass or production of bathypelagic organisms.

The model performance assessment criteria varied between the three models (Fig. E.3.1). They were slightly better for the MIXED model (highest explained deviance, AUC and TSS) suggesting a better performance of models that combined biotic and abiotic variables.

### E.3.2 Sperm whales

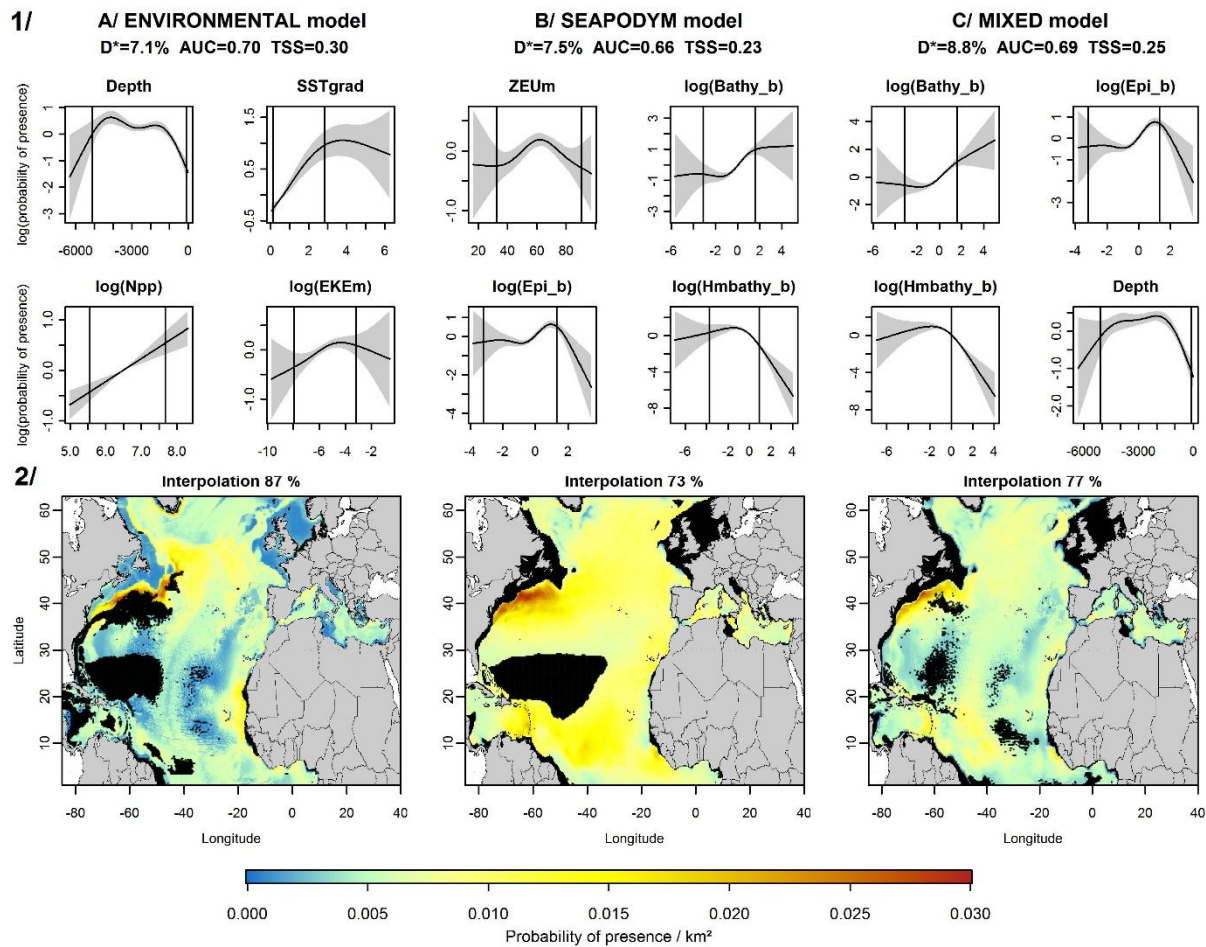
With the ENVIRONMENTAL model, highest sperm whale probabilities of presence were predicted at depths *ca.* 1,500 m and 4,000 m, high spatial gradients of SST, high net primary production and medium eddy kinetic energy (Fig. E.4A). This indicated a preference for habitat associated with high depths, thermal fronts, a high productivity and strong eddies. Highest probabilities of presence were predicted on the western side of the Atlantic Ocean near the Gulf Stream and fairly homogeneously from continental slopes to oceanic waters. In the Mediterranean Sea, predicted probabilities of presence were fairly homogeneous but lower than in the Atlantic Ocean with highest probabilities of presence predicted in the Alboran Sea and on the continental slopes.

With the SEAPODYM model, highest sperm whale probabilities of presence were predicted at euphotic depth *ca.* 60 m, fairly high biomass of bathypelagic and epipelagic organisms and fairly low biomass of highly migrant bathypelagic organisms (FIG. E.4B). This indicated a preference for fairly high productive habitats and resulted in a prediction of highest probabilities of presence near the Gulf Stream in the Atlantic Ocean and in the western Mediterranean Sea.

Highest probabilities of presence were predicted for a high biomass of bathypelagic and epipelagic organisms, at depths *ca.* 1,500 m and 4,000 m and a fairly low biomass of highly migrant bathypelagic organisms with the MIXED model (Fig. E.4C). This indicated a preference highly productive habitats associated with deep depths. This resulted in similar patterns as for the ENVIRONMENTAL model but more homogeneous and slightly lower probabilities of presence.

Interpolation areas varied between the three models and were less extended for the SEAPODYM (73%) and MIXED models (77%) compared to the ENVIRONMENTAL model (87%; Fig. E.4.2).

The model performance assessment criteria little varied between the three models (Fig. E.4.1). They were slightly better for the ENVIRONMENTAL and MIXED models (highest explained deviance and AUC than the SEAPODYM model) but the difference was not really significant.



**Fig. E.4. Functional relationships for the selected variables (1/) and the predicted probabilities of presence of sperm whales (2/) for the ENVIRONMENTAL model (A), the SEAPODYM model (B) and the MIXED model (C).** 1/: Solid lines are the estimated smooth functions, and the shaded regions represent the approximate 95% confidence intervals. The y-axes indicate the probabilities of presence on a log scale, where zero indicates no effect of the covariate. The vertical lines indicate the 2.5<sup>th</sup> and 97.5<sup>th</sup> quantiles of the data. D\*: explained deviance; AUC: Area Under the receiving Curve; TSS: True Skill Statistics. 2/: Black areas on prediction maps represent zones where we did not extrapolate the predictions. Percentages represent the proportion of the study area defined as interpolation with the gap analysis.

### E.1.1 Kogiids

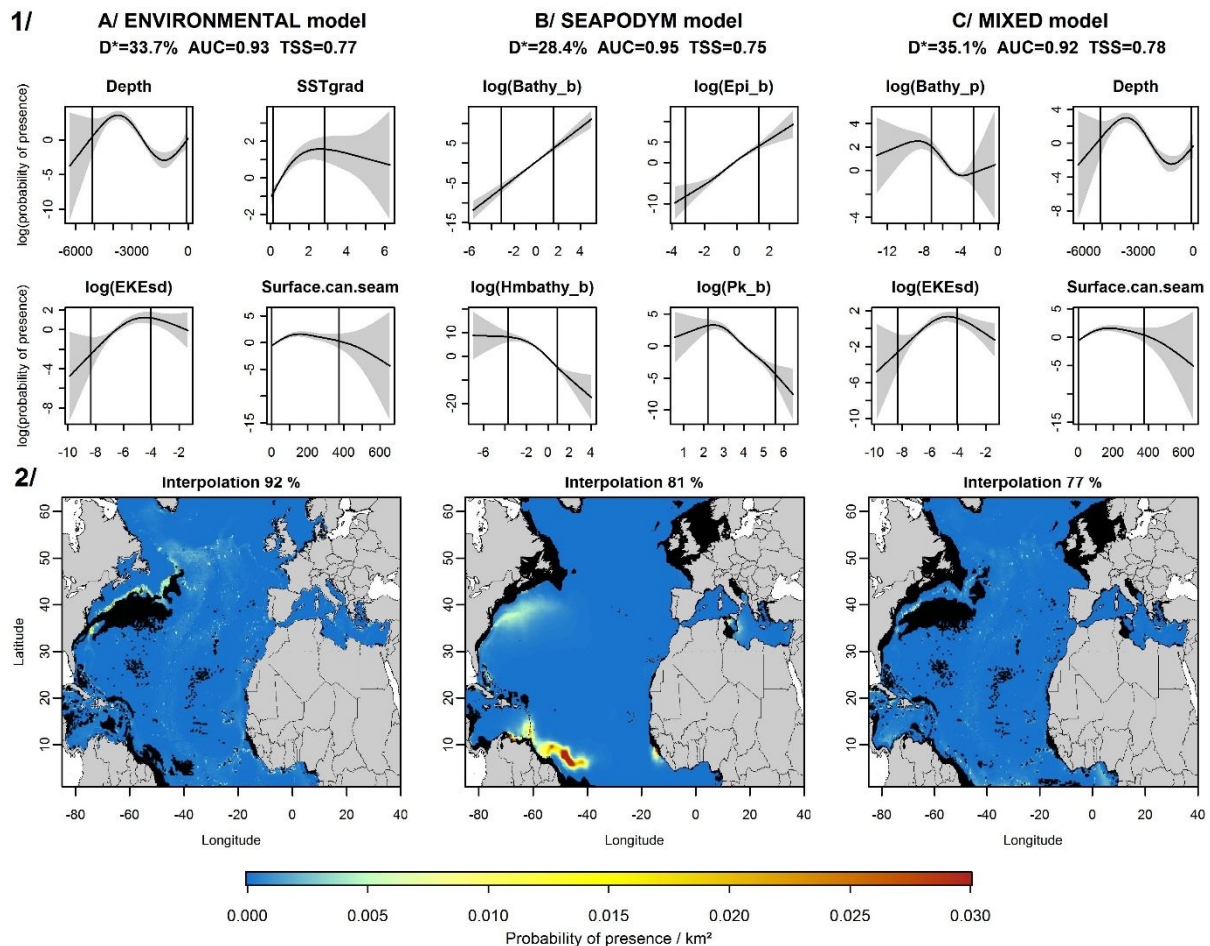
With the ENVIRONMENTAL model, highest kogiid probabilities of presence were predicted at depths *ca.* 4,000 m, high spatial gradients of SST and eddy kinetic energy and medium surfaces of canyons and seamounts habitats (*ca.* 200 km<sup>2</sup>; Fig. E.5A). This indicated a preference for habitat associated with high depths, high thermal fronts, dynamic waters and canyons and seamounts habitats. Highest probabilities of presence were predicted on the western side of the Atlantic Ocean near the Gulf Stream where fronts and canyons are abundant but no individuals were predicted in the northern part of the study area. Despite an absence of sightings in the Mediterranean Sea, probabilities of presence were predicted in the Tyrrhenian and the Ionian Seas and in the north of the Levantine basin.

With the SEAPODYM model, highest kogiid probabilities of presence were predicted for high biomasses of bathypelagic and epipelagic organisms and a low biomass of highly migrant bathypelagic and plankton organisms (Fig. E.5B). This indicated a preference for fairly high productive habitats and resulted in a prediction of highest probabilities of presence near the Gulf Stream and the Ecuador in the Atlantic Ocean.

Highest probabilities of presence were predicted for low production of bathypelagic organisms, at depths *ca.* 4,000 m and high spatial gradients of SST and variations of the eddy kinetic energy with the MIXED model (Fig. E.5C). This indicated a preference for habitat associated with high depths, a low production of deeper organisms, dynamic waters and canyons and seamounts habitats. This resulted in similar patterns as for the ENVIRONMENTAL model but with lower probabilities of presence.

Interpolation areas varied between the three models and were less extended for the SEAPODYM (81%) and MIXED models (77%) compared to the ENVIRONMENTAL model (92%; Fig. E.5.2).

The model performance assessment criteria were slightly better for the MIXED model (Fig. E.5.1).



**Fig. E.5. Functional relationships for the selected variables (1/) and the predicted probabilities of presence of kogiids (2/) for the ENVIRONMENTAL model (A), the SEAPODYM model (B) and the MIXED model (C). 1/:** Solid lines are the estimated smooth functions, and the shaded regions represent the approximate 95% confidence intervals. The y-axes indicate the probabilities of presence on a log scale, where zero indicates no effect of the covariate. The vertical lines indicate the 2.5<sup>th</sup> and 97.5<sup>th</sup> quantiles of the data. D\*: explained deviance; AUC: Area Under the receiving Curve; TSS: True Skill Statistics. 2/: Black areas on prediction maps represent zones where we did not extrapolate the predictions. Percentages represent the proportion of the study area defined as interpolation with the gap analysis.

## References

- Abecassis, M., Senina, I., Lehodey, P., Gaspar, P., Parker, D., Balazs, G., Polovina, J. (2013). A model of loggerhead sea turtle (*Caretta caretta*) habitat and movement in the oceanic North Pacific. *PLoS One* 8(9): e73274.
- Allouche, O., Tsoar, A., Kadmon, R. (2006). Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of applied ecology* 43(6): 1223-1232.
- Austin, D., Bowen, W.D., McMillan, J.I., Iverson, S.J. (2006). Linking Movement, Diving, and Habitat to Foraging Success in a Large Marine Predator. *Ecology* 87: 3095–3108.
- AVISO <https://www.aviso.altimetry.fr/en/data/products/sea-surface-height-products/global/madt-h-uv.html>
- Canada Meteorological Centre <https://podaac.jpl.nasa.gov/dataset/CMC0.2deg-CMC-L4-GLOB-v2.0>
- Clark, M. (2013). Generalized additive models: getting started with additive models in R. Center for Social Research, University of Notre Dame, 35.
- Conchon, A. (2016). Modeling marine zooplankton and micronekton (Doctoral dissertation, La Rochelle University).
- Cotté, C., Guinet, C., Taupier-Letage, I., Mate, B., Petiau, E. (2009). Scale-dependent habitat use by a large free-ranging predator, the Mediterranean fin whale. *Deep Sea Research Part I: Oceanographic Research Papers* 56: 801–811.
- 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(1): 43–57.
- Ferguson, M.C., Barlow, J., Reilly, S.B., Gerrodette, T. (2006). Predicting Cuvier's (*Ziphius cavirostris*) and Mesoplodon beaked whale population density from habitat characteristics in the eastern tropical Pacific Ocean. *Journal of Cetacean Research and Management* 7: 287–299.
- GEBCO <http://www.gebco.net/>
- Harris, P.T., Macmillan-Lawler, M., Rupp, J., Baker, E.K. (2014). Geomorphology of the oceans. *Marine Geology* 352: 4–24.
- Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science* 3: 297-313.
- Hays, G. C. (2003). A review of the adaptive significance and ecosystem consequences of zooplankton diel vertical migrations. In *Migrations and Dispersal of Marine Organisms* (pp. 163-170). Springer, Dordrecht.
- Jaquet, N., and Whitehead, H. (1996). Scale-dependent correlation of sperm whale distribution with environmental features and productivity in the South Pacific. *Marine ecology progress series*: 1-9.
- Jennings, M.D. (2000). Gap analysis : concepts, methods, and recent results. *Landscape Ecology* 15: 5–20.

- Lambert, C., Mannocci, L., Lehodey, P., Ridoux, V. (2014). Predicting cetacean habitats from their energetic needs and the distribution of their prey in two contrasted tropical regions. *PLoS one* 9(8): e105958.
- Lehodey, P., Andre, J.M., Bertignac, M., Hampton, J., Stoens, A., Menkès, C., Memeryn L., Grima, N. (1998). Predicting skipjack tuna forage distributions in the equatorial Pacific using a coupled dynamical bio-geochemical model. *Fisheries Oceanography* 7(3-4): 317-325.
- Lehodey, P., Senina, I., Murtugudde, R. (2008). A spatial ecosystem and populations dynamics model (SEAPODYM) – Modeling of tuna and tuna-like populations. *Progress in Oceanography* 78(4): 304-318.
- 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): 69-84.
- Lehodey, P., Conchon, A., Senina, I., Domokos, R., Calmettes, B., Jouanno, J., Hernandez, O., Kloser, R. (2014). Optimization of a micronekton model with acoustic data. *ICES Journal of Marine Science* 72(5): 1399-1412.
- Mannocci, L., Catalogna, M., Dorémus, G., Laran, S., Lehodey, P., Massart, W., Monestiez, P., Van Canneyt, O., Watremez, P., Ridoux, V. (2014). Predicting cetacean and seabird habitats across a productivity gradient in the South Pacific gyre. *Progress in Oceanography* 120: 383–398.
- OREGONSTATE <http://orca.science.oregonstate.edu/1080.by.2160.8day.hdf.vgpm.m.chl.m.sst.php>).
- Österblom, H., Olsson, O., Blenckner, T., Furness, R.W. (2008). Junk-food in marine ecosystems. *Oikos* 117(7): 967-977.
- Redfern, J.V., Ferguson, M.C., Becker, E.A., Hyrenbach, K.D., Good, C. et al. (2006). Techniques for cetacean – habitat modeling. *Marine Ecology Progress Series* 310: 271–295.
- Virgili, A., Lambert, C., Pettex, E., Dorémus, G., Van Canneyt, O., Ridoux, V. (2017). Predicting seasonal variations in coastal seabird habitats in the English Channel and the Bay of Biscay. *Deep Sea Research II* 141: 212-223.
- Wood, S.N. (2006). On confidence intervals for generalized additive models based on penalized regression splines. *Australian & New Zealand Journal of Statistics* 48: 445–464.
- Wood, S. (2013). mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation. Retrieved 7 July 2014, from <http://cran.r-project.org/web/packages/mgcv/index.html>.





# MODELISER LES DISTRIBUTIONS DES ESPECES MARINES RARES : LES CETACES GRANDS PLONGEURS

**Résumé:** Les cétacés grands plongeurs, cachalots *Physeteridae* et *Kogiidae* et baleines à bec *Ziphiidae*, sont des espèces marines rares. Leur faible densité, aire de distribution étendue et faible disponibilité en surface génèrent de faibles taux d'observations. Cette particularité constitue un défi pour la modélisation d'habitat de ces espèces, préalable à leur conservation. Les modèles doivent gérer l'abondance de zéros qui limitent leur capacité à inférer des résultats écologiquement cohérents. Cette thèse vise donc à trouver une méthodologie adaptée aux jeux de données abondants en zéros, à déterminer comment les variables environnementales influencent la distribution des grands plongeurs et à prédire les zones potentielles qu'ils utilisent. Tester la capacité de prédiction de différents modèles d'habitat confrontés à un nombre décroissant d'observations a permis de souligner la pertinence d'un modèle, même si un minimum de 50 observations est nécessaire pour fournir des prédictions fiables. Des données issues de différentes campagnes visuelles ont été assemblées afin de produire les premières cartes de densités de grands plongeurs à l'échelle de l'océan Atlantique Nord et la mer Méditerranée. Les densités les plus élevées sont prédites dans les eaux entre 1500 et 4000 m de profondeur et près des fronts thermiques, particulièrement le long des pentes continentales et à l'ouest de l'océan Atlantique Nord. Par ailleurs, l'analyse de la transférabilité des modèles a montré une variation des habitats préférentiels en fonction des écosystèmes. Finalement, cette thèse permet de discuter les défis de la modélisation statistique appliquée aux espèces rares et les applications de gestion associées.

**Mots clés:** Espèces rares, modélisation d'habitat, cétacés grands plongeurs, baleines à bec, cachalots, kogiidés, océan Atlantique Nord, mer Méditerranée

# MODELLING DISTRIBUTIONS OF RARE MARINE SPECIES: THE DEEP-DIVING CETACEANS

**Summary:** Deep-diving cetaceans, sperm- and beaked whales *Physeteridae*, *Kogiidae* and *Ziphiidae*, are rare marine species. Due to their low densities, wide distribution ranges and limited presence at the water surface, visual surveys usually result in low sighting rates. This paucity of data challenges the modelling of their habitat, prerequisite for their conservation. Models have to cope with a great number of zeros that weakens the ability to make sound ecological inferences. Consequently, this thesis aimed at finding a methodology suitable for datasets with a large number of zeros, determining how environmental variables influence deep-diver distributions and predicting areas preferentially used by these species. By testing the predictive performance of various habitat models fitted to decreasing numbers of sightings, I selected the most suitable model and determined that at least 50 sightings were needed to provide reliable predictions. However, individual surveys can rarely provide sufficient deep-diver sightings thus I merged many visual survey datasets to produce the first basin-wide deep-diver density maps in the North Atlantic Ocean and the Mediterranean Sea. Highest densities were predicted in waters from 1500-4000 m deep and close to thermal fronts; hotspots were predicted along the continental slopes, particularly in the western North Atlantic Ocean. In addition, a model transferability analysis highlighted that habitat drivers selected by the models varied between contrasted large ecosystems. Finally, I discussed challenges related to statistical modelling applied to rare species and the management applications of this thesis.

**Keywords:** Rare species, habitat modelling, deep-diving cetaceans, beaked whales, sperm whales, kogiids, North Atlantic Ocean, Mediterranean Sea



Centre d'Etudes Biologiques de Chizé  
UMR 7372 Université de La Rochelle – CNRS  
17000 LA ROCHELLE



Crédit photo couverture: © Genavir (Campagne THALASSA, 2005)

Crédit illustrations : © Laura Hedon

## ABSTRACT

Deep-diving cetaceans, sperm- and beaked whales *Physeteridae*, *Kogiidae* and *Ziphiidae*, are rare marine species. Due to their low densities, wide distribution ranges and limited presence at the water surface, visual surveys usually result in low sighting rates. This paucity of data challenges the modelling of their habitat, prerequisite for their conservation. Models have to cope with a great number of zeros that weakens the ability to make sound ecological inferences. Consequently, this thesis aimed at finding a methodology suitable for datasets with a large number of zeros, determining how environmental variables influence deep-diver distributions and predicting areas preferentially used by these species. By testing the predictive performance of various habitat models fitted to decreasing numbers of sightings, I selected the most suitable model and determined that at least 50 sightings were needed to provide reliable predictions. However, individual surveys can rarely provide sufficient deep-diver sightings thus I merged many visual survey datasets to produce the first basin-wide deep-diver density maps in the North Atlantic Ocean and the Mediterranean Sea. Highest densities were predicted in waters from 1500-4000 m deep and close to thermal fronts; hotspots were predicted along the continental slopes, particularly in the western North Atlantic Ocean. In addition, a model transferability analysis highlighted that habitat drivers selected by the models varied between contrasted large ecosystems. Finally, I discussed challenges related to statistical modelling applied to rare species and the management applications of this thesis.

Centre d'Etudes Biologiques de Chizé  
UMR 7372 Université de La Rochelle – CNRS  
17000 LA ROCHELLE



Centre d'Etudes  
Biologiques de  
Chizé

