



Universidad Internacional de La Rioja
Facultad de Ciencias de la Salud

Máster Universitario en Bioinformática

**Modelado Predictivo y comparativa de la
distribución de cetáceos migratorios
(Balaenoptera musculus, Physeter
macrocephalus y Orcinus orca) mediante
Algoritmos de Aprendizaje Automático.**

Trabajo fin de Estudio presentado por:	Marc Casas Sebastián
Tipo de trabajo:	Tipo 3
Director/a:	Ricardo Blazquez Encinas
Fecha:	06 - 06 - 2025

Resumen

Este estudio realiza el desarrollo de modelos predictivos basados en algoritmos de aprendizaje automático para estimar la distribución espacial de tres grandes cetáceos de amplio rango y cuyo comportamiento ecológico es distinto: la ballena azul (*Balaenoptera musculus*), el cachalote (*Physeter macrocephalus*) y la orca (*Orcinus orca*) a través de la integración de datos de presencia obtenidos de plataformas públicas (OBIS) y variables fisicoquímicas del medio marino (Copernicus Marine Service). El objetivo principal del estudio es analizar la respuesta diferencial de estas especies a los factores ambientales, evaluar la capacidad predictiva de distintos algoritmos y poder construir diferentes mapas de patrones migratorios que nos ayuden a entender la relación entre estos grandes mamíferos y su medio, para contribuir a la importancia ecológica de los patrones migratorios de los cetáceos y aportar herramientas útiles para su conservación en un contexto de cambio global.

Palabras clave: (Máximo 3 palabras) Aprendizaje automático, cetáceos, patrones migratorios.

Abstract

This study aims to develop predictive models based on machine learning algorithms to estimate the spatial distribution of three wide-ranging cetacean species with distinct ecological behaviors: the blue whale (*Balaenoptera musculus*), the sperm whale (*Physeter macrocephalus*), and the orca (*Orcinus orca*). The approach integrates presence records obtained from public platforms (OBIS) with physicochemical variables of the marine environment (Copernicus Marine Service). The primary objective is to analyze the species-specific responses to environmental factors, assess the predictive performance of various algorithms, and generate maps of migratory patterns to enhance our understanding of the relationship between these large marine mammals and their habitat. Moreover, the study seeks to improve the ecological role of cetacean migrations and provide valuable tools for their conservation in a context of global change.

Keywords: Machine learning, cetacean, distribution

Índice de contenidos

1. Introducción.....	7
2. Hipótesis.....	12
3. Objetivos	12
3.1. Objetivo general	12
3.2. Objetivos específicos.....	12
4. Metodología	13
4.1. Variables físico-químicas del agua	13
4.2. Datos de presencia de cetáceos.....	17
4.3. Procesamiento de datos.....	19
4.3.1. Carga de datos	19
4.3.2. Limpieza de datos y detección de valores faltantes	19
4.3.3. Exploración visual de la distribución de los datos.....	20
4.3.4. Variables utilizadas	29
5. Resultados	31
5.1. Descripción general del conjunto de datos final	31
5.1.1. Número total de registros y cobertura por especie.....	31
5.1.2. Distribución espacial y temporal.....	32
5.1.3. Balanceo de clases	32
5.2. Modelos de aprendizaje automático	33
5.2.1. Curvas ROC.....	34
5.2.2. Valores AUC	36
5.2.3. Importancia de las variables	38
5.3. Mapas de distribución potencial.....	40
5.4. Modelos de distribución MaxENT	42
6. Discusión.....	49
7. Conclusiones	55
Referencias bibliográficas.....	57

Índice de figuras

Ilustración 1: Utilización de la API para la extracción de los datos históricos de clorofila - A.....	14
Ilustración 2: Distribución de la media de las variables físicas en toda la serie temporal.	20
Ilustración 3: Diagramas de cajas para las variables físicas.....	21
Ilustración 4: Gráficos de distribución (histogramas) para cada variable física.	22
Ilustración 5: Mapa de calor de las variables físicas.	23
Ilustración 6: Histograma de la representación temporal de las diferentes especies.	24
Ilustración 7: Distribución global de cachalote (<i>Physeter macrocephalus</i>) con datos de OBIS.	25
Ilustración 8: Distribución global de orca (<i>Orcinus orca</i>) con datos de OBIS.	26
Ilustración 9: Distribución global de ballena azul (<i>Balaenoptera physalus</i>) con datos de OBIS.....	27
Ilustración 10: Mapa de la distribución global de las tres especies.	28
Ilustración 11: Gráfico comparativo de modelos y especies incluyendo clorofila.	33
Ilustración 12: Curvas ROC para <i>Physeter macrocephalus</i>	34
Ilustración 13: Curvas ROC para <i>Orcinus orca</i>	35
Ilustración 14: Curvas ROC para <i>Balaenoptera musculus</i>	35
Ilustración 15: Comparación de AUC por modelo y especie.	37
Ilustración 16: Importancia de variables por especie.	38
Ilustración 17: Mapa de distribución <i>Physeter macrocephalus</i>	40
Ilustración 18: Mapa de distribución para <i>Orcinus orca</i>	41
Ilustración 19: Mapa de distribución <i>Balaenoptera musculus</i>	42
Ilustración 20: Resultados de MaxENT para <i>Physeter macrocephalus</i>	43
Ilustración 21: Resultados de MaxENT para <i>Balaenoptera musculus</i>	44
Ilustración 22: Resultados de MaxENT para <i>Orcinus orca</i>	44
Ilustración 23: Mapa de distribución MaxENT de <i>Physeter macrocephalus</i>	46
Ilustración 24: Mapa de distribución MaxENT de <i>Orcinus orca</i>	47
Ilustración 25: Mapa de distribución MaxENT de <i>Balaenoptera musculus</i>	48

Índice de tablas

Tabla 1: Contiene las 11 variables del dataset global de datos estadísticos de Copernicus. ...	16
Tabla 2: Comparativa entre mallas 0.1º y 0.5º.....	30
Tabla 3: Valores AUC para cada especie y algoritmo.	36
Tabla 4: Resumen de la importancia de variables MaxENT.	45

1. Introducción

Los grandes cetáceos desempeñan un papel ecológico fundamental en los océanos. Como organismos altamente móviles, longevos y de un elevado nivel trófico, participan activamente en procesos ecológicos a gran escala, desde la regulación de cadenas tróficas hasta el transporte de nutrientes entre diferentes capas oceánicas [1] [2]. A través de sus movimientos migratorios estacionales, contribuyen a mantener un equilibrio ecosistémico de regiones distantes, conectando zonas de alta productividad con áreas de reproducción, hecho que los convierte en organismos clave para el funcionamiento del océano a escala global.

Además de su rol ecológico, los grandes cetáceos son considerados especies paraguas, dado que su protección implica indirectamente la conservación de un gran conjunto de especies asociadas que comparten sus hábitats. A su vez, estos también son especies bandera, carismáticas y fácilmente reconocibles, que generan interés social y político, lo cual favorece la movilización de recursos para la conservación de los hábitats marinos [3]. Proteger las rutas migratorias, zonas de alimentación y cría de especies de grandes mamíferos significa, en última instancia, proteger a comunidades enteras de organismos que dependen directa o indirectamente de estos hábitats.

Este estudio, se centra en tres especies de gran tamaño, movilidad y relevancia ecológica: la ballena azul (*Balaenoptera musculus*), el cachalote (*Physeter macrocephalus*) y la orca (*Orcinus orca*). La ballena azul, el animal más grande que ha existido, es una especie migratoria que recorre largas distancias entre zonas de alta productividad y áreas de reproducción. Su fuerte dependencia de las condiciones tróficas la convierte en un excelente modelo para estudiar el impacto de variables como la temperatura y la productividad [1]. El cachalote, por otro lado, representa una especie de buceo profundo cuya distribución está más asociada a características batimétricas y a la presencia de presas mesopelágicas como los calamares [4], [5]. Finalmente, la orca, aunque más versátil y ampliamente distribuida, ofrece un caso interesante por su variabilidad ecológica y comportamental, lo que permite evaluar la capacidad de los modelos para captar estrategias adaptativas complejas en distintos hábitats marinos [3], [6].

Además, estos organismos actúan como reguladores tróficos superiores, ejerciendo control top-down¹ sobre las presas y afectando indirectamente la estructura de los niveles inferiores de la red alimentaria [7]. Las Orcas, por ejemplo, pueden inducir cascadas tróficas al regular poblaciones de pinnípedos o peces, lo que a su vez altera la dinámica de los niveles tróficos inferiores [7]. De forma análoga, el cachalote afecta las comunidades mesopelágicas mediante la depredación de calamares y otros organismos profundos, lo cual contribuye a modificar las estructuras tróficas verticales [4] [6]. Estos mecanismos reguladores tienen implicaciones críticas para la biodiversidad, la productividad y la resiliencia del ecosistema marino.

Los desplazamientos estacionales entre zonas de alimentación en altas latitudes y áreas de reproducción en aguas más cálidas no solo reflejan adaptaciones fisiológicas complejas, sino que también determinan la distribución de otras especies marinas asociadas y condicionan procesos ecológicos clave, como la redistribución vertical de biomasa y la disipación de nutrientes [5]. Estos movimientos, sin embargo, se ven crecientemente amenazados por presiones antropogénicas como la contaminación acústica, las colisiones con embarcaciones, el cambio climático y la fragmentación de hábitats, lo que convierte a los grandes cetáceos en indicadores sensibles de la salud del océano [8].

Comprender y predecir con precisión los patrones de distribución y migración es, por tanto, esencial para generar estrategias de conservación efectivas, delimitar áreas marinas protegidas y anticipar conflictos emergentes entre la megafauna marina y las actividades humanas en el mar como el tráfico marítimo, la contaminación acústica o la sobrepesca, entre otros [9]. Tradicionalmente, los patrones migratorios de grandes especies, se han basado en métodos tales como avistamientos visuales, marcaje satelital, la foto-identificación o el análisis genético, herramientas valiosas pero muy limitadas por su coste, cobertura espacial y representatividad muestral [1] [10].

En los últimos años, el desarrollo de modelos de distribución de especies (SDMs) ha permitido integrar registros de presencia con variables ambientales; temperatura superficial del mar

¹ Cuando los organismos de niveles tróficos superiores influyen la abundancia de las especies que ocupan el nivel trófico anterior. - https://www.uniovi.es/ranadon/Ricardo_Anadon/docencia/cascaras/Leccion21.html

(SST), batimetría, salinidad o concentración de clorofila-a para estimar la idoneidad del hábitat de las especies marinas [11] [12].

El aprendizaje automático, y en particular los modelos de aprendizaje supervisado, han abierto nuevas vías para abordar problemas ecológicos complejos. Estos algoritmos permiten detectar patrones no lineales y multivariantes que serían difíciles de capturar con métodos estadísticos tradicionales [9]. Entre los algoritmos más prometedores para tareas de clasificación y predicción en ecología marina se encuentran: Random Forest, las Máquinas de Soporte (SVM), los Boosted Regression Trees (GBM) y las redes neuronales profundas (deep learning). Estos pueden entrenarse con grandes volúmenes de datos heterogéneos y, mediante técnicas de validación cruzada, ofrecen resultados robustos y generalizables [11].

Uno de los elementos críticos para modelar la distribución y los patrones migratorios de los cetáceos es la identificación de variables ambientales relevantes. Diversos estudios han mostrado que factores como la temperatura superficial del mar (SST), la salinidad, la batimetría, la productividad primaria (frecuentemente representada por la concentración de clorofila-a) y la velocidad de las corrientes oceánicas, pueden influir significativamente en la presencia y los desplazamientos de estos organismos marinos [11], [12]. Estas variables reflejan condiciones físico-químicas del hábitat que condicionan la disponibilidad de presas y la idoneidad del entorno para funciones clave como la alimentación, reproducción o descanso [5]. Por tanto, su correcta selección y cuantificación son determinantes para generar modelos robustos y fiables que permitan predecir los nichos ecológicos y las trayectorias migratorias de estos organismos.

Recientemente el desarrollo tecnológico y las herramientas bioinformáticas, han permitido el acceso y el auge de las técnicas de aprendizaje automático, permitiendo capturar relaciones no lineales y de alta complejidad entre las condiciones oceánicas y los avistamientos de cetáceos, con resultados prometedores en términos de precisión predictiva y escalabilidad [9]. Estas técnicas, permiten no solo estimar la distribución potencial de las especies estudiadas, sino, identificar regiones críticas para su conservación hecho que mejora la resolución espacial y temporal de los mapas de distribución, optimiza la gestión de áreas marinas protegidas y permite anticipar desplazamientos o conflictos potenciales en contextos de cambio climático o mitigar impactos de las actividades humanas [1], [8]. Además, la combinación de múltiples fuentes de datos, desde datos provenientes de satelitales hasta

bases de datos públicas como OBIS, proporcionan un marco integrado que favorece la toma de decisiones a múltiples escalas.

A pesar de su potencial, la implementación de modelos de aprendizaje automático en ecología marina enfrenta retos importantes. Entre ellos se incluyen la escasez de datos etiquetados de calidad, los sesgos espaciales y temporales de las observaciones, y la necesidad de interpretar biológicamente modelos complejos [9]. En un futuro, las investigaciones deberán centrarse en desarrollar enfoques híbridos que combinen la precisión de los modelos de ML con marcos conceptuales ecológicos sólidos, que fomenten la recopilación estandarizada de datos y garanticen la accesibilidad y reproducibilidad de los modelos generados.

En este contexto, la bioinformática aplicada a la conservación marina se consolida como un campo interdisciplinar con gran potencial para abordar desafíos ecológicos complejos mediante la integración de datos ambientales, herramientas computacionales y algoritmos entrenados [12], [13]. La combinación de modelos predictivos, sensores remotos y bases de datos, posibilita la generación de hipótesis contrastables sobre la distribución de especies altamente móviles, incluso en regiones donde los datos son escasos, sesgados o poco sistematizados [14].

Este trabajo se enmarca en esta línea de investigación, proponiendo un enfoque comparativo basado en algoritmos de aprendizaje supervisado para inferir la distribución espacial potencial de tres cetáceos de amplio rango y diferencias: la ballena azul (*Balaenoptera musculus*), el cachalote (*Physeter macrocephalus*) y la orca (*Orcinus orca*). Esta selección responde tanto a su relevancia ecológica como a su variabilidad comportamental: desde migraciones obligadas condicionadas por la productividad primaria en el caso de la ballena azul [15], hasta comportamientos de buceo profundo y estrategias sociales complejas en el cachalote y la orca, respectivamente [2], [16].

Para ello, se aplicarán algoritmos de aprendizaje supervisado como Random Forest, Support Vector Machines (SVM), XGBoost, Naive Bayes, seleccionados por su capacidad para capturar relaciones no lineales y de alta dimensionalidad, superando en muchos casos a modelos estadísticos tradicionales [17]. La validación cruzada y el uso de métricas como el AUC permitirán evaluar la capacidad predictiva y generalización de los modelos construidos [18].

Durante el trabajo, se explorará la importancia relativa de cada variable predictora para cada especie, lo que permitirá identificar diferencias interespecíficas en la respuesta ecológica a las condiciones oceánicas. Estudios recientes han demostrado, por ejemplo, que el cachalote responde con mayor sensibilidad a variables oceánicas a profundidad (como la temperatura entre 200 y 600 m o la energía cinética de remolinos (EKE)) en comparación con especies que se alimentan en la columna superficial [5]. Esta perspectiva mejora la precisión ecológica del modelado y permite refinar mapas de distribución frente a enfoques basados exclusivamente en variables superficiales.

En conjunto, este estudio busca contribuir al conocimiento ecológico y la distribución de tres grandes especies de cetáceos desde una aproximación computacional y comparativa, demostrando el potencial de los modelos predictivos como herramientas complementarias a los métodos tradicionales de seguimiento. La combinación de datos abiertos, modelado mediante herramientas de aprendizaje automático y el estudio y conocimiento sobre ecología marina, se presentan como una vía prometedora para la gestión adaptativa de estas especies altamente móviles en un océano en constante transformación y movimiento.

2. Hipótesis

Las tres especies estudiadas (Ballena azul, Cachalote y Orca) tienen distintos patrones migratorios que dependen de la especie y los parámetros fisicoquímicos del agua.

Es posible determinar el comportamiento migratorio de las tres especies mediante algoritmos de aprendizaje supervisado y el uso de datos existentes en bases de datos públicas.

3. Objetivos

3.1. Objetivo general

Desarrollo de un modelo predictivo utilizando técnicas de aprendizaje supervisado para analizar y comparar la distribución potencial de la ballena azul, cachalote y orca en función de variables ambientales y datos de presencia de avistamientos.

3.2. Objetivos específicos

- Recopilar y procesar datos ambientales y de presencia georreferenciada para las tres especies de interés.
- Filtrado de datos, análisis exploratorio y construcción de un conjunto de datos robusto.
- Desarrollar modelos de aprendizaje automático para predecir la distribución potencial de cada especie.
- Identificar factores ambientales clave que puedan influir en la distribución de cada especie y evaluar similitudes y diferencias entre ellas.
- Comparar el rendimiento de los modelos generados utilizando métricas estadísticas.
- Visualizar y mapear las áreas críticas de distribución y la eficacia del modelo.

4. Metodología

Para llevar a cabo el trabajo, se realizó una recopilación de datos públicos provenientes de diferentes bases de datos:

- OBIS (Sistema de Información Biogeográfica Oceánica): Proporciona datos de distribución de especies marinas a nivel mundial.
- Copernicus Marine Service: Ofrece datos de variables oceanográficas y ambientales de alta resolución.

De estas fuentes, se construye una base de datos con la información sobre la presencia de las tres especies objetivo (ballena azul, cachalote y orca) y se cruza esta información con una tabla con distintas variables ambientales para determinar relaciones entre las variables ambientales y la distribución de las tres especies estudio.

Todos los scripts que se utilizaran durante este trabajo se recogen en un repositorio público de github². A partir de este punto, cualquier script nombrado se puede encontrar en dicho repositorio para facilitar su replicación en un estudio similar.

4.1. Variables físico-químicas del agua

La extracción de los parámetros físico-químicos del agua se realizó con un sencillo script de Python (`extract_copernicus.py`) que realiza una descarga utilizando la API de Python de Copernicus Marine Service haciendo uso de la librería `copernicusmarine` que permite descargar datasets públicos mediante la referencia del `dataset_id` previamente seleccionado mediante la búsqueda filtrada en su catálogo web (Ilustración 1).

² <https://github.com/markuzkuz/predict-cetaceans.git>

```

1 import copernicusmarine
2
3 copernicusmarine.get(
4     dataset_id = 'cmems_obs-mob_glo_bgc-chl-poc_my_0.25deg-climatology_P1M-m'
5 )

```

Terminal Output:

```

589.2/589.2 kB 43.0 MB/s eta 0:00:00
Downloading six-1.17.0-py2.py3-none-any.whl (11 kB)
Downloading tzdata-2025.2-py2.py3-none-any.whl (1547 kB)
Downloading zipp-3.21.0-py2.py3-none-any.whl (9.8 kB)
Downloading deprecated-1.2.18-py2.py3-none-any.whl (18.0 kB)
Downloading locket-1.0.0-py2.py3-none-any.whl (14.4 kB)
Downloading wrapt-1.17.2-cp311-cp311-macosx_11_0_arm64.whl (38 kB)
Building wheels for collected packages: asciitree
Building wheel for asciitree (project.toml) ... done
Stored in directory: /Users/marc/Library/Caches/pip/wheels/71/c1/d4/228776d3b870d4df7852ed0d1a3c3d3b85d6e6c2b4a87
Successfully built asciitree
Installing collected packages: pytz, asciitree, zipp, wrapt, urllib3, tzdata, typing-extensions, tqdm, toolz, six, setuptools, semver, pyyaml, packaging, numpy, lxml, locke
t, jmespath, idna, fsspec, fasteners, cloudpickle, click, charset-normalizer, certifi, annotated-types, typing-inspection, requests, python-dateutil, pydantic-core, parit,
importlib-metadata, httpx, deprecated, pystac, pydantic, pandas, numcodecs, H5netcdf, dask, botocore, zarr, xarray, s3transfer, boto3, copernicusmarine
Attempting uninstall: setuptools
Found existing installation: setuptools 65.5.0
Uninstalling setuptools-65.5.0:
Successfully uninstalled setuptools-65.5.0
Successfully installed annotated-types-0.7.0 asciitree-0.3.3 boto3-1.37.33 botocore-1.37.33 certifi-2025.1.31 charset-normalizer-3.4.1 click-8.1.8 cloudpickle-3.1.1 coperni
cusmarine-2.0.1 dask-2025.1.0 deprecated-1.2.18 fasteners-0.19 fsspec-2025.3.2 H5netcdf-1.6.1 httpx-3.13.0 idna-3.10 importlib-metadata-8.6.1 jmespath-1.0.1 locket-1.0.0 lm
b-5.1.2 numcodecs-0.15.1 numpy-2.4 packaging-24.2 pandas-2.2.3 parts-1.4.2 pydantic-2.11.3 pydantic-core-2.30.1 pystac-1.12.2 python-dateutil-2.9.0.post0 pytz-2025.2 pyya
ml-6.0.2 requests-2.32.3 s3transfer-0.11.4 semver-3.0.4 setuptools-80.9.0 six-1.17.0 toolz-1.0.0 tqdm-4.67.1 typing-extensions-4.13.2 typing-inspection-0.4.0 tzdata-2025.2
urllib3-2.4.0 wrapt-1.17.2 xarray-2025.3.1 zarr-2.18.7 zipp-3.21.0
[notice] A new release of pip is available: 23.2.1 -> 25.0.1
[notice] To update, run: pip install --upgrade pip
In [vsn] marci:~$python3 extract_copernicus.py
INFO - 2025-04-17T17:28:32 - Downloading Copernicus Marine data requires a Copernicus Marine username and password, sign up for free at: https://data.marine.copernicus.eu/
register
Copernicus Marine username: boxicas@gmail.com
Copernicus Marine password:
INFO - 2025-04-17T17:28:32 - Selected dataset version: "202411"

```

Ilustración 1: Utilización de la API para la extracción de los datos históricos de clorofila - A.

Para este estudio, se han descargado cuatro conjuntos de datos principales: CLOROPHIL_A_MULTIOBS_GLO_BIO_BGC_3D_REP_015_010³ correspondiente a información de la distribución de Clorofila – A (chl), partículas orgánicas (poc) y coeficiente de retrodispersión de partículas (bbp) a nivel global de forma mensual des de 1998–2022 y hasta 1000 metros de profundidad. El segundo dataset con identificador PH_SEALEVEL_GLO_PHY_MDT_008_063⁴ corresponde a un modelo de altura media del mar (MDT - Mean Dynamic Topography) y densidad superficial a nivel global con temporalidad de una única media climatológica estática de 20 años (1993–2013). El tercer dataset con identificador GLOBAL_OMI_HEALTH CARBON_PH_TREND_1985_P20230930⁵ que corresponde a un indicador oceánico (Ocean Monitoring Indicator or OMI) de cambios medios de valores de pH superficial oceánico (año-1) calculado con una resolución de 0,25° × 0,25° desde 1985 a 2024. Este indicador proviene de series temporales mensuales de pH (ph_trend) distribuidas con el producto Copernicus Marine MULTIOBS_GLO_BIO CARBON_SURFACE_REP_015_008 [19]. Para cada celda de la cuadrícula, se ha utilizado una regresión lineal de mínimos cuadrados para ajustar una función lineal de pH en función del tiempo, donde la pendiente (μ) y la desviación estándar residual

³ https://data.marine.copernicus.eu/product/MULTIOBS_GLO_BIO_BGC_3D_REP_015_010/services

⁴ https://data.marine.copernicus.eu/product/SEALEVEL_GLO_PHY_MDT_008_063/services

⁵ https://data.marine.copernicus.eu/product/GLOBAL_OMI_HEALTH_carbon_ph_trend/services

(σ) se definen como estimaciones de la tendencia a largo plazo con su incertidumbre asociada (ph_trend_uncertainty).

Por último, el dataset con identificador GLOBAL_MULTIYEAR_PHY_001_030⁶ que corresponde a un dataset global y multianual que contiene todo un registro de reanálisis físico con datos de temperatura y salinidad mensuales des del 01-01-1993 hasta el 31-06-2021.

Todos estos datos, se encuentran en formato .nc que fueron procesados con el software Python (Version 3. 11) para poder unificar y analizar el conjunto de variables fisicoquímicas y convertirlas a un formato CSV antes de cruzarlas con las variables biológicas de OBIS para poder realizar los distintos modelos predictivos.

Respecto al conjunto de datos GLOBAL_MULTIYEAR_PHY_001_030 que consta de un conjunto de datos espaciales y temporales muy grande (des de 1993 hasta el 2021), el conjunto original de datos era de 447GB que fueron almacenados en un dispositivo NAS Sinology y que debían ser procesados y analizados. Para superar este primer inconveniente, después de comprobar la estructura y disposición de los datos mediante el software Python (open.py) abriendo el primer archivo .nc y haciendo un print de las variables, se identificaron las siguientes once variables principales (Tabla 1).

⁶ https://data.marine.copernicus.eu/product/GLOBAL_MULTIYEAR_PHY_001_030/services

Tabla 1: Contiene las 11 variables del dataset global de datos estadísticos de Copernicus.

Variable	Descripción	Dimensiones	Unidades
mlofst	Espesor de la capa de mezcla oceánica	lon, lat, time	metros (m)
zos	Altura de la superficie del mar	lon, lat, time	metros (m)
bottomT	Temp. potencial en el fondo marino	lon, lat, time	°C
sithick	Espesor de hielo marino	lon, lat, time	metros (m)
siconc	Concentración de hielo marino	lon, lat, time	fracción (0-1)
usi	Velocidad hacia el este del hielo marino	lon, lat, time	m/s
vsi	Velocidad hacia el norte del hielo marino	lon, lat, time	m/s
thetao	Temp. potencial del agua en profundidad	lon, lat, depth, time	°C
so	Salinidad	lon, lat, depth, time	PSU
uo	Velocidad este del agua	lon, lat, depth, time	m/s
vo	Velocidad norte del agua	lon, lat, depth, time	m/s

De este dataset, se realizó un script de Python (nc_to_csv.py) para:

- Iterar sobre las carpetas de los distintos años (1993–2021).
- Leer los 12 archivos .nc por año ignorando los archivos ocultos.
- Para cada archivo y variable seleccionada:
 - Abrir el archivo .nc
 - Encontrar la variable.
 - Extraer los datos.
 - Colapsar las dimensiones no espaciales (profundidad, tiempo).
 - Calcular la media, desviación estándar, IQR (rango intercuartílico), máximo y mínimo por cada variable en cada año.
- Guardar los resultados anuales en archivos csv (comma separated values).

De este modo, se consiguió reducir el dataset original a 6 documentos csv (uno por cada variable) con la representación de los valores descriptivos para cada variable y cada año desde 1993 a 2021.

A continuación, se realizó una extracción similar con el dataset referente a CLOROPHIL_A_MULTIOBS_GLO_BIO_BGC_3D_REP_015_010 que contiene datos desde 1998–2022 hasta 1000 metros de profundidad en forma de promedio mensual donde se extraen las estadísticas descriptivas para cada mes de las distintas variables del dataset (chl, bbb, poc y sus respectivos errores). Este dataset no contiene los datos desglosados anualmente, sino que cada archivo es un promedio mensual basado en los datos de 1998 a 2022, no un año específico (promedio mensual climatológico). Por ello, se extraen estadísticas descriptivas mensuales (media, desviación estándar, IQR, mínimo, máximo) para cada una de las variables incluidas: clorofila-a, bbb, POC y sus respectivos errores (chloro.py).

Por último, el dataset PH_SEALEVEL_GLO_PHY_MDT_008_063 contiene información sobre nivel del mar (MDT) y otras variables relacionadas con la altura de ola y el movimiento de las aguas, lo cual está orientado al análisis de la dinámica oceánica. Este dataset, se usará a modo complementario para poder añadir variables físicas como el nivel del mar (mdt), o u y v para las velocidades, que son campos medios que representan la media del estado del mar durante 1993-2013 mediante el script (mdt_to_csv.py).

4.2. Datos de presencia de cetáceos

La obtención de los datos de presencia de las especies objeto de estudio se realizará a través de Ocean Biodiversity Information System (OBIS)⁷, una plataforma global de acceso libre que recopila y estandariza datos georreferenciados sobre la biodiversidad marina a partir de múltiples fuentes científicas.

Para la descarga de los registros de presencia, se empleó la herramienta mapper disponible en el portal web de OBIS. Esta herramienta permite realizar búsquedas personalizadas

⁷ <https://obis.org/data/access/#>

mediante la introducción del nombre científico de la especie deseada. En este caso, se buscarán las siguientes tres especies de interés:

- *Balaenoptera musculus* (ballena azul)
- *Physeter macrocephalus* (cachalote)
- *Orcinus orca* (orca)

Una vez introducido el nombre científico en el buscador del Mapper, se muestra un mapa interactivo con los registros de distribución conocidos des de 1750 hasta 2025. Desde esta misma interfaz, se procede a la descarga directa de los datos de presencia en formato .tsv (tab-separated values), que incluye coordenadas geográficas, fecha de observación, fuente de los datos e información relevante para su posterior análisis y cruce con las variables ambientales.

Este formato es fácilmente legible por software de análisis de datos como R o Python, lo que facilita su tratamiento posterior. Los archivos .tsv descargados se almacenarán de forma separada para cada especie y se procesarán en fases posteriores del análisis, previa limpieza, normalización y verificación de calidad de los registros (eliminación de duplicados, registros sin coordenadas, etc.). De todos los datos de cada especie disponibles en los archivos .tsv⁸, se seleccionarán las siguientes variables: *id* que corresponde a un identificador único para cada avistamiento, *date_year* que corresponde al año del avistamiento, *decimalLatitude* y *decimalLongitude* las coordenadas decimales de Latitud y Longitud, *bathymetry* o profundidad (si se aplica) basados en los datos de EMODnet Bathymetry y GEBCO⁹, *originalScientificName* que hace referencia a el nombre científico del animal, *basisOfRecord* dato categórico para saber si la observación es HumanObservation (humana) o Occurrence (de presencia), *coordinateUncertaintyInMeters* para saber la incertidumbre de la posición (parseado por OBIS) y *samplingProtocol* que indica el tipo de protocolo de muestreo que se ha utilizado.

De esta forma, se puede trabajar directamente con datos validados de biodiversidad marina a escala global, lo cual es fundamental para el modelado predictivo de la distribución de las especies seleccionadas en función de las variables físico-químicas del océano.

⁸ Campos de datos OBIS: <https://obis.org/data/access/>

⁹ <https://github.com/iobis/xylookup> (Data references).

4.3. Procesamiento de datos

Una vez descargados y convertidos los conjuntos de datos ambientales desde su formato original NetCDF (.nc) a archivos en formato CSV (.csv), se procede a su carga, limpieza y exploración preliminar con el objetivo de preparar los datos para su integración con los datos de presencia biológicos y su posterior modelado predictivo.

Para ello, se utilizó el entorno de programación R (Versión 2024.12.1+563) mediante RStudio, que ofrece un conjunto de herramientas eficaces para el análisis exploratorio y la visualización de datos (TFM_fisicas.R).

4.3.1. Carga de datos

Los archivos .csv generados a partir de los distintos conjuntos de datos ambientales (clorofila, salinidad, temperatura, velocidad del agua, pH) se importarán utilizando funciones como `read.csv()`.

Cada archivo contiene valores agregados anualmente por variable, incluyendo estadísticas descriptivas como la media, desviación estándar, máximo, mínimo y rango intercuartílico (IQR) para cada año en el periodo disponible.

4.3.2. Limpieza de datos y detección de valores faltantes

Una vez cargados, los datasets se someterán a una revisión exhaustiva de su estructura y consistencia, mediante funciones como `str()`, `summary()` y `glimpse()` para validar el formato y tipo de datos de cada columna.

Posteriormente, se procede a la detección de valores ausentes (NA) utilizando funciones como `is.na()` y `rowSums(is.na(datos))`, y se cuantifica su presencia por variable y por año para decidir si es necesario imputar, eliminar o transformar los valores faltantes.

También se procede a la normalización de los datos usando `scale(mean)` para poder escalarlos y poder comparar las variables entre sí.

4.3.3. Exploración visual de la distribución de los datos

Con el objetivo de comprender la distribución de las variables ambientales, se llevaron a cabo diversos análisis exploratorios utilizando el paquete ggplot2. Entre los gráficos iniciales se incluyen:

- Histogramas (`geom_histogram`) para observar la forma de la distribución (asimetría, modalidad, curtosis, etc.).
- Diagramas de caja (`geom_boxplot`) para detectar posibles outliers y comparar la dispersión por año o por variable.
- Gráficos de líneas para observar las tendencias temporales (en caso de variables con datos anuales o mensuales).

Estas representaciones gráficas permitirán identificar patrones, estacionalidades y posibles anomalías en los datos, fundamentales para ajustar los modelos predictivos en fases posteriores al análisis.

Después de la limpieza, filtrado y normalización de los datos para el conjunto de variables físicas globales, se pueden visualizar los distintos tipos de gráficos para observar como se distribuyen las distintas variables durante toda la serie temporal.

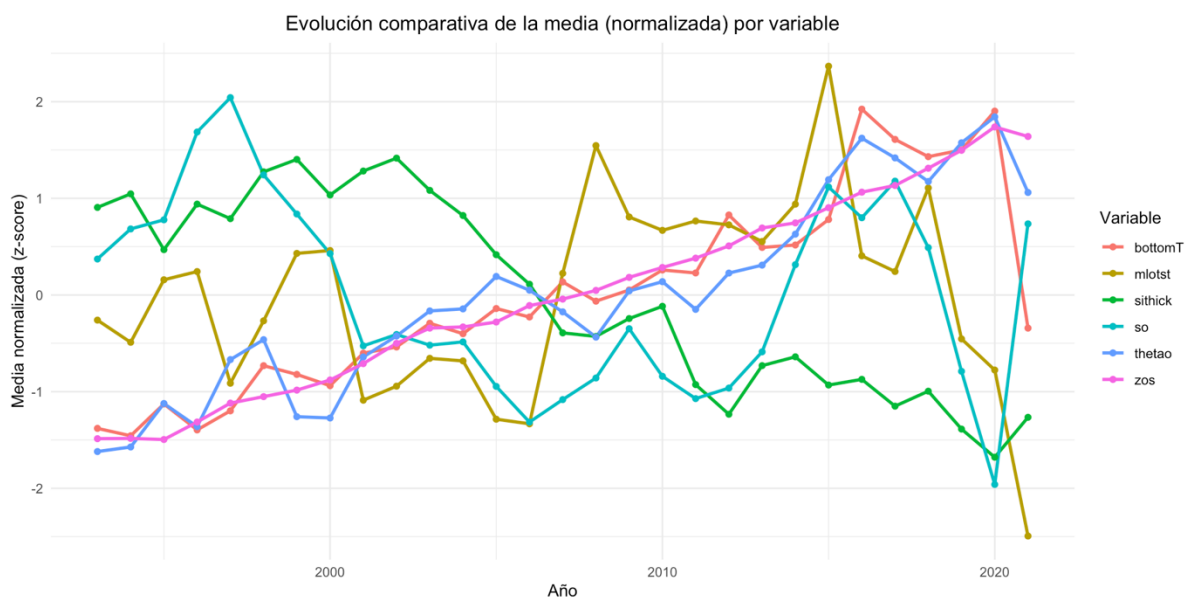


Ilustración 2: Distribución de la media de las variables físicas en toda la serie temporal.

Este primer gráfico (Ilustración 2), nos permite detectar patrones temporales en las distintas variables y observar cómo hay algunas variables que presentan un patrón en diente de sierra ya que aumentan y disminuyen sin un patrón claro cómo el espesor de la capa de mezcla oceánica (mlosts) o la salinidad (so), mientras que se observan otras variables que han ido aumentando conjuntamente a lo largo de los años cómo la temperatura del fondo (bottomT), zos (altura del nivel del mar) y la temperatura del agua en profundidad (thetao). Además, se observa una variable que tiene un patrón claramente descendente en lo largo del tiempo que se trata del espesor del hielo oceánico (sithick).

También se pueden realizar otro tipo de gráficos como “violin plots” o diagramas de cajas con estas mismas variables (Ilustración 3).

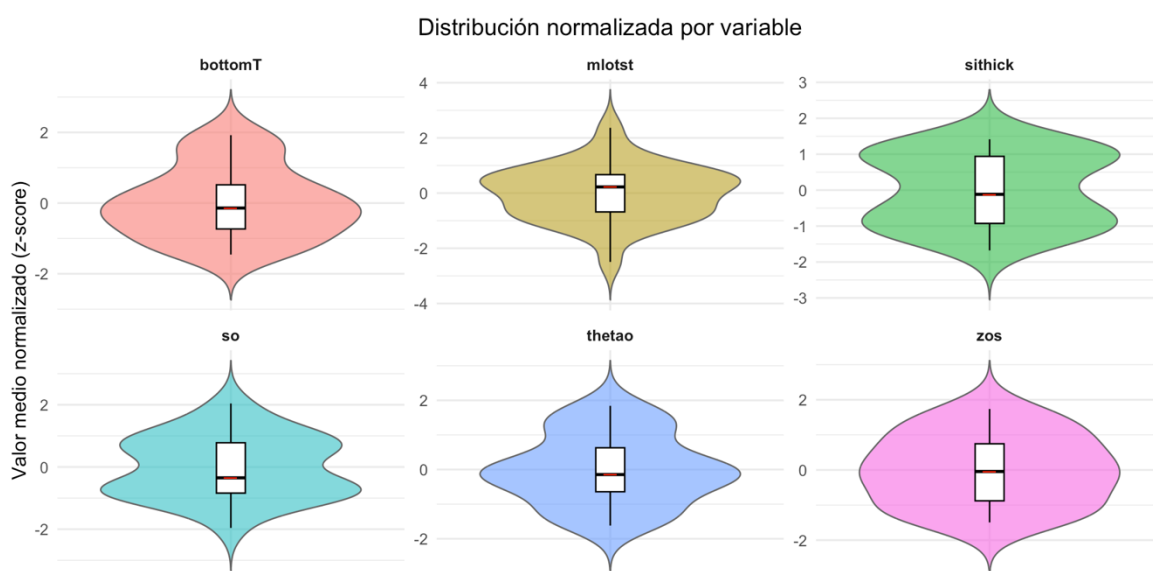


Ilustración 3: Diagramas de cajas para las variables físicas.

En estos diagramas de cajas, se ha optado por generar un solo gráfico por variable y juntarlos con la función `facet_wrap()` y se trabajó con los datos normalizados (Z-score) para facilitar la comparación entre variables ya tienen valores de magnitudes muy diferentes y en un solo gráfico no se observaban bien. Estos gráficos nos resultarán útiles para poder observar la escala normalizada de cada variable y su rango de valores plausible, así como la ubicación de la media, los valores outliers y los máximos y mínimos de cada variable.

A continuación, se realizó un histograma usando el mismo formato que el gráfico anterior con la librería de ggplot2 y la función `geom_histogram` dónde además se añade `geom_density` para conocer la curva de densidad de cada histograma (Ilustración 4).

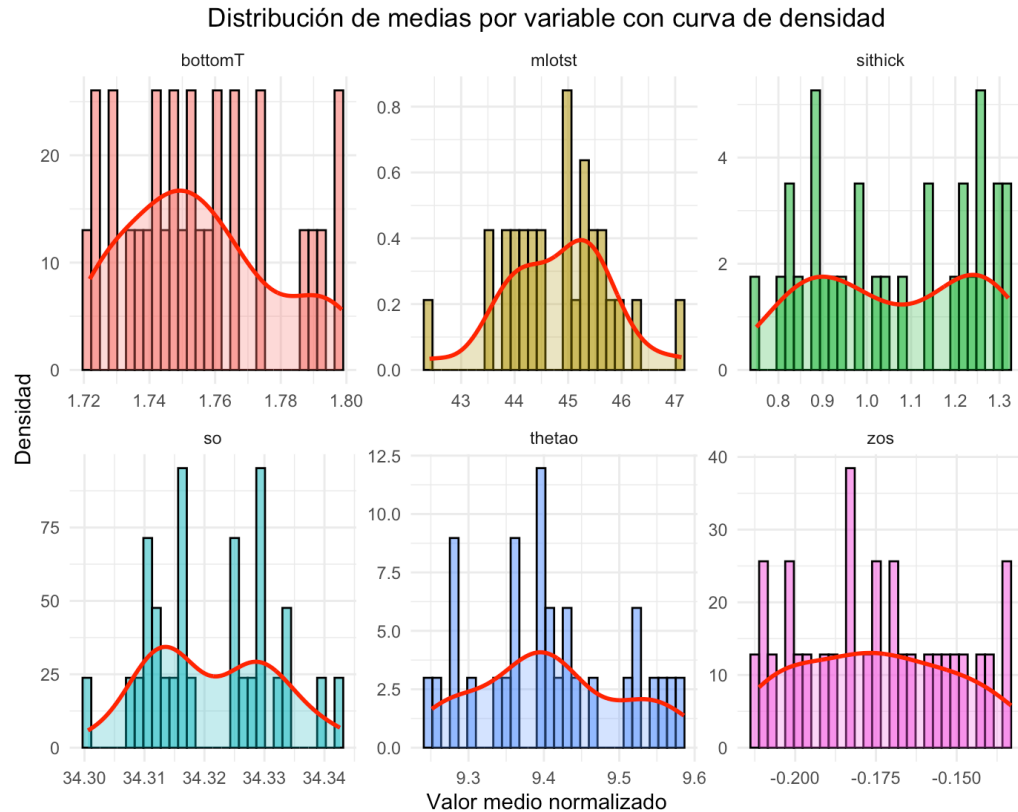


Ilustración 4: Gráficos de distribución (histogramas) para cada variable física.

Este tipo de gráfico de barras o histograma, resulta muy útil para conocer a simple vista la distribución que tiene cada variable dentro del conjunto de datos, así como observar su curtosis, simetría o asimetría y ver si se ajustan a una distribución normal, en forma de campana de Gauss. Hay variables que siguen una distribución normal a lo largo de los años (mlotst, thetao, zos e incluso bottomT) mientras que hay otras que claramente no siguen una distribución normal como so y sithick.

El último gráfico exploratorio de las variables físicas es una matriz de correlación y, a partir de esta, la construcción de un gráfico de calor o heatmap para poder observar visualmente si existe y de qué tipo es la relación entre las distintas variables (Ilustración 5).

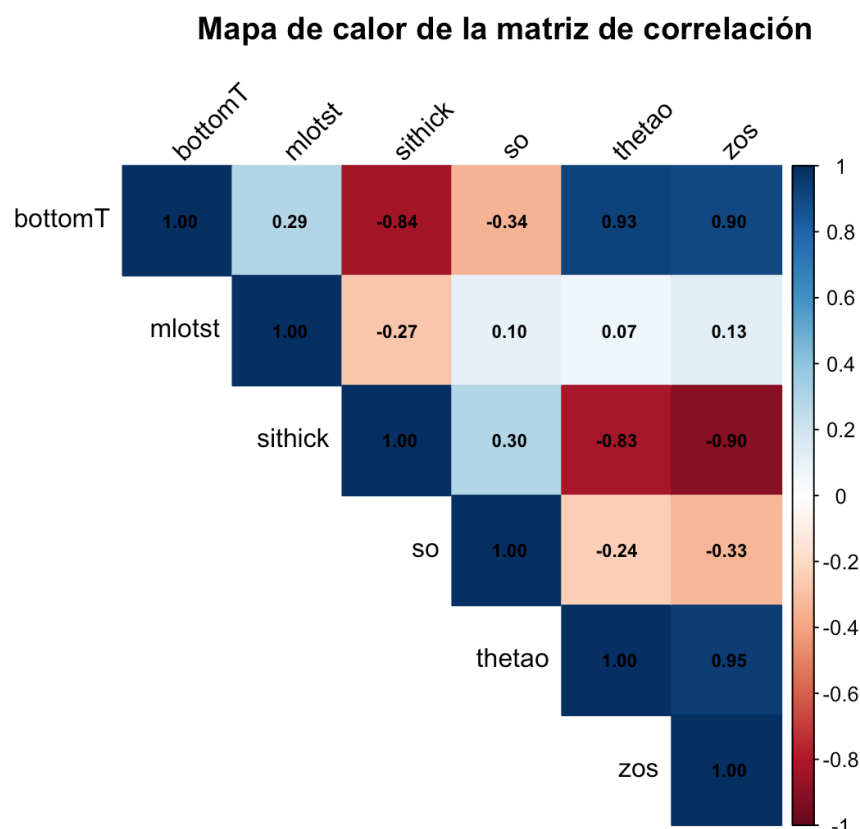


Ilustración 5: Mapa de calor de las variables físicas.

Gracias a este simple gráfico, podemos observar una paleta de colores que van del azul oscuro (máxima correlación) al rojo oscuro (inversamente correlacionados) la representación de cada variable frente a las otras. De esta forma, podemos ver cómo la temperatura del fondo se relaciona muy bien con la temperatura en profundidad (thetao) y altura del nivel del mar (zos) y de forma inversamente proporcional con el espesor de la capa de hielo (sithick).

A continuación, se realizó un proceso similar con los datos de la clorofila dónde se agruparon las distintas variables por mes (bbp, poc, chl y sus errores) y una vez comprobado si existían valores faltantes y tras su normalización, se realizaron diferentes resúmenes estadísticos para una primera exploración de los datos.

Por último, se analizaron los datos de presencia de cetáceos procedentes de OBIS que se descargaron directamente de su portal en forma de archivo .tsv. Una vez descargados, se procede a analizar su estructura y preparar los datos para cruzarlos con los datos espaciales (TFM_biologicas.R).

Los datos se cargaron en Rstudio mediante la función `read_tsv()` y analizando su estructura se observó que los diferentes conjuntos de datos son archivos tabulares muy grandes con un numero de columnas idéntico para las tres especies (282) y con un nombre de observaciones o filas que oscila entre 23.745 o 16.700 para el caso de la ballena azul y la orca y que puede llegar hasta 104.316 en el caso del cachalote. Otra observación referente al conjunto de datos de OBIS, es que pese a tener la misma estructura para las distintas especies, hay columnas que tienen un formato distinto en los distintos archivos y se realizó una conversión de las variables seleccionadas para poder juntar los tres conjuntos de datos de avistamientos en un solo dataframe y continuar así con los distintos análisis.

También se observa que hay bastantes columnas en los diferentes conjuntos de datos que tienen valores nulos (NA) que se excluirán para limpiar el conjunto de datos. El primer gráfico (Ilustración 6) que se realiza con el conjunto de datos debidamente filtrado y organizado, es un histograma que relaciona el número de registros u observaciones de avistamientos de cetáceos junto con las distintas especies representadas en el conjunto de datos global respecto al tiempo (en años) utilizando la librería de ggplot2 y la función `geom_histogram`.



Ilustración 6: Histograma de la representación temporal de las diferentes especies.

Además, se realizaron otros gráficos de carácter exploratorio como gráficos de dispersión, transformaciones logarítmicas y distintos test de normalidad que indican que nuestro conjunto de datos sigue una distribución no normal.

Otro tipo de gráfico que se construye con este mismo conjunto de datos es el gráfico de distribución de las distintas especies en función de su ubicación geográfica (Latitud, Longitud) y la especie en cuestión. Una primera versión de este gráfico fue realizada mediante el software R studio pero se ha optado por exportar los distintos conjuntos de datos de OBIS a .csv para facilitar su visualización mediante el software libre QGIS-LTR (V.3.40) mediante el cual se ha curado la resolución de las capas así como su escala y coordenadas para poder generar los siguientes mapas de puntos (Ilustración 7, 8 y 9).

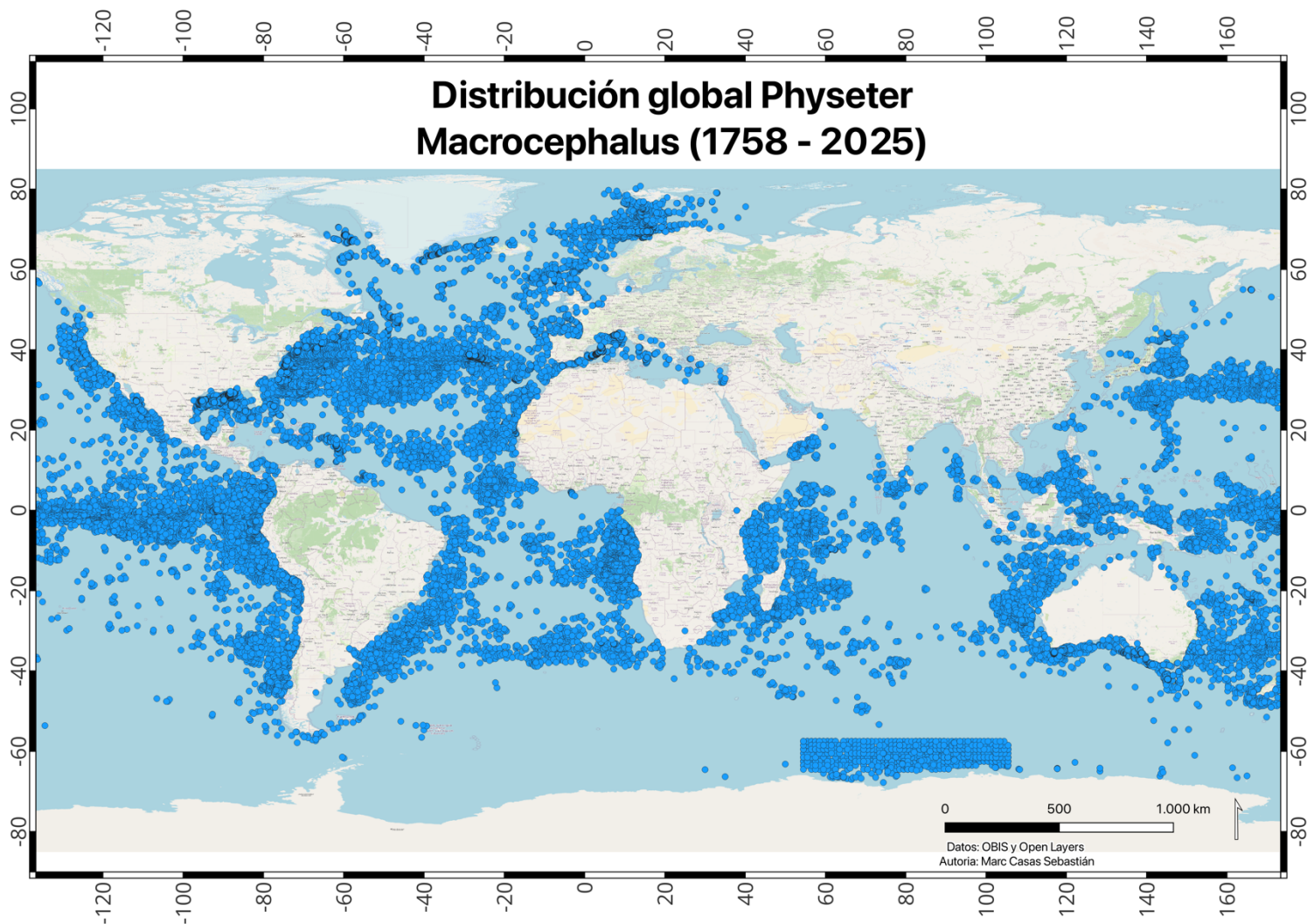
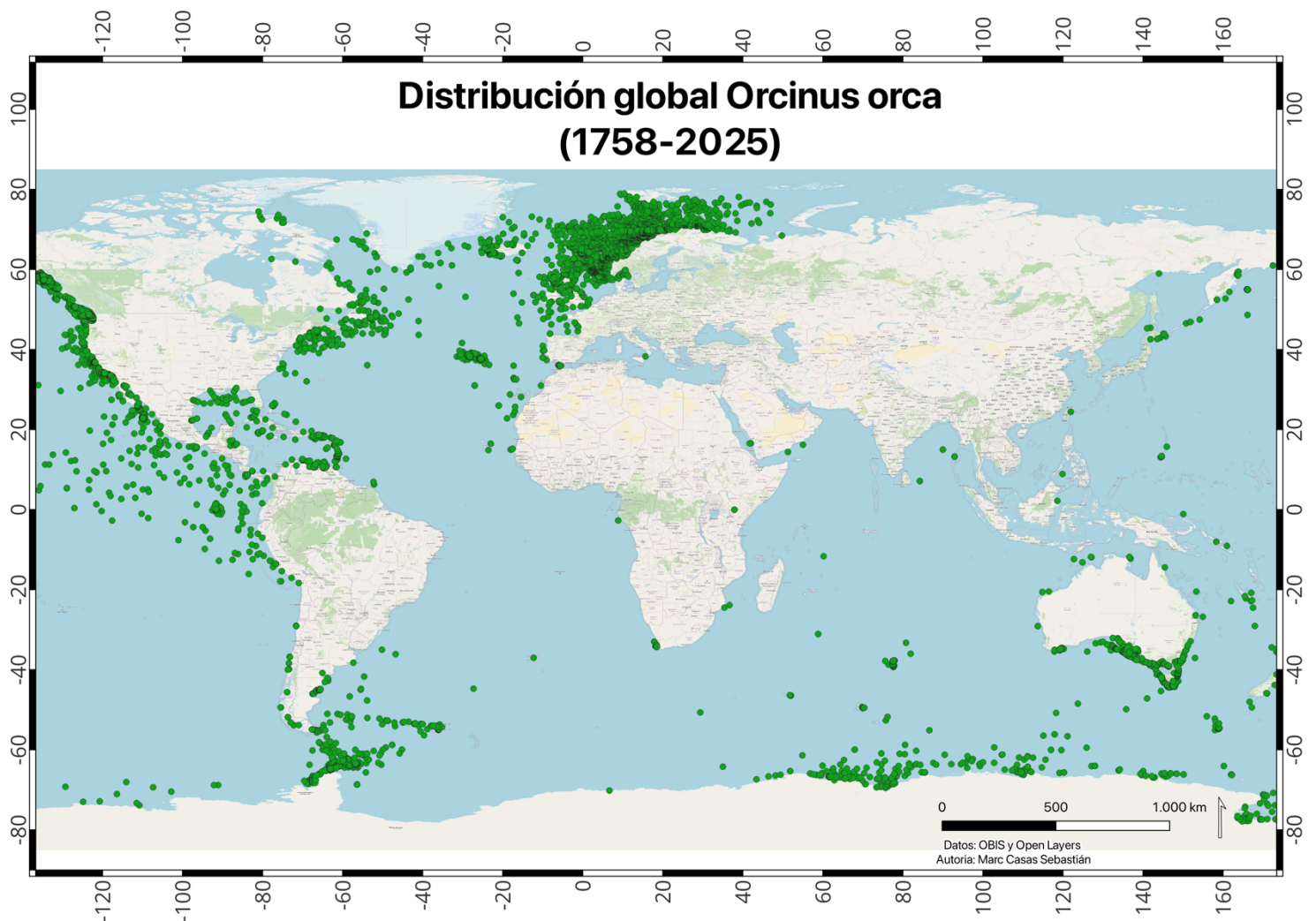


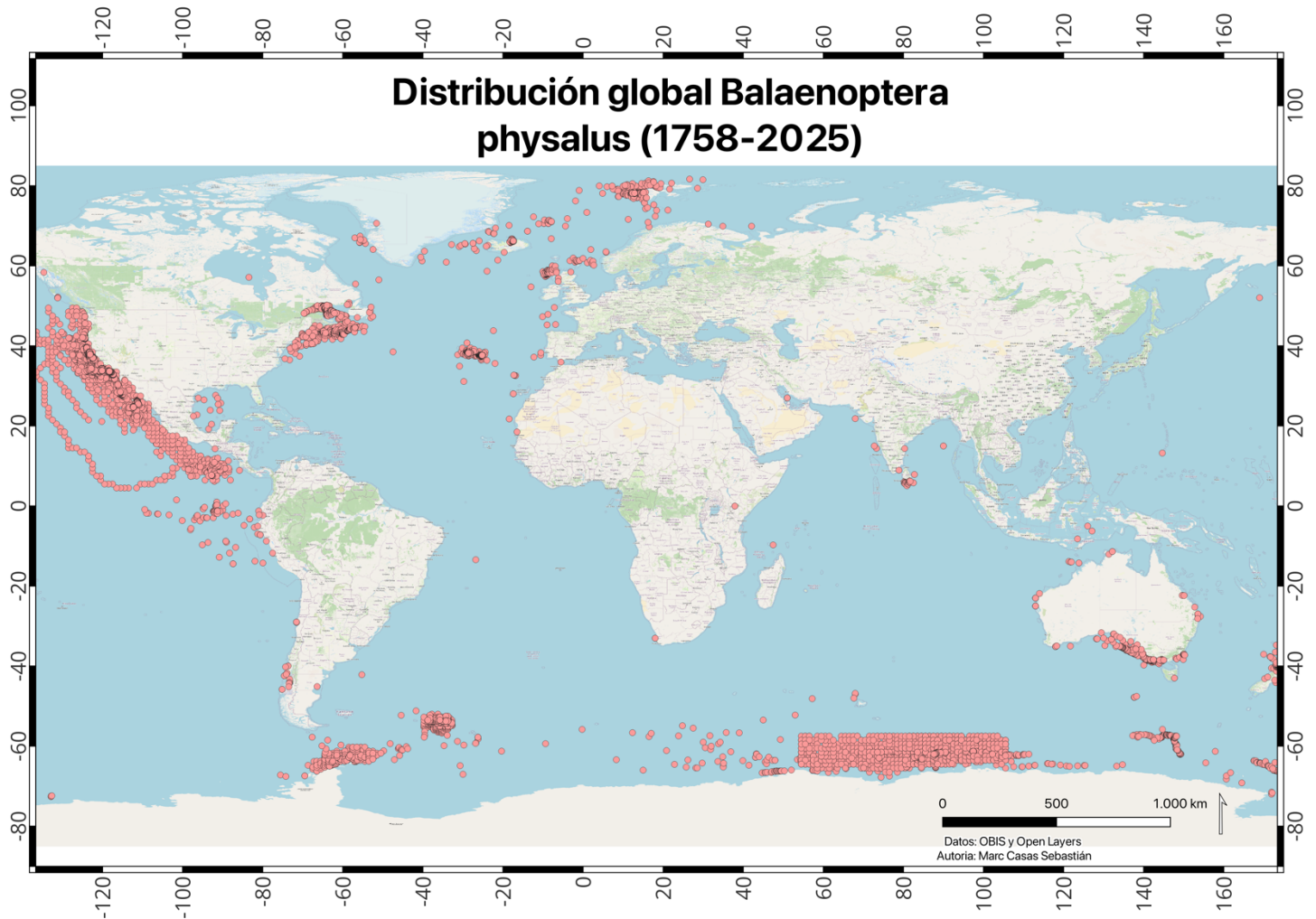
Ilustración 7: Distribución global de cachalote (*Physeter macrocephalus*) con datos de OBIS.

En este primer mapa (Ilustración 7) se observa la distribución a nivel global del cachalote durante los años 1758 – 2025 gracias a los datos obtenidos de OBIS dónde cada punto representa un avistamiento de la especie en cuestión (*Physeter macrocephalus*) en un punto geográfico. Para esta especie tenemos alrededor de 88.000 observaciones con su respectiva información asociada.



*Ilustración 8: Distribución global de orca (*Orcinus orca*) con datos de OBIS.*

Este segundo mapa (Ilustración 8), contiene 16.577 registros de avistamiento de orca desde 1758 hasta 2025 a nivel global dónde cada punto tiene toda la información asociada al avistamiento.



*Ilustración 9: Distribución global de ballena azul (*Balaenoptera physalus*) con datos de OBIS.*

Este tercer mapa de distribución (Ilustración 9), representa los distintos avistamientos de ballena azul (23.059) des de 1758 hasta 2025 con todos sus respectivos metadatos asociados al avistamiento.

Gracias a estos tres mapas realizados con la herramienta QGIS, se pueden empezar a observar diferencias y similitudes en la distribución de las distintas especies que se discutirán en el siguiente apartado junto a los resultados.

Por último, podemos realizar un mapa conjunto de las tres especies de estudio (Ilustración 10).

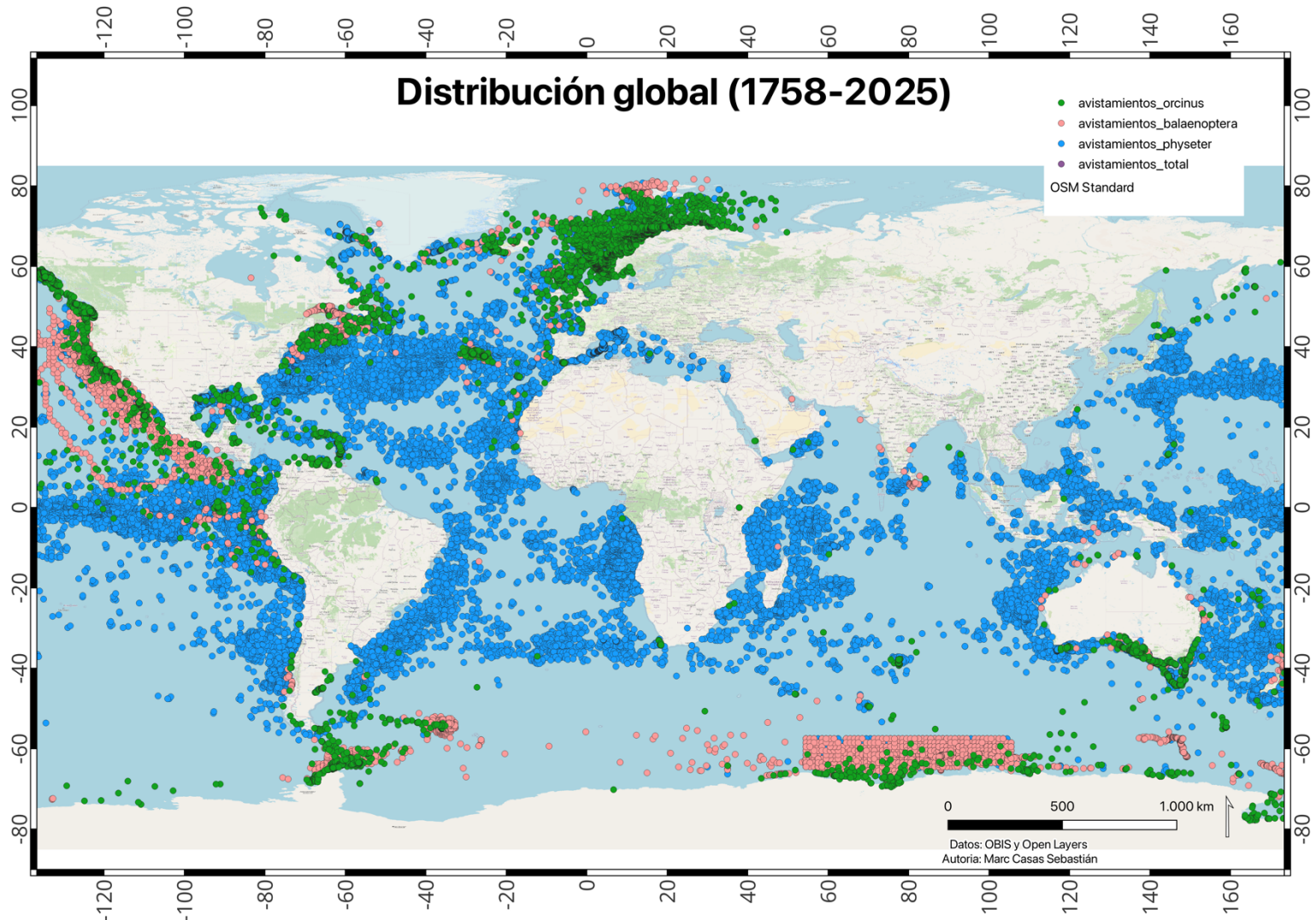


Ilustración 10: Mapa de la distribución global de las tres especies.

Este último mapa, utiliza un código de colores para representar la totalidad de las observaciones de cada especie a nivel global des de 1758 a 2025 hecho que nos facilita una idea generalizada de cómo se distribuyen y si hay patrones aparentes o diferencias entre las especies a estudiar y su localización. A modo de resumen, se puede entrever un cierto patrón similar entre las especies de Orca y Ballena azul mientras que el cachalote difiere bastante de dicho patrón con una distribución mucho más amplia.

4.3.4. Variables utilizadas

4.3.4.1. Variables dinámicas

Una vez se han explorado los distintos conjuntos de datos por separado, en este apartado se realizó la construcción del conjunto de datos final que se utilizará para los distintos modelos. Debido a la naturaleza de los datos, se divide el conjunto de datos en dinámicos para aquellas variables con registros anuales: `m1otst`, `bottomT`, `thetao`, `so`, `sithick`, `zos`, `siconc` dónde se realiza una extracción mediante un script de Python de los datos desde los archivos NetCDF de una temporalidad concreta (1993-2022) y se calculan los valores medios por cada año en cada punto espacial (coordenadas) que serán exportados en un archivo .csv con cuatro columnas por cada variable: **latitude**, **longitude**, **año** y **variable** dónde esta última columna contiene todos los valores de las medianas anuales por cada punto geográfico. Este cálculo, se realizó con otro script de Python usando Dask¹⁰ para poder optimizar el procesamiento en memoria ya que el conjunto de archivos original pesaba más de 400GB.

Los datos de avistamientos de cetáceos (procedentes de OBIS) también contienen una dimensión temporal anual así que se podrán juntar fácilmente al conjunto de datos de las variables dinámicas para la ejecución de los modelos de aprendizaje supervisado.

4.3.4.2. Variables estáticas

Otro tipo de dato que encontramos entre nuestras variables de estudio, son aquellos que provienen de valores medios entre los distintos años (1993-2022) en los diferentes archivos NetCDF: `ph_trend`, `phtrenduncertainty`, `u`, `v`, `mdt`. Estas variables no pueden aportar una dimensión temporal así que se calcularon las medianas anuales por cada punto espacial y fueron exportados en un archivo .csv con tres columnas por cada variable: `latitude`, `longitude` y `variable` dónde esta última columna contiene los valores de las medianas anuales en cada

¹⁰ Dask es una librería open-source escrita en Python que permite la computación paralela. - <https://www.dask.org>

punto. Este cálculo también se realizó mediante un script de Python parecido al de las variables dinámicas.

Además, también se utilizaron los datos del conjunto de clorofila con las variables: chl, bbp, poc, chl_error, bbp_error y poc_error que se trataron como datos estáticos sólo de la temporalidad concreta 1998-2022 y se incorporaron posteriormente como segundo análisis para no perder dimensionalidad temporal en los algoritmos de predicción y poder tener un mayor número de observaciones.

4.3.4.3. Malla espacial

A continuación, se decidió agrupar los distintos conjuntos de datos estáticos y dinámicos (por separado) haciendo uso de otro script de Python para unir los .csv individuales de cada variable en un solo archivo tabular, pero debido a la diferencia entre la precisión de los distintos conjuntos de datos, muchas de las coordenadas extraídas de los .nc no coincidían exactamente. La solución a este problema, fue realizar una agrupación espacial de las coordenadas mediante la creación de una malla geográfica para poder agrupar las coordenadas y los respectivos valores para cada variable y evitar perder el máximo de valores. Para este estudio, se han probado dos resoluciones espaciales distintas: 0.1° y 0.5° de latitud/longitud (Tabla 2).

Tabla 2: Comparativa entre mallas 0.1° y 0.5°

	0.1°	0.5°
Total de filas	~4.6 millones	262,438
Filas con todas las variables	74,694	19,818
Filas con ≤ 2 valores nulos	1,056,876	105,535
Porcentaje de valores nulos	34.8%	32.7%
% valores nulos	~87%	~41.7%

Observando la tabla, se decidió usar la malla de 0.5º ya que reduce sustancialmente la cantidad de celdas con valores ausentes o nulos, conserva el suficiente detalle espacial para poder hacer un análisis con cobertura global y permite cubrir el mayor número de variables en cada celda sin perder precisión relevante. Además, reducía mucho el esfuerzo de computación y hace que los conjuntos de datos sean más manejables.

Una vez agrupado el conjunto de datos en un documento .csv para cada tipo de datos (estáticos y dinámicos) se cargaron en Rstudio para su análisis exploratorio y filtrado y limpieza del dataset final.

A continuación, se limpiaron los conjuntos de datos de las tres especies (Orca, Ballena y Cachalote) para ajustarlos a la temporalidad de los datos fisicoquímicos (1993-2022) y se unieron a este dataset global utilizando el campo de latitud y longitud y año mediante un left join (unión a la izquierda).

Con el conjunto de datos final, se creó un campo binario de tipo numérico con los datos de avistamientos para poder realizar los algoritmos de aprendizaje supervisado y tras un balanceado de las distintas clases (avistamientos o ausencias) se realizaron los distintos modelos de aprendizaje supervisado.

5. Resultados

5.1.Descripción general del conjunto de datos final

5.1.1. Número total de registros y cobertura por especie

Siguiendo los pasos de la metodología, para el desarrollo de modelos predictivos de distribución potencial de cetáceos, se han integrado 72,299 registros georreferenciados con variables ambientales asociadas para las tres especies objetivo. La distribución por especie después de homogeneizar y limpiar el conjunto de datos es:

- *Physeter macrocephalus* (cachalote): 45.730 observaciones (63,3%).
- *Balaenoptera musculus* (ballena azul): 14.445 observaciones (20%).
- *Orcinus orca* (orca): 12.024 observaciones (16%).

Todos estos datos fueron recopilados gracias a la existencia de fuentes abiertas como Copernicus Marine Service y OBIS. La diferencia en el número de registros entre especies, refleja tanto la distribución como sesgos en el esfuerzo de muestreo y se decide realizar los distintos modelos con las especies por separado para poder sacar tanto conclusiones individuales como conjuntas.

No se han incorporado en los modelos ausencias reales. Para los modelos supervisados, se generaron pseudo-ausencias estratificadas mediante un muestreo aleatorio, seleccionando ubicaciones sin registros en celdas de 0.5° y controlando que no se solaparan espacial ni temporalmente con registros conocidos.

La base final de datos combina observaciones georreferenciadas con valores medidos o derivados de variables fisicoquímicas a escala global, ya agregadas en una malla de 0.5° .

5.1.2. Distribución espacial y temporal

Espacialmente, los registros cubren todo el océano global, desde latitudes australes (-78°) hasta septentrionales (81°), y longitudes desde -180° a 180° , reflejando una cobertura oceánica completa que incluye los principales hábitats y zonas de migración conocidas de las tres especies de estudio.

Temporalmente, los datos se distribuyen desde 1993 hasta 2022, con su mediana en el año 2012 y una densidad mayor en años recientes debido a la importancia y necesidad de conservar sus hábitats.

5.1.3. Balanceo de clases

Para cada especie, se han generado pseudo-ausencias mediante submuestreo aleatorio estratificado sobre celdas sin registros de presencia, igualando el número total de ausencias simuladas al de presencias disponibles. Este procedimiento permite ajustar los modelos con clases balanceadas (1:1) y reducir el sesgo hacia la clase mayoritaria, mejorando la interpretabilidad y robustez de los resultados obtenidos.

5.2. Modelos de aprendizaje automático

Para la creación de los modelos de aprendizaje automático se ha realizado en dos partes tal y como se ha ido explicando a lo largo de la metodología: conjunto de datos sin valores de clorofila en una temporalidad 1993-2022 y el mismo conjunto de datos incluyendo los datos de clorofila (1998-2022) con las variables chl, bbp y poc, lo que permite capturar mejor la dimensión biológica del ecosistema.

Tras un proceso de evaluación comparativa de rendimiento mediante la métrica AUC (Área Bajo la Curva ROC) en los principales algoritmos, se observa una disminución significativa del rendimiento cuando se añadían directamente todas las variables de clorofila. Para resolverlo, se implementa un procedimiento de selección de variables mediante Recursive Feature Elimination (RFE)¹¹, que permite identificar el subconjunto óptimo de predictores, eliminando ruido y redundancia del conjunto de datos.

A partir de este refinamiento, los modelos construidos con las variables seleccionadas (incluyendo clorofila) mostraron mejoras notables en el rendimiento predictivo para todas las especies analizadas. Esta mejora, junto con la relevancia ecológica de la clorofila como marcador de disponibilidad trófica para los cetáceos como *Balaenoptera musculus*, justifica su incorporación como parte fundamental del conjunto de datos final utilizado en los análisis.

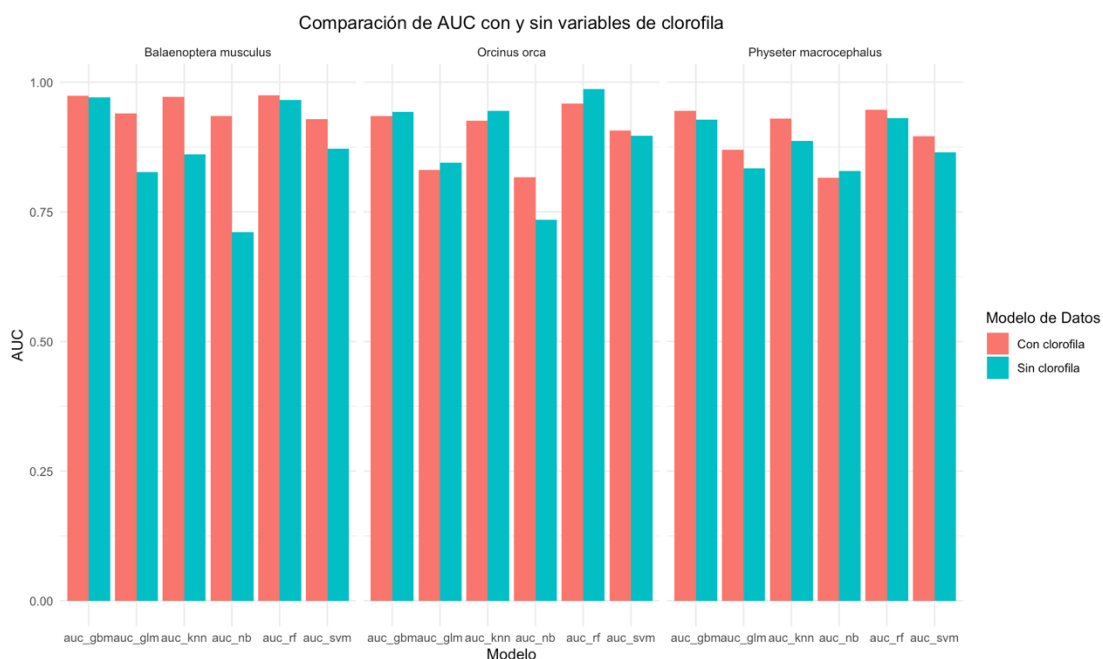


Ilustración 11: Gráfico comparativo de modelos y especies incluyendo clorofila.

¹¹ https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html

5.2.1. Curvas ROC

Para cada una de las tres especies estudiadas: *Balaenoptera musculus*, *Physeter macrocephalus* y *Orcinus orca* se han entrenado cinco algoritmos de clasificación supervisada con el objetivo de predecir la probabilidad de presencia a partir de variables ambientales oceánicas. Los modelos aplicados han sido: regresión logística (GLM), bosques aleatorios (Random Forest, RF), máquinas de vectores soporte (SVM), Naive Bayes (NB), k-vecinos más cercanos (KNN, $k = 5$) y máquinas de soporte de gradiente (GBM) sobre un 80% del conjunto de datos y evaluando sobre el 20% restante.

Las curvas ROC (Receiver Operating Characteristic) permiten evaluar el rendimiento de los modelos de clasificación binaria (presencias:pseudoausencias) al graficar la tasa de verdaderos positivos (sensibilidad) frente a la tasa de falsos positivos ($1 -$ especificidad). Cuanto más cerca esté la curva de la esquina superior izquierda del gráfico, mejor es el desempeño del modelo.

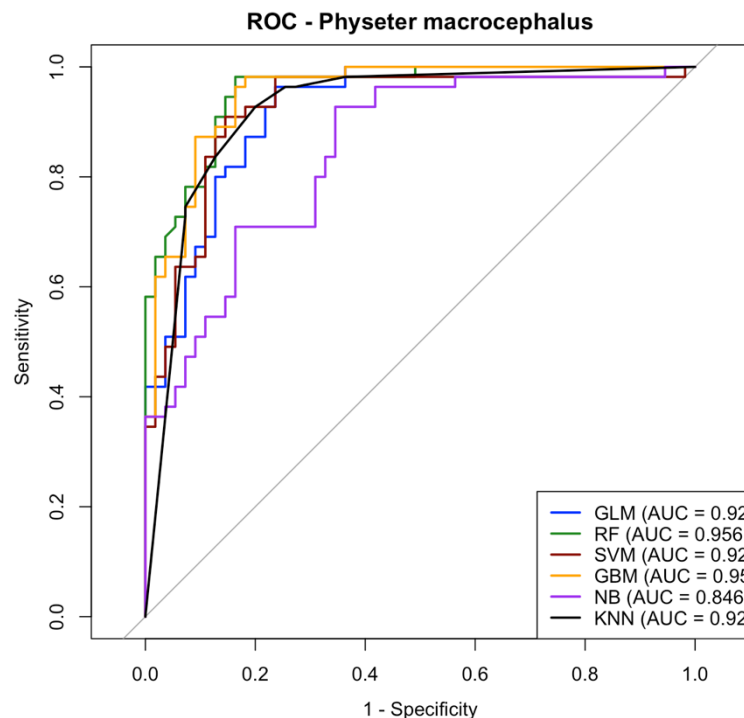


Ilustración 12: Curvas ROC para *Physeter macrocephalus*.

Observando la Ilustración 12 podemos ver cómo la mayoría de los algoritmos utilizados tienen una buena clasificación para *Physeter macrocephalus* siendo RF y GBM los que mejor clasifican para esta especie.

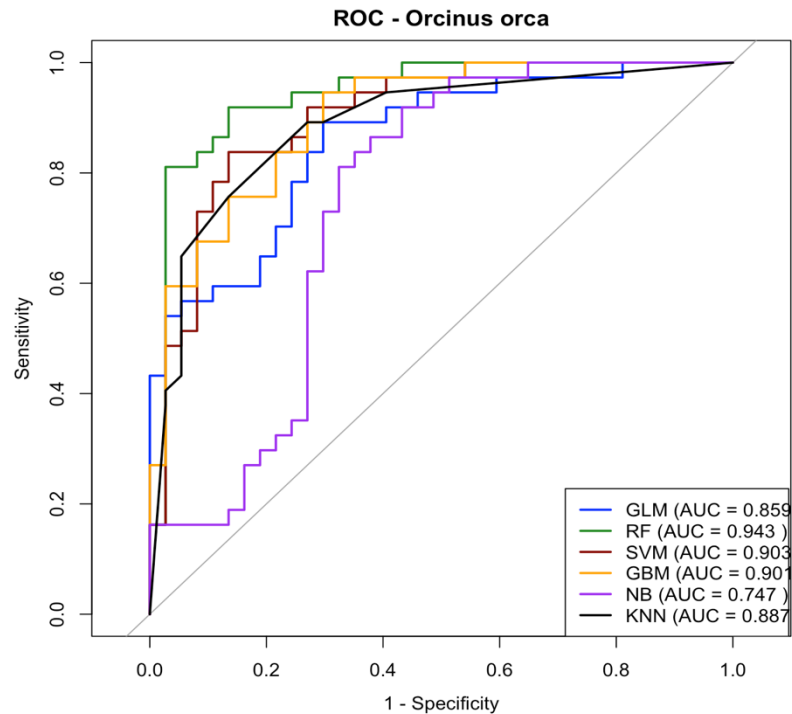


Ilustración 13: Curvas ROC para *Orcinus orca*.

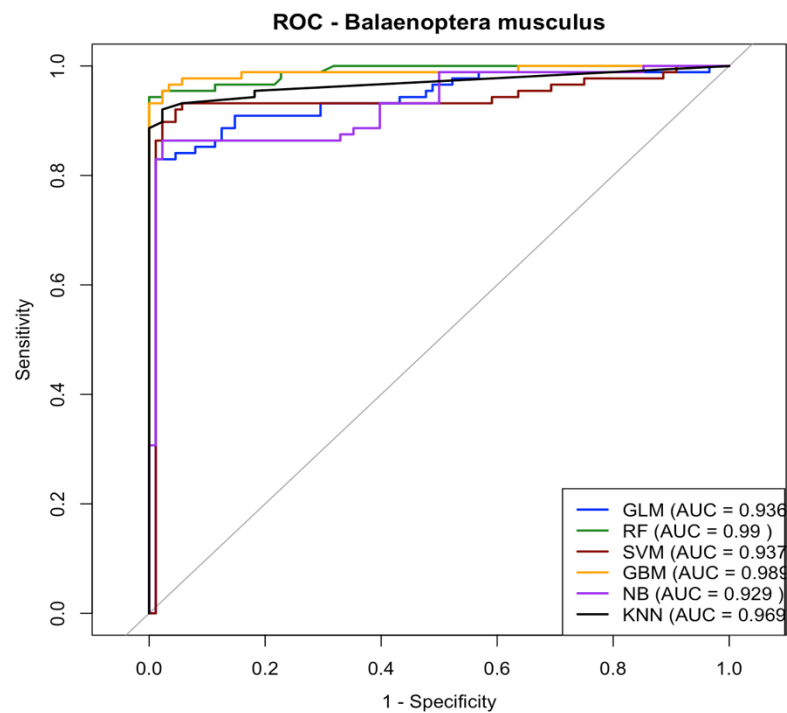


Ilustración 14: Curvas ROC para *Balaenoptera musculus*.

En estas dos ilustraciones 13 y 14, se observan las distintas curvas ROC para cada modelo y especie donde se aprecia una buena clasificación de los distintos algoritmos para *Balaenoptera musculus* siendo RF y GBM los que mejor clasifican. Por otro lado, observamos como *Orcinus orca* tiene una peor capacidad de predicción con algoritmos que no clasifican nada bien (NB, KNN) en comparación con RF o GBM.

5.2.2. Valores AUC

Cada modelo se entrenó sobre un 80% de los datos y se evaluó sobre el 20%. La métrica de desempeño utilizada ha sido el Área Bajo la Curva ROC (AUC), que oscila entre 0.5 (rendimiento aleatorio) y 1 (clasificación perfecta). Los valores de AUC para cada modelo y especie se resumen en la siguiente figura (Tabla 3).

Tabla 3: Valores AUC para cada especie y algoritmo.

Especie	GLM	RF	SVM	GBM	Naive Bayes	KNN
<i>Physeter macrocephalus</i>	0.922	0.956	0.923	0.952	0.845	0.921
<i>Orcinus orca</i>	0.859	0.943	0.903	0.901	0.747	0.887
<i>Balaenoptera musculus</i>	0.936	0.990	0.937	0.989	0.929	0.969

Observando la tabla de resultados, se confirma que los modelos de Random Forest (RF) y Gradient Boosting Machine (GBM) presentan el mejor rendimiento para las tres especies analizadas. Ambos algoritmos alcanzan valores de AUC superiores a 0.90, lo que indica una alta capacidad discriminativa entre presencias y pseudo-ausencias.

Para *Physeter macrocephalus*, el modelo RF logra el mayor AUC (0.956), seguido muy de cerca por GBM (0.952) y SVM (0.923). Estos valores reflejan una fuerte relación entre las variables ambientales seleccionadas y la distribución de esta especie, con algoritmos basados en árboles mostrando un comportamiento robusto.

Para el caso de *Orcinus orca*, RF también destaca con un AUC de 0.943, seguido por SVM (0.903) y GBM (0.901). Aunque los valores siguen siendo elevados, se aprecia una ligera

disminución en comparación con *Physeter*, posiblemente atribuida a una mayor complejidad o variabilidad en los patrones de distribución de esta especie.

Balaenoptera musculus tiene los mejores resultados absolutos: el algoritmo de RF alcanza un AUC de 0.990, seguido de GBM con un valor de 0.989 y SVM en 0.937. Este alto rendimiento sugiere que las variables seleccionadas, incluyendo la clorofila, son altamente informativos para predecir la distribución de esta especie, probablemente debido a su fuerte dependencia de zonas de alta productividad.

Por otro lado, GLM ofrece una capacidad predictiva razonable en *Physeter* (0.922) y *Balaenoptera* (0.936), pero es inferior en *Orcinus* (0.859), lo que podría indicar que la relación entre variables y presencia no es completamente lineal en este último caso.

El modelo Naive Bayes es el modelo de menor rendimiento, especialmente en *Orcinus orca* (0.747), lo que sugiere que sus supuestos simplificados sobre independencia de variables no son adecuados para datos ecológicos complejos como estos.

Finalmente, KNN muestra un desempeño intermedio, con valores AUC que oscilan entre 0.887 y 0.969, siendo competitivo pero inferior a los modelos RF y GBM.

Además, también podemos realizar un histograma para visualizar los distintos valores AUC de todos los algoritmos realizados para cada especie (Ilustración 15).



Ilustración 15: Comparación de AUC por modelo y especie.

En conjunto, los resultados refuerzan la elección de Random Forest y GBM como los algoritmos preferentes para modelar la distribución de cetáceos en función de variables oceanográficas seleccionadas, tras un proceso cuidadoso de limpiado y procesado de los datos seguido de KNN o SVM.

5.2.3. Importancia de las variables

Una vez ajustados los modelos de clasificación supervisada y determinada la eficacia de los modelos Random Forest y GBM por encima del resto, se ha calculado la importancia relativa de cada variable en la predicción de presencia de cada especie estudiada. Para hacerlo, primero se utilizaron técnicas de reducción de dimensionalidad como el Análisis de Componentes Principales (PCA), que transforman las variables originales en combinaciones lineales abstractas pero al no representar bien el conjunto de variables, se han extraído directamente de los modelos de aprendizaje automático que permiten interpretar directamente la influencia individual de cada variable original sobre el resultado, lo que es crucial desde una perspectiva ecológica y de gestión y planificación espacial. De esta manera,

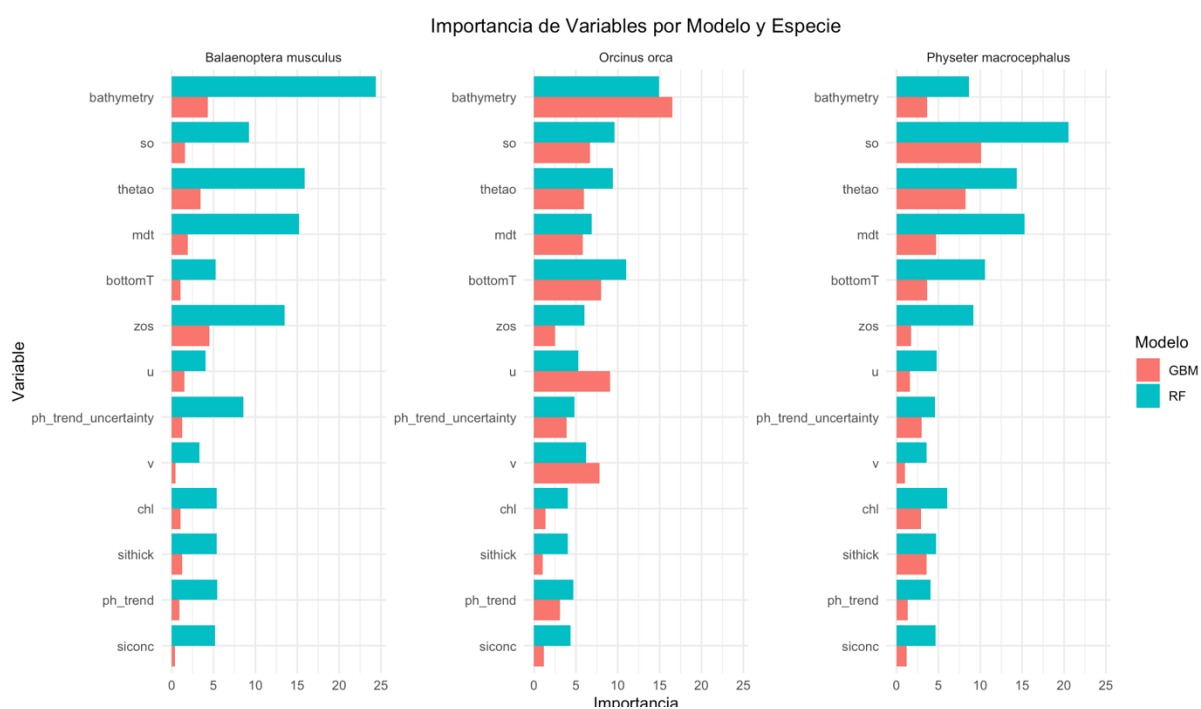


Ilustración 16: Importancia de variables por especie.

a partir de las importancias calculadas para RF y GBM se puede extraer el siguiente gráfico comparativo (Ilustración 16).

Observando la imagen, podemos decir que la bathymetry (batimetría) es la variable más importante para las tres especies en ambos modelos. Esta consistencia subraya el papel fundamental que juega la topografía submarina en la distribución de los grandes cetáceos, al influir en la productividad y accesibilidad de presas en las tres especies. Ha sido señalada como un factor predictivo principal para múltiples especies, incluyendo odontocetos y mysticetos, por su relación con cañones submarinos, taludes y zonas de afloramiento [14].

Además, observamos variables relacionadas con la estructura del océano como so (salinidad), thetao (temperatura potencial del agua), y mdt (topografía dinámica media) que también representan una alta importancia en los tres casos, reflejando una dependencia común de estos animales por hábitats específicos relacionados con el movimiento de las masas de agua y las corrientes oceánicas. Estas variables describen masas de agua, estabilidad de columna de agua y procesos como corrientes y remolinos que afectan a la distribución trófica [5].

Fijándonos en la ballena azul (*Balaenoptera musculus*), se observan muy accentuadas las variables bathy, so, y mdt, pero también destaca ph_trend_uncertainty, posiblemente reflejando una sensibilidad más aguda a tendencias de acidificación oceánica relacionado con su hábitat alimenticio. Además, es la única especie donde zos (nivel del mar) tiene un peso comparable al resto, lo que podría relacionarse con su preferencia por áreas con surgencias (upwhelings) o regiones productivas que han sido vinculadas a las áreas de alimentación de *B. musculus*, especialmente en regiones como el Pacífico sudeste y el norte del Índico [20].

De forma similar, la orca (*Orcinus orca*), comparte relevancia de bathy, so, y mdt, pero se observa una mayor dispersión en la importancia relativa de las variables, lo que podría deberse a la mayor plasticidad ecológica de la orca como depredador generalista que explota múltiples hábitats. La variable u (componente zonal de corriente) gana relevancia para esta especie, posiblemente por el uso de corredores oceánicos y su adaptación a distintos patrones hidrodinámicos (incluyendo componentes zonales del flujo) puede reflejar una estrategia de caza dependiente del transporte horizontal de presas [21].

Por último, el cachalote (*Physeter macrocephalus*) como era de esperar, muestra un patrón similar al de *Balaenoptera*, pero con mayor peso relativo en variables como thetao y zos. Esto

puede estar asociado a su estrategia de buceo profundo y selección de hábitats más estables para la alimentación y reproducción. Para esta especie, chl (clorofila) tiene cierta importancia, aunque menor, lo cual puede deberse a su relación indirecta de productividad primaria como señal de presencia de presas (calamares) en las capas profundas del océano [5].

5.3. Mapas de distribución potencial

Tras la evaluación de los modelos de clasificación y la identificación de las variables ambientales más relevantes, se procede a la elaboración de mapas de distribución potencial para las tres especies de cetáceos analizadas: *B. musculus*, *P. macrocephalus* y *O. orca*.

Para ello, se utilizan los modelos de aprendizaje automático entrenados previamente, seleccionando aquellos con mejor rendimiento en función del AUC (Random Forest y Gradient Boosting Machines (GBM)). Estos modelos han sido aplicados sobre el conjunto de datos ambientales espaciales y temporales, generando una predicción de la probabilidad de presencia en cada punto de la malla 0.5° x 0.5°.

Los mapas obtenidos, permitirán observar los patrones espaciales de probabilidad de ocurrencia, visualizando las áreas con mayor probabilidad de encontrar cada especie. Este enfoque resulta especialmente útil para tareas de conservación marina, gestión espacial y análisis de cambio climático, al facilitar una representación continua y visual de la distribución más probable de estas especies.

A continuación, se mostrarán los tres mapas de distribución creados a partir de RF para las tres especies de estudio (Ilustración 17, 18 y 19).

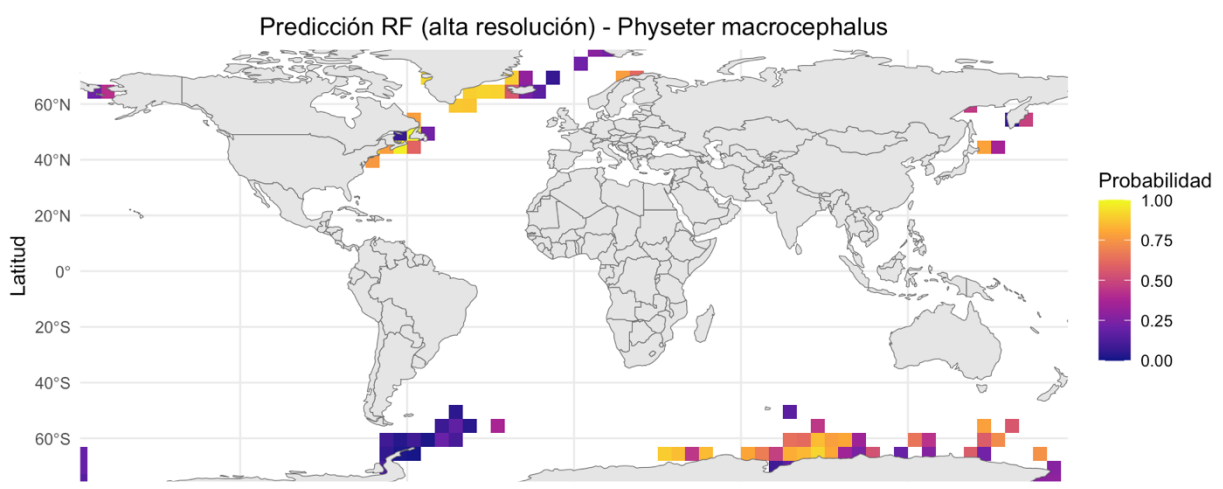


Ilustración 17: Mapa de distribución *Physeter macrocephalus*.

En este mapa de distribución a escala global a partir del modelo predictivo RF se puede observar la probabilidad de encontrar cachalotes (*Physeter macrocephalus*) en un valor que oscila entre una escala del 0-1 siendo los sitios de más probabilidad la costa este de EEUU y Canada, así como cerca del océano austral en la longitud 60°E.

A continuación, se representa la distribución espacial para la especie *Orcinus orca* (Ilustración 18):

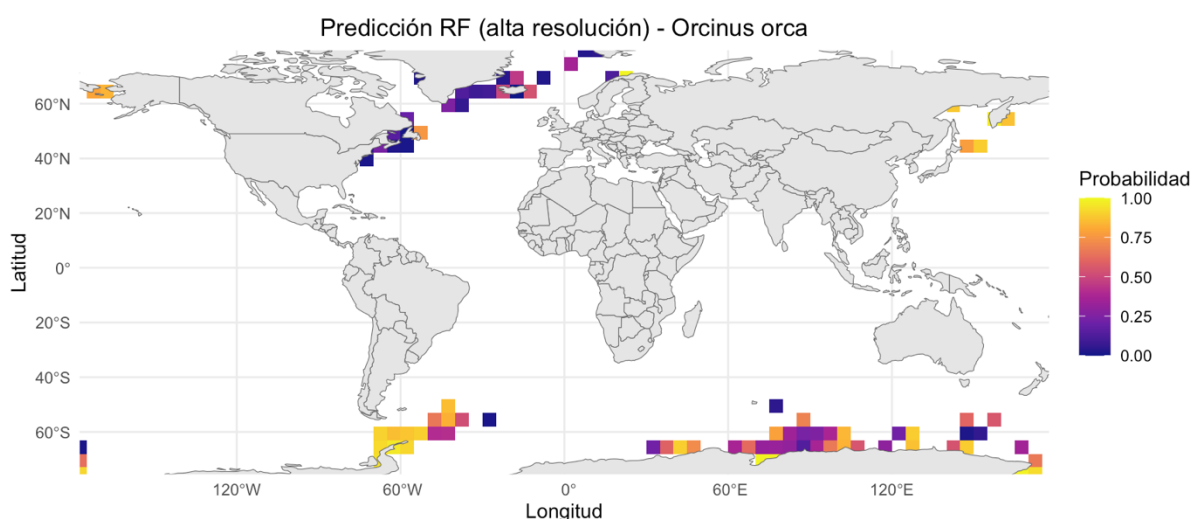


Ilustración 18: Mapa de distribución para Orcinus orca.

Observando el siguiente mapa de distribución y de manera opuesta al cachalote se observa la distribución de *Orcinus orca*, mucho más localizada en la península Antártida y cerca del estrecho del Drake, así como en el océano austral y parte del pacífico Norte con avistamientos mucho menos frecuentes en el Atlántico norte.

Por último, se observa el mapa de distribución de *Balaenoptera musculus* (Ilustración 19):

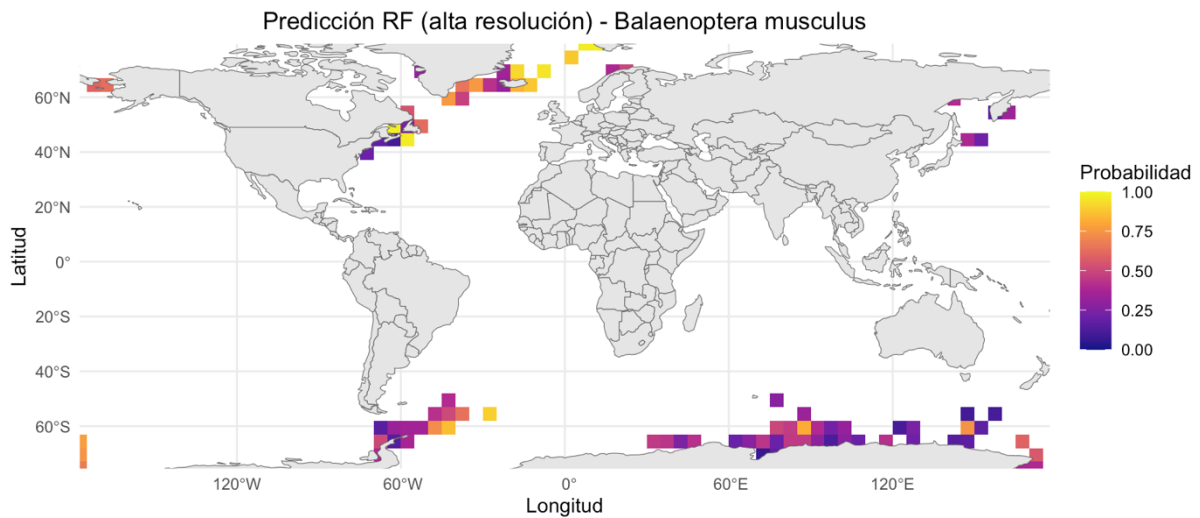


Ilustración 19: Mapa de distribución Balaenoptera musculus.

En este último mapa, podemos observar la distribución espacial de la ballena azul que ha resultado ser una mezcla entre las dos distribuciones anteriores donde se encuentra una probabilidad alta de avistamiento de esta especie en el atlántico sur, pero más alejada de la península Antártida con avistamientos poco probables en el océano austral y con más probabilidad de avistamiento en el norte-este de Canadá para Latitudes superiores a los 60°N.

5.4. Modelos de distribución MaxENT

Como complemento a los modelos supervisados clásicos, se utilizó el algoritmo MaxEnt¹² (Maximum Entropy Modeling) para predecir la distribución potencial de cada una de las tres especies de cetáceos estudiadas: *Physeter macrocephalus*, *Balaenoptera musculus* y *Orcinus orca*. MaxEnt es un método de modelado de nicho ecológico ampliamente utilizado en ecología de la conservación por su capacidad para trabajar exclusivamente con datos de presencia y generar predicciones robustas basadas en la teoría de máxima entropía que

¹² https://biodiversityinformatics.amnh.org/open_source/maxent/

explica que la distribución más probable de un sistema en la naturaleza es la que tiene más entropía.

Para cada especie, se extrajeron los registros únicos de presencia georreferenciada a partir de la base de datos combinada de OBIS, filtrando aquellos con coordenadas válidas y variables ambientales completas, es decir, sin valores nulos. Posteriormente, se emplearon las mismas variables predictoras físico-químicas y de productividad marina integradas espacialmente en una malla global de 0.5° de resolución. Entre las variables utilizadas se encuentran temperatura potencial a distintas profundidades (thetao, bottomT), salinidad (so), contenido de hielo marino (siconc, sithick), altura dinámica media (zos), profundidad de la capa de mezcla (mlofst), y variables de clorofila y materia particulada (chl, poc, bbp).

Los modelos se implementaron en R mediante el paquete dismo, con la versión 3.4.3 de MaxEnt, integrando una validación cruzada de 5 replicaciones (replicates = 5, replicatetype = "crossvalidate").

Una vez realizados los modelos, su calidad fue evaluada mediante el área bajo la curva ROC (AUC), obteniendo valores medios de 0.899 para *Physeter macrocephalus* (Ilustración 20), 0.755 para *Balaenoptera musculus* (Ilustración 21) y 0.815 para *Orcinus orca*, lo que refleja un rendimiento razonable en el cachalote y la orca y bastante menor capacidad predictiva en la ballena azul *B. musculus*. Se debe tener en cuenta que MaxEnt modela la idoneidad ambiental y no la abundancia ni el uso del hábitat, y que la interpretación puede verse afectada por sesgos en el muestreo de presencia. Además de los valores AUC, los resultados de MaxENT

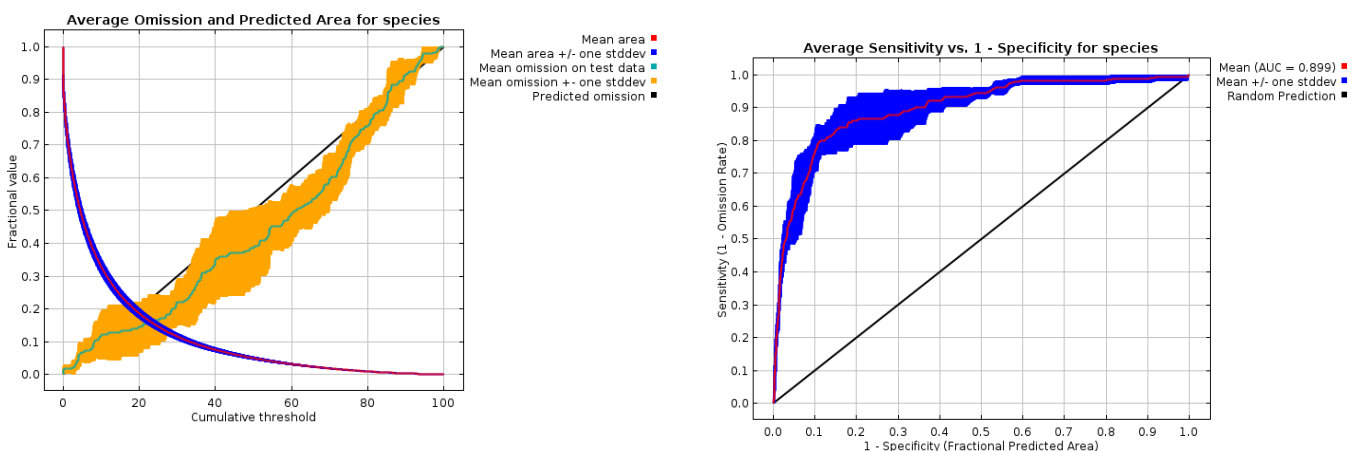


Ilustración 20: Resultados de MaxENT para *Physeter macrocephalus*.

vienen acompañados de un gráfico sobre el área predicha para cada una de las especies y su curva AUC.

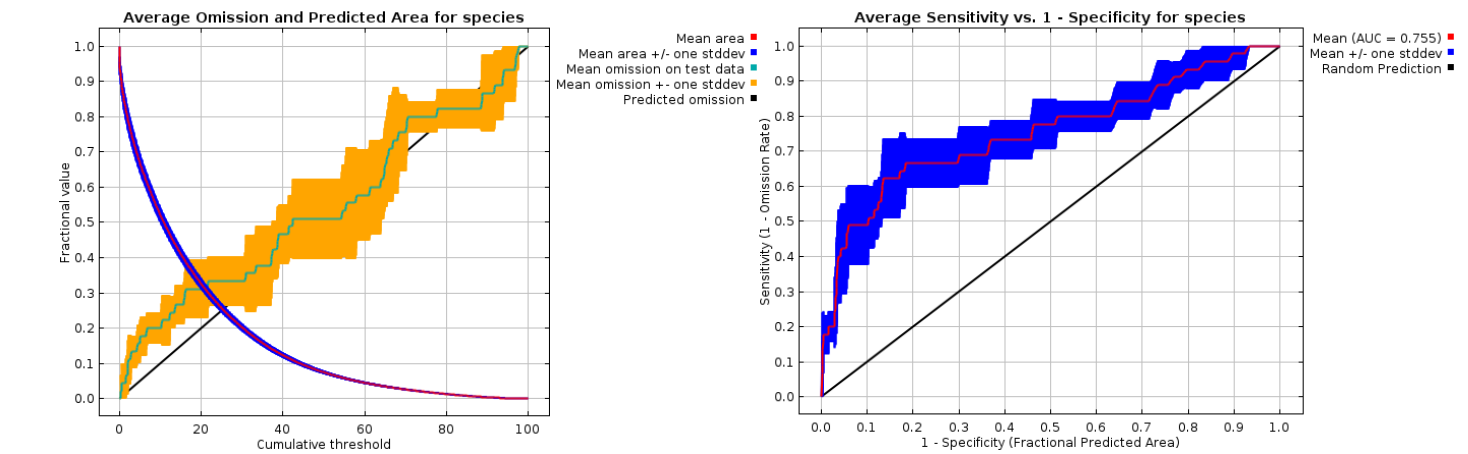


Ilustración 21: Resultados de MaxENT para *Balaenoptera musculus*

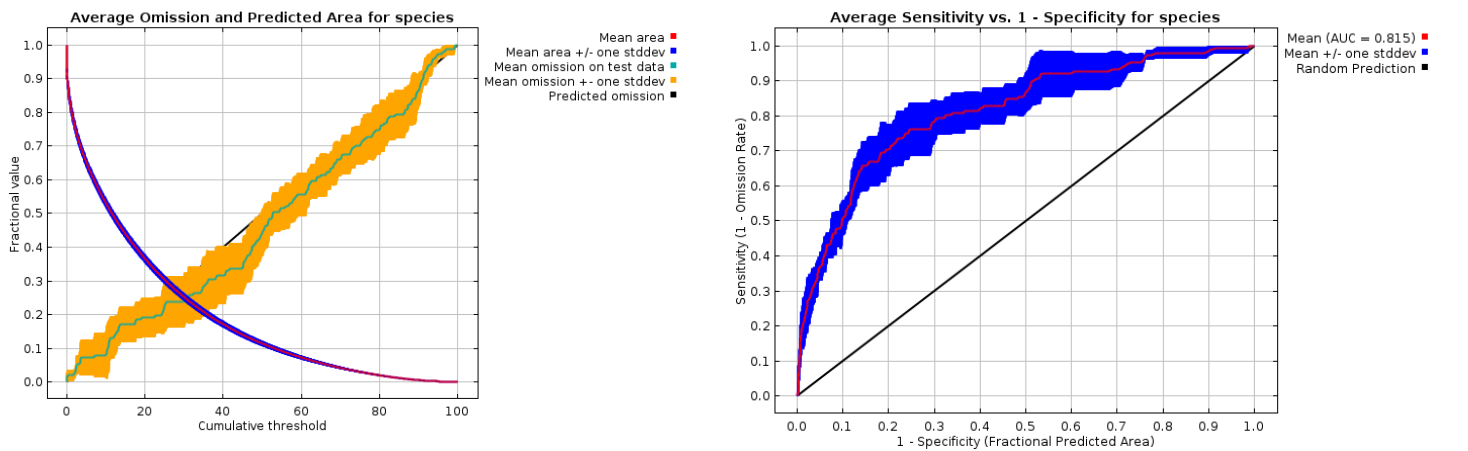


Ilustración 22: Resultados de MaxENT para *Orcinus orca*.

Adicionalmente a estos resultados, la aplicación de modelos MaxENT también analiza la importancia o contribución de cada una de las variables predictoras sobre cada uno de los modelos que se distribuirán en la siguiente tabla para poder compararlas con los modelos realizados mediante algoritmos de aprendizaje supervisado (Tabla 4):

Tabla 4: Resumen de la importancia de variables MaxENT.

Variable (%)	Physeter macrocephalus	Orcinus orca	Balaenoptera musculus
mdt	42.1	18.5	9.6
zos	14.4	6.5	3.8
ph_trend	13.6	7.4	0.9
chl	7.6	17.8	47
siconc	6.8	4.8	2.9
bottomT	4.0	6.1	2.1
thetao	3.5	7.5	19.9
sithick	3.5	3.5	3.0
so	2.3	6.2	2.1
v	0.8	2.6	2.4
m1otst	0.7	1.3	1.6
u	0.6	3.4	0.6
poc	0.1	3.4	3.8
bbp	0.0	0.1	0.4

Tal y como podemos observar en la siguiente tabla resumen, podemos ver como para el cachalote parece que la variable predictora que más contribuye al modelo MaxEnt es el modelo digital del terreno (42.1% mdt) y la altura dinámica (zos) mientras que para la orca

parece ser entre la chlorofila y el mdt (17.8% chl y 18.5% mdt) y para la ballena azul parece ser claramente la chlorofila (47% chl) seguida por la temperatura (19.9% thet_{ao}).

Para finalizar y complementar la información de los modelos obtenidos, se generaron archivos raster GeoTIFF con las predicciones para cada especie, reorientados y ajustados espacialmente para facilitar su análisis e interpretación en sistemas de información geográfica (SIG) como QGIS. Asimismo, se analizaron las contribuciones relativas de las variables a cada modelo, destacando la importancia consistente de la altura dinámica del mar (zos) y la temperatura del fondo (bottomT) en todas las especies, coherente con su comportamiento migratorio y preferencias batimétricas. Para cada especie se seleccionó el primer modelo replicado (modelo_rep1) y se aplicó sobre las capas raster de predictores ambientales para generar mapas de probabilidad de presencia.

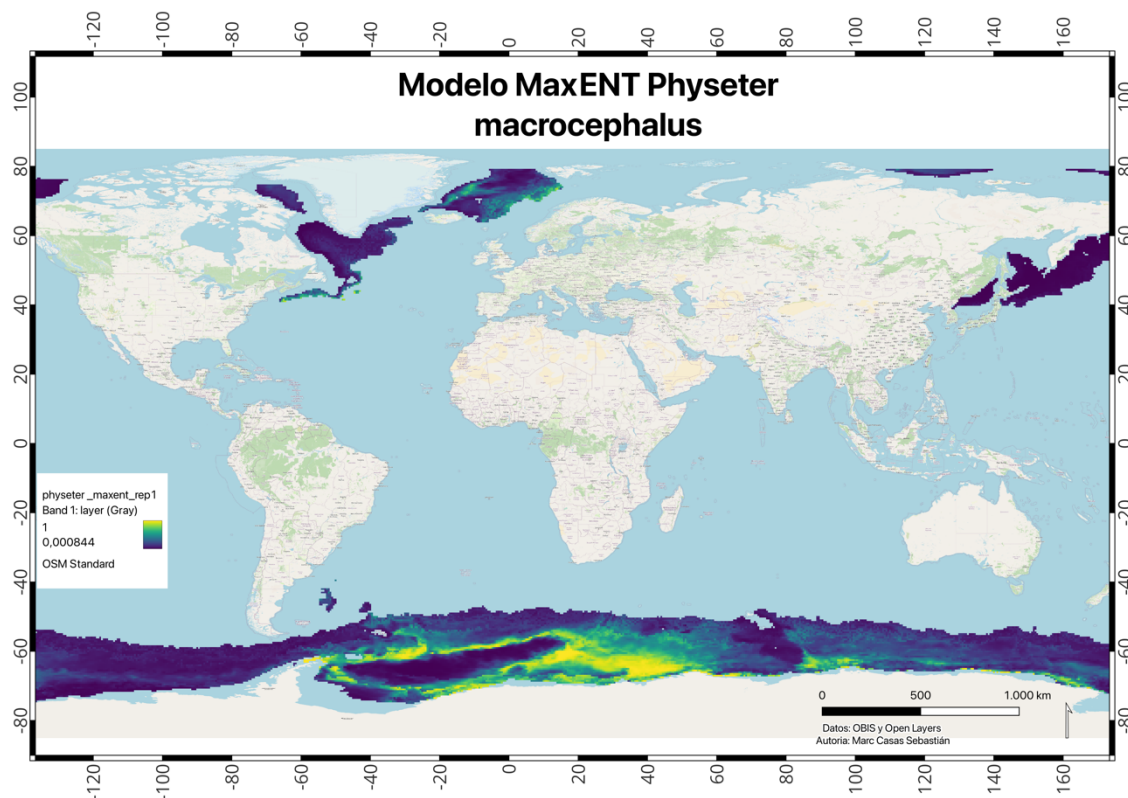


Ilustración 23: Mapa de distribución MaxENT de Physeter macrocephalus.

En el mapa de distribución de *Physeter macrocephalus* resultado del modelo MaxENT (Ilustración 23), se puede determinar una mayor área de distribución predicha por el modelo en las áreas cercanas a la Antártida, concretamente en latitudes entre -60° y 60° S. También

se aprecian puntos de distribución cerca de Islandia entre latitudes entre -20° y 20° N. Además, vemos también una pequeña área de distribución cerca de la costa Este de Canada y la parte superior de la costa Este de los EEUU.

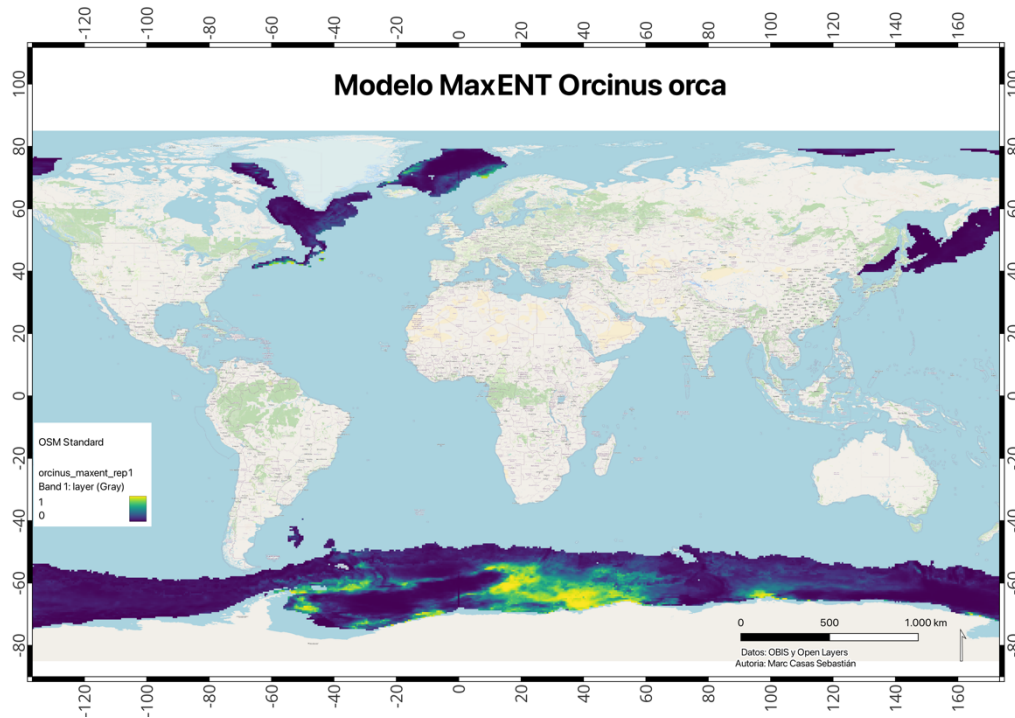


Ilustración 24: Mapa de distribución MaxENT de Orcinus orca.

Observando el siguiente mapa de distribución (Ilustración 24), podemos ver como el modelo ha predicho una distribución bastante similar para la orca dónde se observa una menor acentuación de las presencias tanto en las regiones Sur y Norte, aunque el patrón de presencias coincide bastante exactamente con el cachalote, aunque con una extensión ligeramente menor para la región Sur (cerca de la Antártida) y del océano austral.

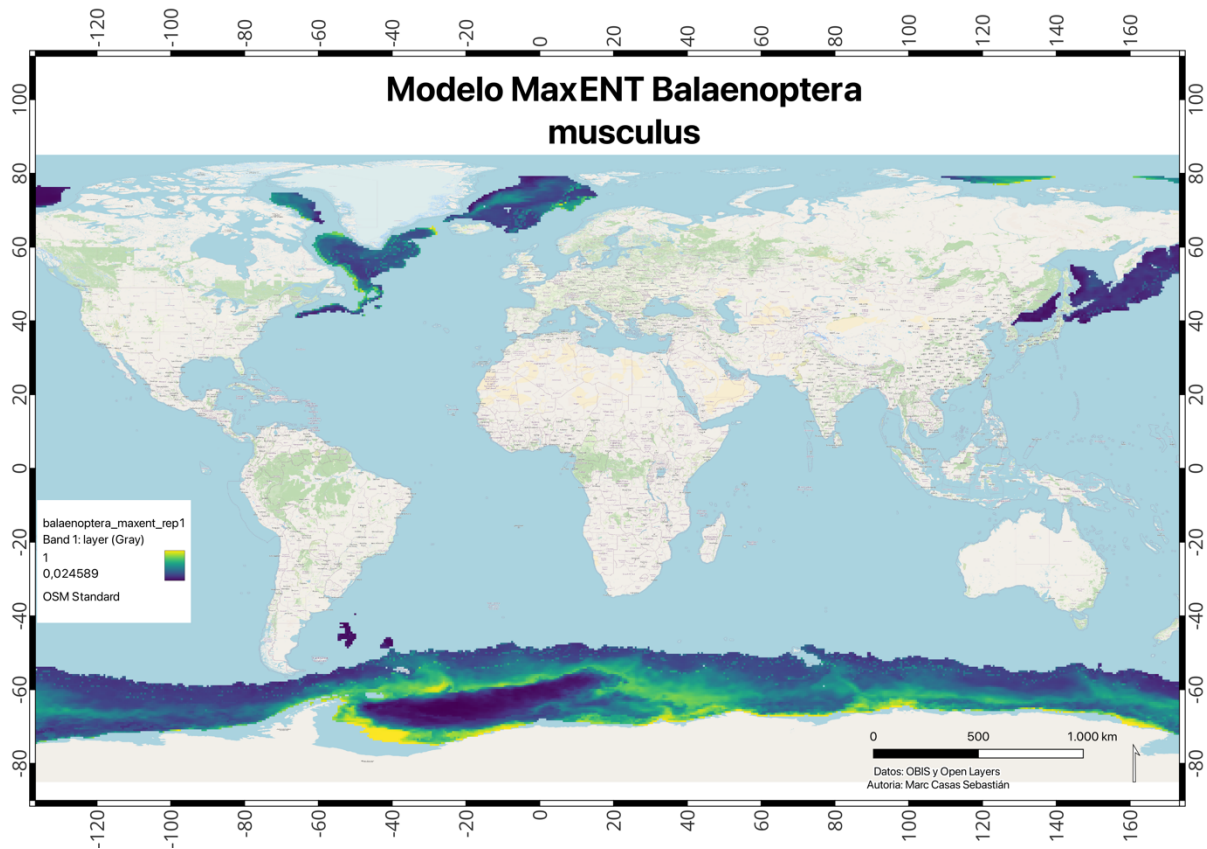


Ilustración 25: Mapa de distribución MaxENT de Balaenoptera musculus.

Por último, observamos el mapa de distribución realizado con MaxENT para la Ballena azul (Ilustración 25) dónde sí que se observan diferencias sustanciales con los otros dos mapas de distribución. Además de ser un mapa mucho más suavizado y con menos ausencias visibles, se observan presencias más probables tanto en todo el océano austral, con mayor concentración cerca de la península Antártica -60º -20ºS, también se observan presencias entre Canadá y Groenlandia y en la misma zona Norte que las otras dos entre Islandia y Noruega con un destacable valor de probabilidad de presencia respecto a las otras dos. Por último, un área de distribución que se diferencia de los otros modelos es en las latitudes Norte 110º - 180ºN cerca de Rusia y la Siberia.

6. Discusión

El modelado predictivo de la distribución de cetáceos migratorios representa un reto metodológico debido a la cantidad y variedad de datos y variables a tener en cuenta. A lo largo de este estudio se han integrado datos de múltiples fuentes, escalas y naturalezas distintas, así que la construcción de modelos robustos ha exigido no solo una rigurosa limpieza, preprocesado y tratamiento de los datos, sino una buena reflexión sobre los factores ecológicos y oceanográficos que influyen en la distribución de estas especies.

Desde su inicio, este trabajo se ha planteado como una combinación entre el rigor ecológico y el analítico de la bioinformática, lo que ha implicado una serie de decisiones técnicas cuyo impacto se refleja directamente en los resultados obtenidos y en su interpretación.

Uno de los aspectos metodológicos más exigentes ha sido, sin duda, la necesidad de combinar fuentes de datos de naturaleza, resolución y origen distintos. Además, la base de datos de presencias obtenida de OBIS proporciona una cobertura espacial significativa, pero carece de información sobre esfuerzo de muestreo, lo que complica la interpretación de las ausencias y limita la capacidad de generalización de los modelos. A este efecto, se suma las variables ambientales procedentes de Copernicus Marine Service, cuya resolución anual y profundidad vertical varían entre distintas variables, y los datos de clorofila-a, solo están disponibles como valores medios mensuales para el periodo 1998–2022. Este desajuste temporal con las series de variables físico-químicas, disponibles desde 1993, obligó a implementar flujos de trabajo paralelos y a tomar una serie de decisiones, como el uso de medianas anuales y la agregación espacial mediante mallas de 0.5°, en lugar de resoluciones más finas que habrían podido aumentar la resolución espacial del dataset. Estas decisiones, lejos de ser triviales, responden a recomendaciones ampliamente aceptadas en estudios de modelado de distribución de especies marinas, donde la calidad de los datos y la coherencia espacio-temporal son condiciones imprescindibles para la robustez del modelo [22].

Dentro de este marco de limitaciones, la inclusión de la clorofila-a como variable predictora merece una consideración especial. A pesar de su carácter mensual y su menor rango temporal, su valor como proxy (indicador) trófico para evaluar la productividad primaria, justifica su incorporación en modelos ecológicos centrados en especies herbívoras como la ballena azul. Su influencia se reflejó especialmente en los modelos de *Balaenoptera musculus*,

donde la presencia de clorofila-a se asocia a mayores probabilidades de ocurrencia. No obstante, su inclusión redujo ligeramente la capacidad predictiva en otros modelos, lo que evidencia la búsqueda del equilibrio entre precisión estadística y relevancia ecológica. En este sentido, resulta clave entender que un modelo estadísticamente eficiente no siempre es un modelo ecológicamente explicativo, y que las variables ecológicas clave pueden tener efectos complejos o no lineales que no se reflejan fácilmente en métricas como el AUC [13].

La elección de los distintos algoritmos aplicados en este trabajo, se han fundamentado en la necesidad de capturar relaciones no lineales, tolerar la multicolinealidad y adaptarse a datos con distribución no balanceada y sin esfuerzos. Los algoritmos basados en árboles, como Random Forest y XGBoost, demostraron un rendimiento superior en términos de estabilidad y poder de discriminación, especialmente en los modelos de *Physeter macrocephalus* y *Balaenoptera musculus*, donde los valores de AUC superaron consistentemente el valor de 0.90. La capacidad de estos modelos para manejar grandes volúmenes de variables y para estimar la importancia relativa de cada predictor los convierte en herramientas idóneas para problemas ecológicos complejos, tal y como se ha documentado extensamente en la literatura [23], [17]. Frente a ellos, los modelos más tradicionales como GLMs y SVMs ofrecieron resultados más modestos, mostrando una mayor sensibilidad a la selección de variables y distribución de las distintas clases. En particular, SVM presentó un rendimiento inestable con *Orcinus orca*, posiblemente debido a la menor consistencia espacial de sus presencias y su distribución como especie altamente móvil y con una amplia extensión.

Una cuestión metodológica central en este trabajo ha sido la generación de pseudoausencias, necesaria en ausencia de datos de esfuerzo y ausencias reales para los modelos. Para ello, se optó por una estrategia basada en la selección de celdas con registros comunes para las tres especies, lo que permite aproximar condiciones desfavorables para la ocurrencia. Sin embargo, este método conlleva riesgos importantes, como la inclusión de falsas ausencias o la sobreestimación del rendimiento del modelo en función de una distribución artificial de las clases [24]. Alternativas más robustas, como el uso de mapas de esfuerzo (con las líneas de la navegación) o la generación estratificada de pseudoausencias ponderadas por accesibilidad, podrían mejorar la inferencia en futuros estudios, especialmente si se desea extrapolar a áreas con menor cobertura de datos o evaluar la idoneidad del hábitat en escenarios proyectados [25].

Desde el punto de vista ecológico, el análisis de los modelos resultantes ha evidenciado diferencias significativas en la respuesta de las especies a los gradientes ambientales. *Physeter macrocephalus* mostró una fuerte asociación con variables como batimetría, temperatura de fondo y salinidad, lo cual concuerda con su comportamiento de inmersión profunda y preferencia por zonas meso y batipelágicas [5]. *Balaenoptera musculus*, en cambio, se correlacionó principalmente con variables epipelágicas como la SST y la clorofila, confirmando su dependencia de áreas de alta productividad superficial y su comportamiento migratorio estacional [6]. En el caso de *Orcinus orca*, los modelos mostraron un rendimiento más variable, posiblemente reflejo de su plasticidad trófica, sus diferencias intraespecíficas (ecotipos¹³) y una mayor dependencia de factores no ambientales como la estructura social o la distribución de presas específicas [21]. Esta comparación multi-especie aporta un valor añadido al trabajo, ya que permite identificar patrones generales y específicos en la relación entre especies altamente móviles y su entorno físico, lo que refuerza la validez ecológica de los resultados obtenidos.

Si bien los mapas generados permiten visualizar zonas de alta probabilidad de presencia, como los márgenes continentales, zonas de transición térmica y regiones de surgencia, es imprescindible interpretarlos con cautela. La falta de datos de esfuerzo limita la extrapolación a regiones no muestreadas, y los modelos solo pueden predecir dentro del rango ambiental representado en los datos de entrenamiento. Este fenómeno, conocido como “truncamiento del nicho”, está ampliamente documentado en estudios de MaxEnt (Máxima entropía) y otros enfoques similares [11], y debe ser considerado al aplicar los resultados a escalas globales o a planes de conservación espacialmente explícitos. Por estos motivos, se ha decidido utilizar a modo complementario, modelos de máxima entropía MaxEnt ya que son muy útiles en contextos de datos de presencia solamente [11], especialmente cuando se busca interpretar patrones más que realizar predicciones absolutas.

¹³ Subespecie de una especie adaptada genéticamente a un hábitat pero que puede cruzarse con otros miembros de la especie no pertenecientes a esa subespecie - <https://dicciomed.usal.es/palabra/ecotipo>

MaxEnt no predice presencias reales sino idoneidad ambiental, y su fortaleza radica en identificar patrones de uso potencial del hábitat sin necesidad de pseudoausencias. En este estudio, la aplicación de este modelo, ha ofrecido un rendimiento razonable para *Physeter macrocephalus* (AUC = 0.899), con un valor intermedio para *Orcinus orca* (AUC = 0.815) y más limitado para *Balaenoptera musculus* (AUC = 0.755).

Comparado con los algoritmos supervisados como RF o GBM (con AUCs superiores a 0.95 en algunos casos), MaxEnt demuestra una menor capacidad predictiva general. Sin embargo, desde un punto de vista ecológico, el análisis de importancia de variables ofrece una perspectiva complementaria muy valiosa. Por ejemplo, en modelos de RF la batimetría y la salinidad dominaban para el cachalote, mientras que MaxEnt reveló una alta dependencia de la altura dinámica media (mdt) y la salinidad superficial (zos). Asimismo, la variable clorofila (chl) fue la más determinante en el modelo MaxEnt de *Balaenoptera musculus*, destacando aún más su relevancia como proxy de productividad [16].

Esta diferencia entre modelos pone en evidencia una característica central en la modelización y que aún se ve más afectada cuando hablamos de distribución de especies y hábitats: no hay un único “mejor” modelo, sino aproximaciones complementarias que capturan distintos aspectos del nicho ecológico. Mientras los modelos de machine learning supervisados destacan en capacidad predictiva, MaxEnt ofrece interpretaciones robustas del nicho potencial, particularmente útil cuando no se dispone de ausencias reales. Además, el fenómeno del truncamiento del nicho, ampliamente discutido en la literatura [26], limita la extrapolación de cualquier modelo a condiciones no representadas en los datos de entrenamiento, especialmente en áreas poco muestreadas. Este sesgo puede resultar en mapas de predicción poco fiables fuera de las regiones ambientales conocidas.

En cuanto a los patrones espaciales, los modelos MaxEnt han mostrado una resolución espacial sustancialmente superior a la obtenida mediante los algoritmos supervisados como Random Forest o GBM. Esto se traduce en mapas más detallados y con contornos de idoneidad ambiental mejor definidos, lo que permite identificar con mayor precisión áreas de transición ecológica, márgenes continentales y regiones de surgencia. A pesar de que todos los modelos coinciden en resaltar zonas de alta probabilidad de presencia, como plataformas continentales profundas y bordes dinámicos oceánicos, las predicciones derivadas de MaxEnt

tienden a ser más refinadas y localizadas, mientras que las de RF y GBM presentan patrones más difusos o sobregeneralizados, especialmente en áreas con menor densidad de datos.

Este fenómeno puede explicarse, en parte, por la tendencia de los algoritmos de aprendizaje supervisado a sobreajustarse al conjunto de entrenamiento, especialmente cuando se usan pseudoausencias generadas sin mapas de esfuerzo ni ponderación por accesibilidad. En contraste, MaxEnt, al estar diseñado específicamente para trabajar con datos de presencia solamente y operar bajo el principio de máxima entropía, tiende a generar predicciones más conservadoras, limitadas al espacio ambiental verdaderamente muestreado, pero menos propensas a la extrapolación [11]. En este estudio, las predicciones de MaxEnt se ajustan de forma más realista al patrón de distribución conocido, lo que sugiere que, a pesar de su menor capacidad predictiva (con valores AUC más bajos), podría estar capturando con mayor fidelidad el nicho ecológico de las especies estudiadas.

Esta diferencia en los patrones espaciales predichos no debe interpretarse como una debilidad de uno u otro tipo de modelo, sino como una señal de que cada modelo aporta una visión complementaria del mismo problema ecológico. Mientras que RF y GBM permiten explorar relaciones complejas y obtener alto rendimiento predictivo en presencia de muchas variables, MaxEnt ofrece una herramienta poderosa para modelar idoneidad con datos de presencia solamente, que resulta especialmente útil cuando el objetivo es la inferencia ecológica y no la predicción masiva. En contextos de conservación, esta diferencia puede tener implicaciones sustanciales, ya que el hecho de generalizar o extrapolar fuera del dominio ambiental muestreado puede llevar a decisiones erróneas sobre zonas prioritarias de protección [18].

En definitiva, este estudio demuestra el potencial de la bioinformática como herramienta para abordar preguntas ecológicas complejas mediante el uso combinado de datos ambientales de alta resolución, técnicas estadísticas avanzadas y algoritmos de aprendizaje automático. Lejos de limitarse a un ejercicio de predicción, este enfoque ha permitido interpretar patrones de distribución espacial de tres especies de cetáceos con marcadas diferencias ecológicas, identificando variables clave que condicionan su presencia y evaluando las limitaciones derivadas del muestreo, la escala y los datos disponibles. La elección deliberada de modelos como Random Forest, XGBoost y MaxEnt no solo ha respondido a criterios de rendimiento, sino también a su capacidad para capturar diferentes características del nicho ecológico de las especies, desde la idoneidad ambiental hasta el ajuste al espacio muestreado. Esta

complementariedad, unida a un tratamiento riguroso de los datos, ha permitido alcanzar niveles de precisión más que aceptables y extraer inferencias ecológicas consistentes.

En un contexto de cambio global, en el que las condiciones físico-químicas del océano están siendo alteradas a ritmos sin precedentes, el desarrollo de modelos predictivos robustos se convierte en una necesidad urgente para anticipar desplazamientos, redefinir áreas de alta sensibilidad ecológica y apoyar la conservación basada en la evidencia. La bioinformática, como disciplina transversal, no solo proporciona las herramientas necesarias para este tipo de análisis, sino que también establece un marco reproducible, escalable y adaptable a nuevas fuentes de datos y escenarios futuros. Trabajos como este no solo contribuyen al comportamiento ecológico de especies clave, sino que permiten construir las bases para una gestión más eficaz, activa y científicamente rigurosa sobre los ecosistemas marinos y su futuro.

7. Conclusiones

Este estudio ha demostrado que la aplicación de algoritmos de aprendizaje automático constituye una herramienta eficaz para modelar la distribución espacial de cetáceos migratorios, permitiendo extraer patrones diferenciales entre especies y vincularlos con distintas variables ambientales clave.

El uso de medianas anuales y utilizar una resolución espacial determinada ha permitido armonizar los conjuntos de datos de distinta naturaleza y reducir el ruido derivado de la variabilidad interanual, contribuyendo a una mayor estabilidad en la predicción. En este contexto, se confirma la relevancia de la clorofila-a como proxy de productividad primaria, especialmente para *Balaenoptera musculus*, cuya presencia muestra una correlación positiva con zonas de alta productividad superficial.

Además, se ha podido verificar que algoritmos basados en ensamblados, como Random Forest y XGBoost, presentan un rendimiento superior en términos de AUC frente a métodos lineales, lo que evidencia la existencia de relaciones no lineales entre las variables predictoras escogidas y la distribución de las especies. Asimismo, se valida el uso de las curvas ROC y valores AUC como métricas de evaluación robustas.

Un aspecto crítico identificado en el estudio es la influencia del método de generación de pseudo-ausencias en el rendimiento de los modelos. Se concluye que la selección aleatoria no estratificada puede inducir sesgos significativos, especialmente en especies con distribuciones restringidas, por lo que se recomienda incorporar técnicas de muestreo con esfuerzos que puedan proporcionar datos fiables tanto de las presencias como de las ausencias.

A nivel ecosistémico, se han detectado patrones diferenciales en la respuesta ambiental de las tres especies. *Balaenoptera musculus* mostró preferencia por áreas de alta productividad superficial y gradientes batimétricos moderados. *Physeter macrocephalus* se asoció con zonas de mayor profundidad y baja variabilidad térmica, reflejando su estrategia de buceo profundo. *Orcinus orca*, por su parte, presentó un patrón más amplio y disperso, compatible con su alta plasticidad ecológica y estrategia trófica versátil.

Los modelos basados en MaxEnt, aunque tienen valores predictivos inferiores respecto a los modelos tradicionales, han permitido una representación espacial más detallada del nicho ecológico de cada especie, evidenciando su utilidad en contextos donde se dispone únicamente de datos de presencia.

Finalmente, los mapas generados permiten visualizar áreas de alta probabilidad de presencia que podrían corresponder con hábitats críticos o corredores migratorios. Estos resultados proporcionan una base sólida para el diseño de medidas de conservación más específicas, la delimitación de áreas marinas protegidas y las evaluaciones de riesgo e impacto de actividades humanas en estas zonas.

Se concluye, por tanto, que el enfoque propuesto no solo permite avanzar en la comprensión de la relación entre tres de los grandes cetáceos marinos y su entorno, sino que además sienta las bases para su aplicación en escenarios de gestión y conservación marina. No obstante, su potencial puede incrementarse en futuras investigaciones mediante la incorporación de más datos de mayor resolución temporal (diaria o semanal), la integración de más variables dinámicas y la inclusión de datos de ausencias verificadas o esfuerzos de muestreo, que permitirían mejorar la calidad y representación de las predicciones de los distintos modelos y, en última instancia, reflejarían con más exactitud el nicho ecológico de cada especie.

Referencias bibliográficas

- [1] Chambault P, Fossette S, Heide-Jørgensen MP, Jouannet D, Vély M. Predicting seasonal movements and distribution of the sperm whale using machine learning algorithms. *Ecol Evol* 2021;11:1432–45. <https://doi.org/10.1002/ece3.7154>.
- [2] Whitehead H. Density-dependent habitat selection and the modeling of sperm whale (*Physeter macrocephalus*) exploitation. vol. 57. 2000.
- [3] Sergio F, Newton I, Marchesi L. Top predators and biodiversity: Much debate, few data. *Journal of Applied Ecology* 2008;45:992–9. <https://doi.org/10.1111/j.1365-2664.2008.01484.x>.
- [4] Clarke MR. Cephalopods as prey. III. Cetaceans. *Philosophical Transactions of the Royal Society B: Biological Sciences* 1996;351:1053–65. <https://doi.org/10.1098/rstb.1996.0093>.
- [5] Virgili A, Teillard V, Dorémus G, Dunn TE, Laran S, Lewis M, et al. Deep ocean drivers better explain habitat preferences of sperm whales *Physeter macrocephalus* than beaked whales in the Bay of Biscay. *Sci Rep* 2022;12. <https://doi.org/10.1038/s41598-022-13546-x>.
- [6] Roman J, Estes JA, Morissette L, Smith C, Costa D, McCarthy J, et al. Whales as marine ecosystem engineers. *Front Ecol Environ* 2014;12:377–85. <https://doi.org/10.1890/130220>.
- [7] Estes JA, Terborgh J, Brashares JS, Power ME, Berger J, Bond WJ, et al. Trophic Downgrading of Planet Earth. n.d.
- [8] Sahri A, Putra MIH, Mustika PLK, Murk AJ. A treasure from the past: Former sperm whale distribution in Indonesian waters unveiled using distribution models and historical whaling data. *J Biogeogr* 2020;47:2102–16. <https://doi.org/10.1111/jbi.13931>.
- [9] Maglietta R, Saccotelli L, Fanizza C, Telesca V, Dimauro G, Causio S, et al. Environmental variables and machine learning models to predict cetacean abundance in the Central-eastern Mediterranean Sea. *Sci Rep* 2023;13. <https://doi.org/10.1038/s41598-023-29681-y>.

- [10] Castañeda OG. ESTIMACIÓN DE LA PRESENCIA Y DISTRIBUCIÓN DE CETÁCEOS EN EL ALTO GOLFO DE CALIFORNIA n.d. <https://doi.org/10.13140/RG.2.2.18324.53127>.
- [11] Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ. A statistical explanation of MaxEnt for ecologists. *Divers Distrib* 2011;17:43–57. <https://doi.org/10.1111/j.1472-4642.2010.00725.x>.
- [12] Guisan A, Zimmermann NE. Predictive habitat distribution models in ecology. vol. 135. 2000.
- [13] Elith J, Leathwick JR. Species distribution models: Ecological explanation and prediction across space and time. *Annu Rev Ecol Evol Syst* 2009;40:677–97. <https://doi.org/10.1146/annurev.ecolsys.110308.120159>.
- [14] Redfern J V, Ferguson MC, Becker EA, Hyrenbach KD, Good C, Barlow J, et al. Techniques for cetacean-habitat modeling. vol. 310. 2006.
- [15] Anderson RC, Branch + TA, Alagiyawadu ^ A, Baldwin R, Marsac F. Seasonal distribution, movements and taxonomic status of blue whales (*Balaenoptera musculus*) in the northern Indian Ocean. n.d.
- [16] Pirotta E, Brotons JM, Cerdà M, Bakkers S, Rendell LE. Multi-scale analysis reveals changing distribution patterns and the influence of social structure on the habitat use of an endangered marine predator, the sperm whale *Physeter macrocephalus* in the Western Mediterranean Sea. *Deep Sea Res 1 Oceanogr Res Pap* 2020;155. <https://doi.org/10.1016/j.dsr.2019.103169>.
- [17] Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. *Journal of Animal Ecology* 2008;77:802–13. <https://doi.org/10.1111/j.1365-2656.2008.01390.x>.
- [18] Elith J, Franklin J. Species Distribution Modeling ☆. Reference Module in Life Sciences, Elsevier; 2017. <https://doi.org/10.1016/b978-0-12-809633-8.02390-6>.
- [19] Chau TTT, Gehlen M, Chevallier F. A seamless ensemble-based reconstruction of surface ocean pCO₂ and air-sea CO₂ fluxes over the global coastal and open oceans. *Biogeosciences* 2022;19:1087–109. <https://doi.org/10.5194/bg-19-1087-2022>.
- [20] Branch TA, Stafford KM, Palacios DM, Allison C, Bannister JL, Burton CLK, et al. Past and present distribution, densities and movements of blue whales *Balaenoptera musculus*

- in the Southern Hemisphere and northern Indian Ocean. *Mamm Rev* 2007;37:116–75. <https://doi.org/10.1111/j.1365-2907.2007.00106.x>.
- [21] Dietz R, Rikardsen AH, Biuw M, Kleivane L, Noer CL, Stalder D, et al. Migratory and diurnal activity of North Atlantic killer whales (*Orcinus orca*) off northern Norway. *J Exp Mar Biol Ecol* 2020;533. <https://doi.org/10.1016/j.jembe.2020.151456>.
- [22] Oppel S, Meirinho A, Ramírez I, Gardner B, O’Connell AF, Miller PI, et al. Comparison of five modelling techniques to predict the spatial distribution and abundance of seabirds. *Biol Conserv* 2012;156:94–104. <https://doi.org/10.1016/j.biocon.2011.11.013>.
- [23] Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, et al. RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY. vol. 88. 2007.
- [24] Jiménez-Valverde A, Lobo JM, Hortal J. Not as good as they seem: The importance of concepts in species distribution modelling. *Divers Distrib* 2008;14:885–90. <https://doi.org/10.1111/j.1472-4642.2008.00496.x>.
- [25] Yackulic CB, Chandler R, Zipkin EF, Royle JA, Nichols JD, Campbell Grant EH, et al. Presence-only modelling using MAXENT: When can we trust the inferences? *Methods Ecol Evol* 2013;4:236–43. <https://doi.org/10.1111/2041-210x.12004>.
- [26] Anderson RP, Martínez-Meyer E, Nakamura M, Araújo MB, Peterson AT, Soberón J, et al. *Ecological Niches and Geographic Distributions (MPB-49)*. Princeton University Press; 2011. <https://doi.org/10.1515/9781400840670>.