



**HAL**  
open science

# Species distribution modelling of the Southern Ocean benthos: a review on methods, cautions and solutions.

Charlène Guillaumot, Bruno Danis, Thomas Saucède

► **To cite this version:**

Charlène Guillaumot, Bruno Danis, Thomas Saucède. Species distribution modelling of the Southern Ocean benthos: a review on methods, cautions and solutions.. Antarctic Science, 2021, 33 (4), pp.349-372. 10.1017/S0954102021000183 . hal-03347312

**HAL Id: hal-03347312**

**<https://hal.science/hal-03347312>**

Submitted on 12 Feb 2024

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

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



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

# **Species Distribution Modelling of the Southern Ocean benthos: a review on methods, cautions and solutions**

**Running title: Review of methods and limits for Southern Ocean SDM**

Guillaumot Charlène<sup>1,2\*</sup>, Danis Bruno<sup>1</sup>, Saucède Thomas<sup>2</sup>

<sup>1</sup> Université Libre de Bruxelles, Marine Biology Lab. Avenue F.D. Roosevelt, 50. CP 160/15 1050  
Bruxelles, Belgium

<sup>2</sup> UMR 6282 Biogéosciences, Univ. Bourgogne Franche-Comté, CNRS, 6 bd Gabriel F-21000 Dijon,  
France

\* charleneguillaumot21@gmail.com

## **ORCID numbers:**

Charlène Guillaumot: <https://orcid.org/0000-0002-5507-511X>

Bruno Danis: <https://orcid.org/0000-0002-9037-7623>

Thomas Saucède: <http://orcid.org/0000-0001-6056-4447>

## ABSTRACT

Species Distribution Modelling (SDM) studies the relationship between species occurrence records and their environmental setting, providing a valuable approach to predict species distribution in the Southern Ocean (SO), a challenging region to investigate due to remoteness and extreme weather and sea conditions. The specificity of SO studies, including restricted field access and sampling, the paucity of observations and difficulties to conduct biological experiments, limit the performance of SDMs.

In this review, we discuss some issues that may influence model performance when preparing datasets and calibrating models, namely the selection and quality of environmental descriptors, the spatial and temporal biases that may affect the quality of occurrence data, the choice of modelling algorithms, and the spatial scale and limits of the projection area.

We stress the importance of evaluating and communicating model uncertainties, and most common evaluation metrics are reviewed and discussed accordingly. Based on a selection of case studies on SO benthic invertebrates, we highlight important cautions to take, and pitfalls to avoid when modelling the distribution of SO species, and provide some guidelines along with potential methods and original solutions that can be used for improving model performance.

**Keywords:** Antarctica, marine benthic invertebrates, modelling performance, biases, limits

## 1. Introduction

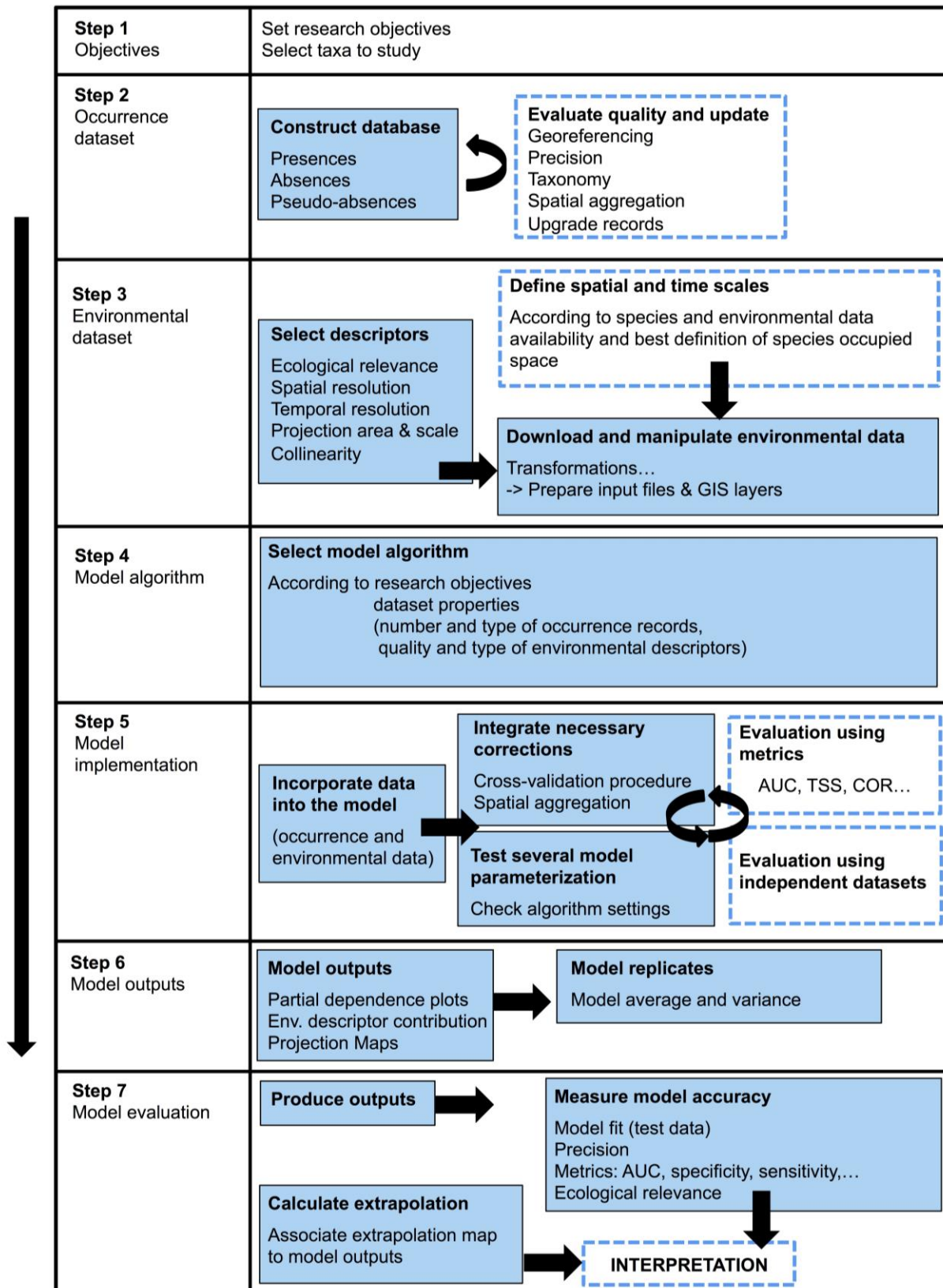
Due to its remoteness and extreme weather and sea conditions, the Southern Ocean (SO) is a challenging region to carry out biological studies (Kaiser et al. 2013, Gutt et al. 2017). It is also one of Earth's regions where we observe the most rapid and dramatic environmental changes in marine ecosystems, motivating the study of these marine communities (Turner et al. 2014, Ashton et al. 2017, Clark et al. 2019). Ecological modelling approaches are now well-established and can be used to predict spatial patterns of organisms', populations' and species' distributions and assess their environmental drivers (Peterson et al. 2011). Based on field observations and experimental datasets, ecological modelling encompasses valuable approaches to help analyse biological data and interpolate our knowledge of species distribution in relation to environmental descriptors (Kennicutt et al. 2014).

Species distribution models (SDMs) are ecological models that study the statistical relationship between species occurrence records and environmental factors, determining the set of environmental conditions that is suitable to species distribution (Elith et al. 2006, Elith and Leathwick 2009, Peterson et al. 2011). They represent the species realised niche (Pearson

2007, Sillero 2011), being the ensemble of abiotic conditions for which the species survives and reproduces, adding into consideration the influence of biotic interactions (competition, predation, parasitism, symbiosis...) (Hutchinson 1957). SDMs have been widely used in various fields of ecology, such as conservation biology, biogeography, paleoecology and global change biology (Pearson 2007). In the recent years, a growing number of ecological studies have been using SDMs to analyse the distribution of marine pelagic and benthic species in the SO (e.g. marine invertebrates, fish, sea birds and marine mammals) and determine species environmental preferences (Loots et al. 2007, Pierrat et al. 2012, Xavier et al. 2016, Nachtsheim et al. 2017), compare ecological niche predictions as a response to changing environments (Basher and Costello 2016, Gallego et al. 2017, Guillaumot et al. 2018b, Jerosch et al. 2019) or identify diversity hotspots for conservation purposes (Pinkerton et al. 2010, Hibberd 2016, Thiers et al. 2017).

However, the quality of ocean-wide models is often limited by the heterogeneity, amount and spatial distribution of data, along with limited temporal and spatial resolutions. For all these reasons, both modelling methods and model construction should be tested for accuracy and robustness prior to interpretation, and these indicators should be transparently communicated to ensure that model outputs are relevant given the specificities of datasets used for modelling.

In the present paper, we review the most common methodological issues encountered in species distribution modelling applied to the SO, following the above flowchart (Fig.1). Challenges regarding occurrence and environmental datasets peculiarities are described. The choice of SDM algorithm, and procedures to implement and evaluate models are addressed. Based on benthic invertebrate case studies, we stress important precautions to take, and pitfalls to avoid at common steps of SDM implementation. Finally, we aim to provide some guidelines with a set of potential methods and original solutions that can be used for improving model performance.



**Figure 1.** Flow chart of the SDM construction process. Steps 1 to 4 concern data collection, and treatment. Steps 5 to 7 integrate procedures for model implementation and evaluation. Dashed rectangles allow for a possible step backward when assessing model uncertainties or evaluating model performance.

## 2. Quality of datasets

### 2.1.1. *Environmental datasets: field data.*

Preparing environmental datasets is the first encountered challenge when generating models (Gutt et al. 2012, De Broyer et al. 2014). The SO, here defined as waters south of 45°S latitude, covers an extensive area of over 20 million km<sup>2</sup> (Breitzke 2014). Having access to environmental data with good temporal and spatial resolutions at such a broad scale is challenging, an issue common to all broad scale oceanographic studies (Robinson et al. 2017). Broad scale is defined here as the entire SO, regional scale as smaller areas of a few hundreds square kilometers and local scale as a few square kilometers to square meters (Gage 2004). Oceanographic data acquisition in the field is strongly conditioned by weather and sea conditions along with the seasonality of polar regions (polar night and dense sea ice coverage in winter) that prompt recurring gaps in the acquisition of environmental data in the SO. Data are also much more sampled close to research stations and along main sailing routes (Guillaumot et al. 2019). This is particularly striking in regions such as the southwestern Weddell Sea, along the shores of the Western Antarctic Peninsula, and in the Bellingshausen and Amundsen seas (Clarke et al. 2007, Griffiths et al. 2014).

### 2.1.2. *Environmental datasets: satellite derived-data.*

Satellite derived-data form a significant source of information for SO oceanographic studies. Providing valuable environmental indicators at broad spatial scale, they can give details about continuous and long-term measurements of water masses including sea-ice coverage, extent, and duration, sea surface temperatures and salinities, biogeochemical parameters, sea level, primary production, and typical meteorological parameters (El Mahrad et al. 2020).

The accuracy of satellite data however should be considered with care, given detection limits, interpolations that reduce the influence of atmospheric particulate scatter and the use of interpolation and gap filling methods that smooth raw data at broad spatial and temporal scales (Pope et al. 2017, Stock et al. 2020).

Whenever possible, it is recommended to validate environmental data derived from satellite products at regional and local scales, by comparing pixels on a satellite image with 'real' field observation data (Henson et al. 2015, Trull et al. 2018). Simple correlation analyses, or more complex ground-truth processes are available to compare satellite and *in situ* data and secure the interpretation of satellite-derived products (White-Newsome et al. 2013, Allan 2014). This however constitutes a huge task and is not performed generally before implementing SDMs.

### 2.1.3. *Environmental datasets: access to datasets.*

Environmental data generated at the scale of the entire SO can be accessed for free through different web portals such as the NASA's OceanColor Web (<https://oceancolor.gsfc.nasa.gov/>), where satellite-derived data, averaging different temporal measurements down to 4 km resolution are available at the scale of the entire SO, since 2000. These images are post-processed to characterise sea surface temperature or ocean color as a proxy of surface productivity.

The NOAA's data center ([WOCE2013, https://www.nodc.noaa.gov/OC5/woa13/woa13data.html](https://www.nodc.noaa.gov/OC5/woa13/woa13data.html)) also makes available post-processed data of ocean temperature, salinity, oxygen concentration and nutrients at different grid formats, down to 0.25° resolution, averaging six decades (from 1955 to 2012). Bio-Oracle (<https://www.bio-oracle.org/>) compiles a large panel of marine data layers at 1° spatial resolution for different depth layers and time periods, for the present (2005-2012) and the future (2040-2050; 2090-2100) ([Assis et al. 2018](#)). Finally, GEBCO (<https://www.gebco.net/>) is the reference platform for very high-resolution bathymetry data (~500 m resolution) of the world's ocean.

Several works also make available compilation of these SO datasets dedicated to ecological modelling in the SO; they represent a valuable source of information to start with data preparation and modelling ([Raymond 2012](#), [Fabri-Ruiz et al. 2017b](#), [Guillaumot et al. 2018c](#)).

An increasing number of environmental data collected during SO oceanographic campaigns were made accessible for regional-scale studies. Several web portals aggregate all these field measurements and provide them open access (e.g. <https://www.marine-geo.org/collections/#!/collection/USAP#summary>; <https://www.pangaea.de/>).

#### *2.1.4. Environmental datasets: spatial and temporal resolutions.*

Most environmental data are accessible through broad-scale maps from the aforementioned data portals and available with a finest spatial resolution of ~ 4 km, if not coarser ([Raymond 2012](#), [De Broyer et al. 2014](#), [Fabri-Ruiz et al. 2017b](#), [Guillaumot et al. 2018c](#)). This low resolution strongly hampers the precise assessment of relationships between species occurrences and environmental descriptors ([Pittman 2017](#), [Staveley et al. 2017](#)) and consequently the accuracy of model predictions ([Connor et al. 2018](#)), because the relevance of environmental descriptors is a trade-off between their resolution and their spatial and temporal coverage ([Guisan et al. 2007](#), [Seo et al. 2009](#), [Lauzeral et al. 2013](#), [Vale et al. 2014](#)). It is recommended that the resolution of environmental descriptors used in SDM should be in line with the scale of ecological processes at play and for which species ecophysiological responses show the highest variations, if models are expected to capture most of species-environment relationships ([Austin and van Niel 2011](#)).



The published environmental datasets are often averaged over relatively long periods of time (from years to decades for WOCE2013 or Bio-Oracle). The analysis of inter-annual variations can complement the interpretation of model predictions: the absence of such information does not preclude running models but this should be kept in mind when it comes to interpreting model outputs (Guillaumot et al. 2018a). Important environmental variations within a reference time period may not satisfy the equilibrium criterium between species distribution and environmental conditions, which is a strong prerequisite of SDM (Elith et al. 2006) and may affect the relevance and accuracy of model predictions (Guillaumot et al. 2018a). In this respect, an alternative to improve modelling performance would be using seasonal averages or extreme values as environmental descriptors, rather than pluri-annual to annual averages (Franklin 2009, Bradie and Leung 2017).

#### *2.1.5. Environmental datasets: cartographic projections.*

Considering the poles in numerical analyses has long been a source of difficulties in spatial modelling as the convergence of meridians distorts shapes, surfaces, angles or distances towards high latitudes when using standard cylindric representations such as the Mercator projection (Deleersnijder et al. 1993, Eby and Holloway 1994, Murray 1996). Working with conical or azimuthal projections (e.g. polar stereographic system) helps maintain the consistency of angles and shapes, and therefore better meets requirements of SO studies, although areas and distances are progressively distorted when moving away from the pole (Mulcahy and Clarke 2001).

Mapping environmental descriptors and projecting model predictions can be carried out with either square or hexagonal pixels. Each option does not alter image quality and hexagonal shapes may even offer some advantages (Kamgar-Parsi and Sander 1989, Tirunelveli et al. 2002). However, some contrasts may be present between images using square or hexagonal pixels, as each pixel measure the average environmental conditions in the considered surface (Vanden Berghe et al. 2013).

Subdividing the study area into subregions and using different pixel shapes can be a good solution to improve the relevance of representations (Vanden Berghe et al. 2013, Cryer 2015). It can also be suggested to evaluate the accuracy of environmental values captured both in square and hexagonal pixels using baseline *in situ*, field measurements. This is yet to be tested for ecological modelling studies for the SO.

#### *2.1.6. Environmental datasets: future forecasts.*

Since 1992, future climate models are constantly being updated through the efforts of the Coupled Model Intercomparison Projects (CMIP) featured by the Intergovernmental Panel on



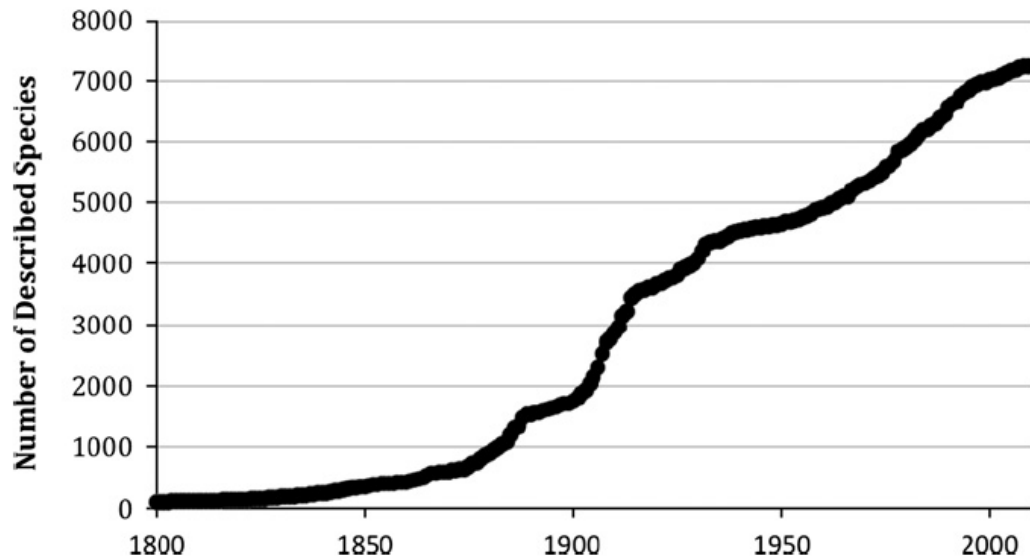
Climate Change (IPCC) Assessment Reports (AR) with the aim of providing a plausible representation of future climate in link with potential anthropogenic impacts (IPCC 2000, Mearns et al. 2001). Last updates (CMIP5, CMIP6) of climate models were driven by different possible future greenhouse gas emission scenarios (Representative Concentration Pathways RCP2.6, RCP4.5, RCP6.0, RCP8.5, from the least to the most pessimistic scenario for CMIP5 and Shared Socioeconomic Pathways, SSP1 to SSP5 for CMIP6) and are built upon the average of an ensemble of simulations (Hayhoe et al. 2017). Future climate models for the SO are available through two main online platforms, Bio-Oracle (<https://www.bio-oracle.org/>, Assis et al. 2018) and the NOAA's portal (<https://psl.noaa.gov/ipcc/ocn/>).

The relevance of using future predictions based on the global assessment scenarios for marine studies has been widely questioned (Flato et al. 2014, Frölicher et al. 2016, de la Hoz et al. 2018), including their use for SDM, given that, climate models mainly rely on untestable assumptions (Beaumont et al. 2008, Gotelli and Stanton-Geddes 2015, Cavanagh et al. 2017, Freer et al. 2018), future layers are not always available for oceanographic studies (Fabri-Ruiz 2018, Guillaumot et al. 2018a, 2018b), discrepancies between present observations and future predictions can be problematic (Jiménez-Valverde et al. 2020), and models are based on a representation of the climate system that has a complex cascading effect on ecological processes (Cavanagh et al. 2017). Cavanagh et al. (2017) examined how well IPCC-class models reproduced sea-ice conditions. By subsetting CMIP5 models that best describe spatial extent and temporal ice cover, they improved the precision of projected future sea ice distribution which better suited to ecological analyses. Extending this method to other key oceanographic parameters should contribute to improving the accuracy of future climate models for the SO and their relevance to ecological studies.

## **2.2. Occurrence datasets**

### *2.2.1. Occurrence datasets: historical compilation.*

Biological sampling in the SO began with the first expeditions of the HMS *Challenger* (1873-1876). Sampling effort has considerably increased over the second part of the 20<sup>th</sup> century and during the last decades in particular, following technological advances that have enabled the access to remote regions and sample processing (Fig. 2).



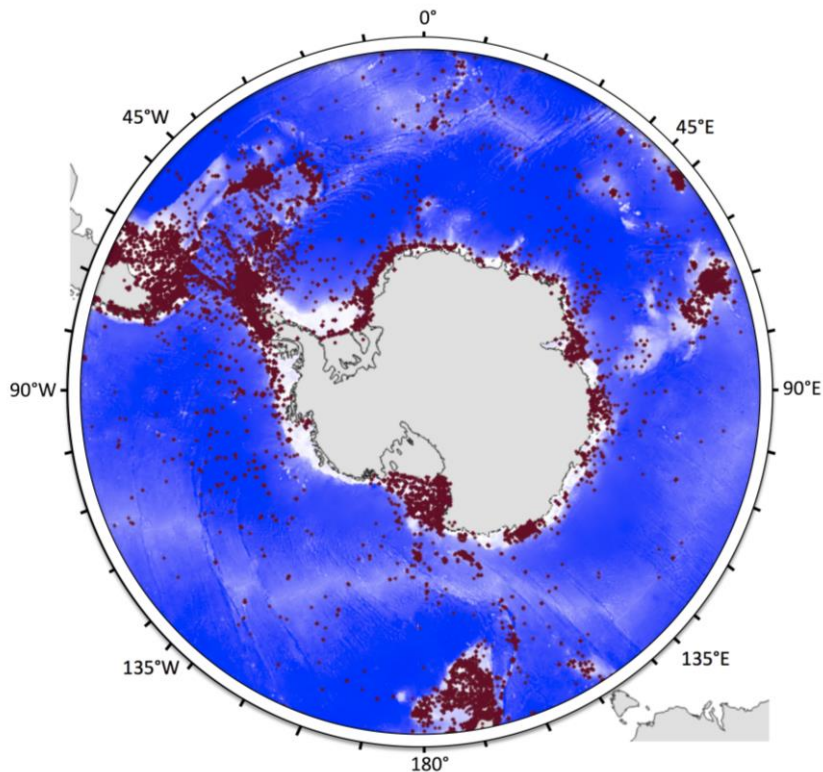
**Figure 2.** Cumulative number of Antarctic species described over time, according to data available in the Register of Antarctic Marine species (until March 2010). From De [Broyer and Danis \(2011\)](#).

This long lasting and irregular effort in biogeographic (occurrence) data collection has had an impact on data compilation and resulted in heterogeneous datasets, as observed in several datapapers and associated IPT databases (Integrated Publishing Toolkit) such as [Guillaumot et al. \(2016\)](#), [Fabri-Ruiz et al. \(2017a\)](#) or [Moreau et al. \(2018\)](#), or in the general platform [biodiversity.aq \(2020\)](#) webportal.

The historical compilation of biological data include (1) taxon misidentifications and taxonomic inconsistencies due to the various taxonomic revisions published through time, (2) errors in the georeferencing of occurrence records due to contrasting nomenclatures used to report latitude and longitude, (3) the accumulation of errors in metadata through the different generations of curation and (4) errors due to the use of different coordinate projection systems. Finally, in cases where species distributions may have shifted with time, species environmental preferences may have changed or non-contemporaneous environmental or occurrence datasets are used, discrepancies between occurrence records and environmental conditions can be present and violate the environment-occurrence equilibrium assumption necessary to generate SDMs. All these side effects were reviewed in detail by [Newbold \(2010\)](#). The impact on species niche definition and SDM predictions have been reported in many works ([Ensing et al. 2012](#), [Lahoz-Monfort et al. 2014](#), [Monk 2014](#), [Aguir et al. 2015](#), [Tessarolo et al. 2017](#), [Guillaumot et al. 2018a](#)) that all advice to thoroughly check datasets for quality management prior to running models.

### 2.2.2. Occurrence datasets: spatial aggregation.

Most species occurrence data were collected in the vicinity of research bases or their surroundings, or along recurrent maritime routes leading to clear spatial aggregation patterns in biological datasets (Griffiths et al. 2014, Guillaumot et al. 2019) (Fig. 3).



**Figure 3.** Distribution of benthos sampling sites (red dots) in the Southern Ocean (SO, < 45°S). Sampling sites are not evenly distributed in the SO, showing important spatial aggregation in the Scotia Arc region and Western Antarctic Peninsula with several clusters along the Antarctic shelf, and over the Kerguelen and Campbell plateaus. In contrast, deep-sea regions and remote areas of the Antarctic shelf are under-sampled. From Guillaumot et al. (2019), updated from Griffiths et al. (2014).

Spatial aggregation can affect model accuracy, as aggregated presence records do not fully and homogeneously represent the entire environment that is occupied by given species. This aggregation also violates an initial assumption of SDM that requires independence between records (Araújo and Guisan 2006, Hijman 2012). This may bias model predictions (Luoto et al. 2005, Segurado et al. 2006, Dormann 2007, Kühn 2007, Crase et al. 2012), leading to statistical artefacts and generating inaccurate patterns (Bahn and McGill 2007, Currie 2007).

Spatial aggregation of data and the effect of this spatial aggregation on model outputs can be quantified using the Moran's I index, which estimates the spatial autocorrelation between presence records used to build the model and predicted presence probabilities (Luoto et al. 2005). This spatial autocorrelation implies that close pixels are expected to present more similar predicted probabilities than distant ones, due to the short geographic distance between records

rather than to environmental similarities alone. Testing and correcting for this bias should help reduce its impact on model predictions (see section 3.3) (Diniz-Filho et al. 2003, Kühn 2007).

### 2.2.3. Occurrence datasets: presence-only records.

SDMs based on presence/absence data are recognized as having better predictive performance than models using presence-only data (Zaniewski et al. 2002, Brotons et al. 2004, Lobo et al. 2010, Wisz and Guisan 2009, Smith 2013, Carvalho et al. 2015, Peel et al. 2019). However, except for some local-scale studies (e.g. Robinson et al. 2011), in most oceanographic studies species absence records are usually not available for SDM, and working with presence-only records is the only alternative (Lobo et al. 2010). SDMs are then built by associating presence-only records to a random selection of background records that will be used to characterise the full environmental conditions (Franklin 2010, Barbet-Massin et al. 2012). Background records should not be mistaken with pseudo-absence records that are artificial absence data, where the species is supposed (but not confirmed) to be absent. Pseudo-absence records do not represent the overall conditions of the study area. Presence/pseudo-absence models consist in another modelling approach, rather predicting occupied and unoccupied habitats than suitable and less suitable habitats for presence-background modelling (Sillero and Barbosa 2020).

Presence-only datasets may contain several uncertainties that can bias model predictions. (1) Working on rare or cryptic species is generally prone to taxonomic misidentifications that may either contract or, alternatively, expand the extent of predicted species distributions (Costa et al. 2015, Aubry et al. 2017). Such biases due to taxonomic errors were shown to be highly variable and depend on experts identifying specimens as suggested by Beale and Lennon (2012) who worked on a compilation of several collections. (2) Sampling gear may have an impact on species detection. Inaccurate species observations may generate false positives (species predicted as present while it was not sampled or observed in the field) and false negatives (species predicted as absent while it was sampled or observed in the field) during model initialization (Guillera-Arroita 2016). Species presence records should be carefully scrutinized prior to modelling (Lozier et al. 2009) or at least, records should be categorized into different subsets of data verifiability (Aubry et al. 2017). (3) Georeferencing errors are a frequent issue in databases (Murphey et al. 2004, Maldonado et al. 2015). This is especially the case in large databases compiling independent datasets using species presences recorded with varying levels of precision (Graham et al. 2008, Bloom et al. 2018). Several studies simulated virtual random georeferencing errors and showed that these errors lead to a significant drop in model performances, inconsistencies in the respective contributions of environmental descriptors contribution also influencing model interpretation (Graham et al. 2008, Osborne and Leitão

2009, Naimi et al. 2011). These side effects seem to be minimized in local-scale models, here again advocating for the use of local-scale models whenever possible (Mitchell et al. 2017).

#### *2.2.4. Occurrence datasets: dealing with small datasets.*

Usually, the number of species presence records available for modelling is relatively limited considering the wide geographic extent of the SO (De Broyer et al. 2014). Generating SDMs with small datasets may include many pitfalls. (1) It reduces the potential of SDMs to transfer in space and time (Hernandez et al. 2006, Raes 2012), (2) it truncates predicted distribution and niche definition (Hortal et al. 2007, 2008, Rocchini et al. 2011, Sánchez-Fernández et al. 2011, Titeux et al. 2017, El-Gabbas and Dormann 2018), (3) it reduces modelling goodness-of-fit as the model may wrongly represent the reality (Stockwell and Peterson 2002, McPherson et al. 2004, Pearson et al. 2007, Wisz et al. 2008, Liu et al. 2019), (4) it increases instability between model replicates (Guillaumot et al. 2018a), (5) it gives rise to methodological constraints on threshold selection (Jiménez-Valverde and Lobo 2007, Bean et al. 2012) or (6) on the application of evaluation metrics (Pearson et al. 2007), (7) it complicates the identification of model optimal complexity (Galante et al. 2018) and (8) it leads to a reduction in model accuracy because presence and background datasets would not differ markedly (Luoto et al. 2005).

Alternatives are being developed to produce more accurate models based on a limited amount of presence records. One solution is generating several models performed on restricted areas and datasets with more detailed information and then averaging them with a weighted ensemble approach. This “ensemble of small models” approach had higher performance compared to single models (Lomba et al. 2010, Breiner et al. 2015, 2018).

Another alternative is to restrict the prediction area according to where occurrence records are found and ensuring upstream that the number of records is sufficient to precisely characterise the species environmental preferences: a trivial advice surprisingly neglected as recently pointed out by Morales et al. (2017) and Araújo et al. (2019).

#### *2.2.5. Occurrence datasets: definition of species occupied environmental space.*

Spatial aggregation, along with heterogeneity, limited size and uncertainties in datasets can strongly bias the quantification of the species occupied environmental space (Hortal et al. 2008, Newbold 2010, Tassarolo et al. 2017). However, accurately defining species occupied space is the cornerstone of SDM initialization (Elith et al. 2006, Boulanger et al. 2018).

Moreover, SDM supposes that species are in equilibrium with the environmental conditions they inhabit. SDM does not take into consideration potential vagrants that have dispersed out of their usual environmental range or populations that could momentarily survive in unsuitable habitats, because it violates the equilibrium assumption between species distribution and environmental

conditions (Beale and Lennon 2012). These elements should be cautiously considered when preparing datasets prior to generating models, by removing any atypical record.

Over the last two decades, field data acquisition was expanded by the use of biologging technology with electronic devices attached on seabirds and marine mammals to access the position of species all year long (Raymond et al. 2015, Ropert-Coudert et al. 2020). These data uncover the hidden behaviour of marine animals, constitute a powerful way of better estimating species occupied space; they can also be used to validate and refine the understanding of environmental conditions prevailing in those species distribution areas (Arthur et al. 2017, Nachtsheim et al. 2017, Hindell et al. 2020).

### **3. Adapting model implementation to datasets**

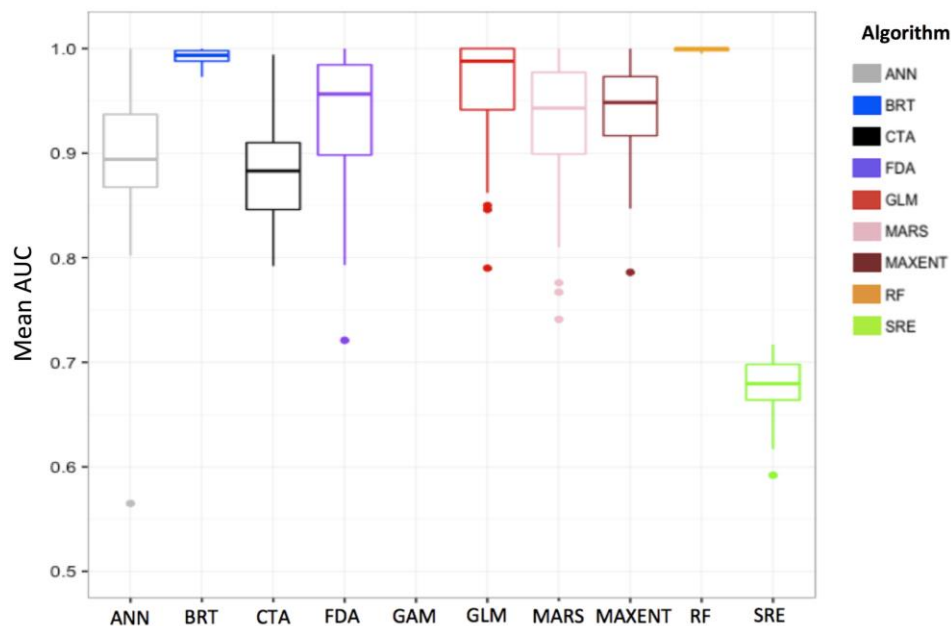
#### *3.1. The choice of modelling algorithms*

To run performant SDMs, several assumptions must be tested and computing methods adapted to each case study (Austin 2002, de la Hoz et al. 2019). Among them, the choice of the modelling algorithm should be of major concern, since no algorithm works best for all species, in all areas, at all spatial scales, and all time periods (Jarnevich et al. 2015, Qiao et al. 2015). The selection and parameterization of modelling algorithms were proved to be major causes of variation between SDM predictions (Diniz-Filho et al. 2009, Dormann et al. 2008, Buisson et al. 2010, Watling et al. 2015, Boulanger et al. 2018). Each algorithm is particularly suited to deal with a specific type and quality of data (Guisan and Zimmermann 2000, Austin 2002, Elith et al. 2006, Peterson 2011, Guisan et al. 2017), which will determine final model outputs (Aguirre-Gutiérrez et al. 2013, Beaumont et al. 2016).

When modelling species distribution, it is necessary to select appropriate algorithms that have good transferability performances that is, have good abilities to correctly transfer predictions to other geographic space and time periods (Randin et al. 2006), and that they limit overfitting (i.e. mitigate model complexity) while being flexible in integrating complex environmental relationships. Machine-learning algorithms (e.g. MaxEnt, Boosted Regression Trees BRT, Random Forests RF, support vector machines SVM, Vapnik 1998, Breiman 2001, Elith et al. 2008, 2011) give access to important computing performances (Zhou 2012), and are relevant approaches for handling complex relationships between species occurrences and the environment (Olden et al. 2008, Elith and Leathwick 2009). The BRT and RF algorithms are particularly suited to complex and heterogeneous datasets (Fig. 4, Guillaumot et al. 2020a). They were proved efficient in generating performant models with limited overfitting (Elith et al. 2006, Wisz et al. 2008, Wenger and Olden 2012). They can automatically select the most



informative features among a large set (Merow et al. 2014, García-Callejas and Araújo 2016, Guillaumot et al. 2020a) and perform well to generalize predictions in the absence of information or conversely, to deal with redundant information provided by correlated factors (Breiman 1984, De'ath and Fabricius 2000, Friedman 2001).



**Figure 4.** Compared Area Under the Curve (AUC) performances of SDMs generated with different algorithms (ANN=Artificial Neural Network, BRT=Boosted Regression Trees, CTA=Classification Tree Analysis, FDA=Flexible Discriminant Analysis, GAM=Generalised Additive Model, GLM=Generalised Linear Model, MARS=Multivariate Adaptive Regression Splines, MAXENT=Maximum Entropy, RF=Random Forest, SRE=Surface Range Envelope) to predict the distribution of the sea urchin *Sterechinus diadema* in the Southern Ocean. Results show a good performance for BRT and RF, adapted to small, historically compiled datasets (temporally heterogeneous) and spatially aggregated presence-only data. Models were calibrated with presence-only data and 200 background data randomly sampled in the study area. Average scores of 100 model replicates. See Guillaumot et al. (2018b) for details.

The different fields of application and the respective performance of existing algorithms have been extensively compared in various works based on (1) a single species (Pearson et al. 2006: plants in South Africa, Elith and Graham 2009: plant distribution in Southern Australian landscapes, Marmion et al. 2009: european butterflies, Lorena et al. 2011: plants in South America, Beaumont et al. 2016: mammals in Australia), or (2) an ensemble of worldwide distributed terrestrial (Elith et al. 2006) or marine species (Ready et al. 2010), for (3) certain regions only (Guisan et al. 2007: trees in Switzerland, Tsoar et al. 2007: snails, birds and bats in Israel, Reiss et al. 2011: benthic marine species in the North Sea, Bucklin et al. 2015:



vertebrates of Florida) or (4) using virtual species ([Meynard and Quinn 2007](#), [García-Callejas and Araujo 2016](#), [Qiao et al. 2015](#)).

However, to generate such comparisons (Fig. 4) it is important to specifically adjust each algorithm on the case study. Algorithms all perform differently with regards to overfitting, spatial aggregation, and transferability, and comparing model performances using different parameter settings is challenging ([Merow et al. 2014](#)) given that model parameterization has strong effects on the quality of model outputs ([Anderson and Gonzalez 2011](#), [Rodda et al. 2011](#), [Warren and Seifert 2011](#), [Yackulic et al. 2013](#), [Radosavljevic and Anderson 2014](#), [Moreno-Amat et al. 2015](#), [Halvorsen et al. 2016](#), [Galante et al. 2018](#), [Lieske et al. 2018](#)).

Initially developed in the 1990's, ensemble modelling has been increasingly used since then ([Hansen and Salamon 1990](#), [Schapire 1990](#)). Ensemble modelling consists in combining several algorithms ([Zhou 2012](#)), input datasets (occurrence or environmental descriptors datasets) or parameterizations ([Araújo and New 2007](#), and see [Hao et al. 2019](#) for a review of applications). The approach seems interesting as it can provide predictions that take into account the variability of several models ([Araújo and New 2007](#), [Hao et al. 2019](#)).

Ensemble modelling has been used for various studies in SDM ([Araújo and New 2007](#), [Marmion et al. 2009](#), [Thuiller et al. 2009](#), [Buisson et al. 2010](#), [Luedeling et al. 2014](#), [Trolle et al. 2014](#), [Carvalho et al. 2015](#), [Scales et al. 2016](#), [Jerosch et al. 2019](#)) and benefited from the development of R packages to implement them (*Biomod*: [Thuiller et al. 2009](#), *BiodiversityR*: [Kindt et al. 2017](#), *biomod2*: [Thuiller et al. 2018](#), *sdm*: [Naimi et al. 2019](#)).

The main benefits of using ensemble models lie in the fact that the different algorithms will perform differently for various input cases (despite their overall performance). The models thus complement each other, avoiding some biases that may have resulted from using a single algorithm ([Marmion et al. 2009](#), [Knutti 2010](#), [Zhou 2012](#)). However, model interpretation is much more difficult when mixing algorithms implemented differently, with contrasting ways of presenting outputs ([Sillero 2011](#)) and different definitions of thresholds for identifying habitat suitability ([Perrault-Hébert 2019](#)), requiring the normalization of predictions that is rarely applied ([Zhang and Mahadevan 2019](#)). This is the main limitation to the approach and could offset the gain in model performance ([Crimmins et al. 2013](#), [Zhu and Peterson 2017](#), [Hao et al. 2020](#)), that was contested as it was assessed without using an independent evaluation dataset ([Hao et al. 2019](#)). Combining predictions of different models generated with contrasting assumptions is therefore tricky when interpreting the results ([Perrault-Hébert 2019](#)). Optimizing the parameterization of one single algorithm (that could be correctly evaluated) may therefore constitute a more valuable option ([Perrault-Hébert 2019](#)). Comparing the performance of different algorithms can be helpful in a first stage of the modelling process, in order to select the most suited algorithm and calibrate the models ([Massada et al. 2013](#)).

### 3.2. *The choice of environmental descriptors*

The selection of environmental descriptors is also a crucial step of the modelling process (Franklin 2010, Austin and Van Niel 2011, Petitpierre et al. 2017). Ideally, environmental descriptors should be selected for their ecological relevance to the studied organisms (Austin and van Niel 2011, Dormann et al. 2012, Bradie and Leung 2017), they must capture environmental discontinuities and constraints in the distribution area (Jarnevich et al. 2015), and should also be detailed enough to represent the habitat complexity and variability to allow good SDM accuracy and performance (Elith and Leathwick 2009, Barbet-Massin et al. 2012, Bucklin et al. 2015, Petitpierre et al. 2017).

In most studies, the final number of descriptors selected to depict species environment is generally close to 10 (Pierrat et al. 2012, Mormède et al. 2014, Guillaumot et al. 2018a, Fabri-Ruiz et al. 2019). Overall, a small number of descriptors will allow generating less complex models and facilitate interpretation (Austin and van Niel 2011, Braunisch et al. 2013, Bucklin et al. 2015, Petitpierre et al. 2017). In contrast, increasing the number of descriptors potentially increases the effect of collinearity between them (i.e. correlation between values of descriptors), which may lead to statistical artefacts in model predictions if algorithms cannot handle information redundancy (Dormann et al. 2012, Merow et al. 2014). Therefore, collinearity is usually tested beforehand and collinear descriptors are adjusted (practically, one descriptor of a pair is removed) before running the model (Dormann et al. 2012, Merow et al. 2013, Fois et al. 2018). However, Guillaumot et al. (2020a) showed that model complexity, transferability, and accuracy do not significantly change between models generated with different sets including from 4 to 58 collinear descriptors when using the BRT algorithm. BRT automatically keep the most relevant descriptors to describe species distribution and can deal with redundant information (De'ath and Fabricius 2000, Whittingham et al. 2006, Elith et al. 2008), which is not the case for all algorithms (Merow et al. 2014).

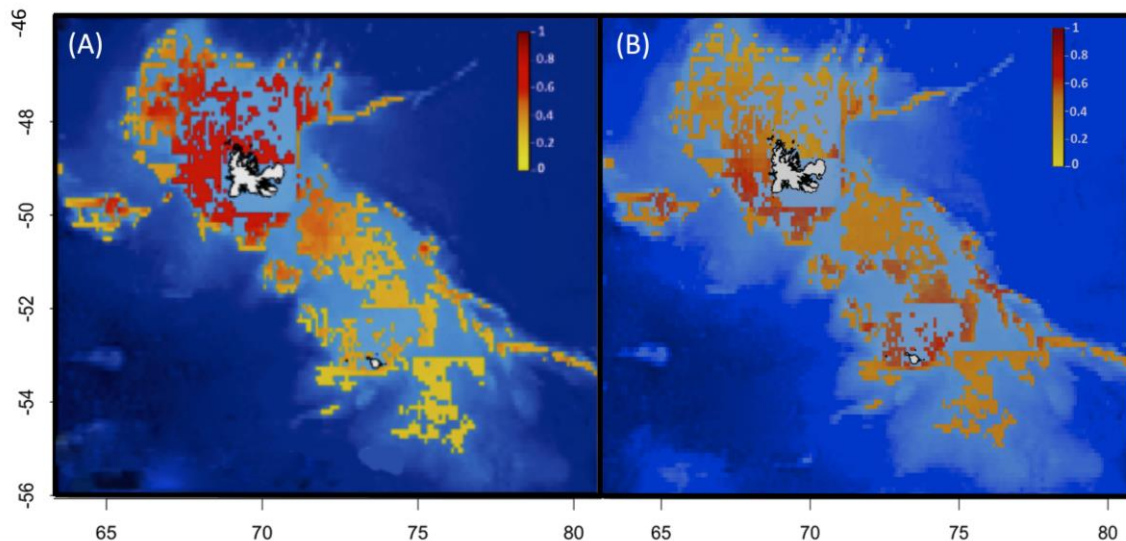
Selecting environmental descriptors therefore implies that several tests should be performed upstream to determine the best set to be used depending on research objectives. Fois et al. (2018) recommended to first calibrate models with a large set of descriptors of different nature (proximal vs. distal descriptors) that will be stepwise pruned, after analysing their ability to accurately describe the habitat and testing collinearity (El-Gabbas and Dormann 2018). Generating, testing, and comparing several sets of descriptors is a widespread strategy to stepwise target the set that allows the best predictive accuracy (Snickars et al. 2014, Bucklin et al. 2015, Bradie and Leung 2017, Petitpierre et al. 2017). Replacing environmental descriptors by Principal Components of a Factorial Analysis was also proved to be efficient, because complex environmental gradients of the study area are simplified in fewer, orthogonalized

components (Kühn 2007, Petitpierre et al. 2017). So far, this latter method has never been applied to SO case studies and should be tested to evaluate the interpretability of model results.

### 3.3. Correcting spatial biases

Generating a model based on spatially aggregated presence-only records may bias predictions with higher probability of occurrence predicted in highly sampled areas (Dormann 2007, Guillaumot et al. 2018a). To compensate for such a bias, a first approach is to sample background records according to the spatial bias introduced by the aggregated presence records themselves (Phillips et al. 2009). The background dataset is used to define the environmental background: its boundaries and variability constitutes essential information to build and project model outputs (Wisiz and Guisan 2009, Barbet-Massin et al. 2012). The choice of the number of background records to be sampled and the extent of their distribution should be considered carefully when calibrating a model, because it can strongly influence model predictions (Chefaoui and Lobo 2008, Lobo et al. 2010, Barbet-Massin et al. 2012, Jarnevich et al. 2017). This number should respect the prevalence score, being the ratio between the species occupied space (represented by presence record locations) and the total surface of the study area (represented by background locations) (McPherson et al. 2004). Some advice are provided in Barbet-Massin et al. (2012) for selecting a correct number of background records according to prevalence scores.

Targeting background records has been extensively tested and several procedures were developed to significantly improve the relevance of models (Fig. 5). Background records can be sampled within pre-defined areas (i.e. 'disks' or 'buffers') close to presence records (Hengl et al. 2009, Phillips et al. 2009, Fourcade et al. 2014, Bertrand et al. 2016), following the presence or absence of other species (Phillips et al. 2009, Syfert et al. 2013, Iturbide et al. 2015, Molloy et al. 2017, Phillips et al. 2017, Ranc et al. 2017), according to probabilities given by a kernel density estimator (KDE) of the sampling frequency (Fourcade et al. 2014, Jarnevich et al. 2017, Guillaumot et al. 2018a, Fabri-Ruiz et al. 2019) or according to additive descriptors of accessibility and sampling effort (El-Gabbas and Dormann 2018). Once again, the selected method should be adapted to each case study and its efficiency tested prior to model interpretation (Støa et al. 2018).



**Figure 5.** Comparison of predicted distribution probabilities (between 0 and 1) of the sea urchin *Ctenocidaris nutrix* on the Kerguelen Plateau: (A) without compensating for sampling bias; (B) with a kernel density estimator (KDE) correction: more background data are sampled in highly sampled areas. The spatial aggregation of presence-only records near the shoreline of the Kerguelen Islands strongly biases model predictions. The KDE correction was proved efficient to correct for such a bias and provides more relevant predictions. From [Guillaumot et al. \(2018a\)](#).

A second method consists in filtering available presence data to reduce the influence of the clustering of species records ([Segurado et al. 2006](#), [Kramer-Schadt et al. 2013](#), [Boria et al. 2014](#)). This is an efficient method compared to the background targeted sampling approach detailed above but the remaining number of presence records after filtering should be sufficient to correctly determine species occupied space ([Kramer-Schadt et al. 2013](#)). Reliable information should be also available to characterise the bias in species occurrence data ([Aiello-Lammens et al. 2015](#), [Sillero and Barbosa 2020](#)). The filtering protocol requires meeting many pre-requisites, but the priority is given to keeping presence data independent and minimizing records clustering ([Alagador D. pers. comm.](#)).

Overall, if several methods were developed to correct for the effect of spatial aggregation on model outputs, it is recommended to interpret model projections performed for poorly sampled areas with great caution ([Phillips et al. 2009](#), [Iturbide et al. 2018](#)).

## 4. Model outputs

### 4.1. Taxonomic bias and population variability

SDMs are usually parameterized using all presence records available for a species and all environmental conditions prevailing at the species records ([Elith and Leathwick 2009](#)). When

modelling species distribution at broad spatial scale, it is often assumed that all populations of a species have the same relationship to environmental conditions over the entire distribution area (Pierrat et al. 2012, Xavier et al. 2016, Guillaumot et al. 2018b, Fabri-Ruiz et al. 2019). However, occurrence datasets may include a set of populations with different phenotypic plasticities (Chevin et al. 2010), transgenerational adaptations (Dixon et al. 2015) or simply different habitat selection in the case of vagile species. Therefore, the modelled species can actually present different abilities to respond to environmental changes. In particular, physiological performances of populations are likely to vary in marine species with wide distribution range and high dispersal capabilities over long distances (Thatje 2012). This is particularly relevant with regards to future predictions that do not integrate inter-population variability in the potential acclimation of species, and may lead models to alternatively over- or under-estimate the distribution of species suitable environments (Cacciapaglia and van Woesik 2017, Thyrring et al. 2017).

Phylogeographic studies have also regularly revealed the existence of cryptic species in the SO benthos, which show similar morphologies for distinct genotypes and potentially, distinct ecological requirements and geographic distributions (Lozier et al. 2009). Such studies often stressed the need of taxonomic revisions (González-Wevar et al. 2019, Ocaranza-Barrera et al. 2019, Moreau et al. 2020). SDMs can be generated based on a spatial subdivision of presence records according to the genetic structure of taxa, and in a second step, the different predictions be merged together to the broader scale (Knowles et al. 2007, Marcer et al. 2016, Cacciapaglia and van Woesik 2017, Ikeda et al. 2017, Roberts et al. 2017, Pardo-Gandarillas et al. 2018). However, defining the genetic structure of benthic species in the SO is a long-term endeavour that requires a constantly renewed, sampling effort considering the extent and complexity of the study area (Moreau et al. 2017, Fraser et al. 2018, Moore et al. 2018). Awaiting for taxonomic revisions and enhanced sampling effort to best depict relationships between genetic units and environmental conditions (Vandersteen 2011), combining SDM with experimental data or mechanistic approaches can be an alternative to take into account the possible physiological contrasts between populations (Kearney and Porter 2009, Kearney et al. 2010, Buckley et al. 2010, Fordham et al. 2013, Briscoe et al. 2016, Feng and Papes 2017, Lopez/Guillaumot et al. in press.).

#### *4.2. Definition of the projection area*

The limitations in the current knowledge of species distribution also affect the quality of information available to estimate their potential distribution (Thuiller et al. 2003). When the limits of species environmental ranges are not fully captured, this uncertainty can significantly impact the accuracy of SDM predictions (Hortal et al. 2007, 2008, Rocchini et al. 2011, Sánchez-Fernández et al. 2011, Titeux et al. 2017, El-Gabbas and Dormann 2018). It reduces the

applicability of models for predictive purposes (Thuiller et al. 2004), induces model overfitting (Tsoar et al. 2007, Barve et al. 2011, Guillaumot et al. 2018b) and can lead to overestimating the extent of suitable areas (Anderson and Raza 2010).

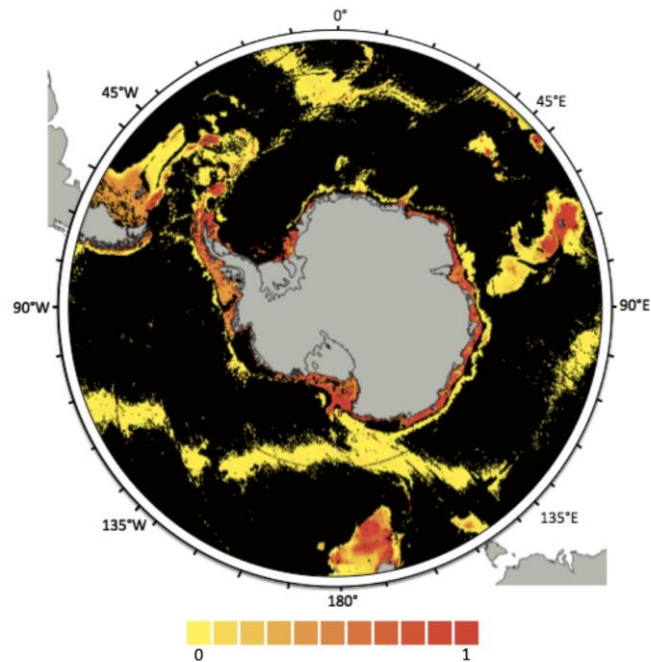
This bias can be partly overcome by reducing the extent of the projection area to the known distribution of occurrence records available (Anderson and Raza 2010), and by increasing knowledge on species ecology and physiology to identify the environmental conditions that are unsuitable for their survival or development (Byrne et al. 2016).

#### *4.3. Model extrapolation*

Models are said to extrapolate when a portion of the predicted area includes environmental conditions that are outside the range of values for which the model was calibrated. Model extrapolation may occur when model predictions are transferred, either in space or time. When extrapolated, model predictions are in non-analog conditions compared to initial, calibration conditions, because calibration data may not encompass the entire environmental range of each of the predictors (Guillaumot et al. 2020b). The set of projected environmental conditions can otherwise still be within the range of conditions, but specific combinations of environmental descriptors may become new, leading to extrapolation too (Mesgaran et al. 2014). In such conditions, predictions might be ecologically and statistically invalid and model interpretation inaccurate (Randin et al. 2006, Williams and Jackson 2007, Williams et al. 2007, Fitzpatrick and Hargrove 2009, Owens et al. 2013).

Among different approaches, Elith et al. (2010) propose to estimate and quantify model extrapolation using the Multivariate Environmental Similarity Surface (MESS) index to identify the most influential descriptors that lead to extrapolation. Grid-cell pixels for which at least one environmental descriptor has a value outside the range of environmental values defined by presence-only records (calibration range) is considered to be extrapolation. In these cases, MESS gets negative values and the ensemble of pixels containing negative values defines the extrapolation area (Elith et al. 2010, Guillaumot et al. 2020b). Most often, for SDMs performed at the scale of the SO, the number of records available to define the environmental space occupied by species is limiting, and the resolution of environmental descriptors relatively low (see section 2). As a consequence, SDM projections sometimes include wide extrapolation areas that may cover over 75% of the predicted regions (Fig. 6) (Guillaumot et al. 2020b).



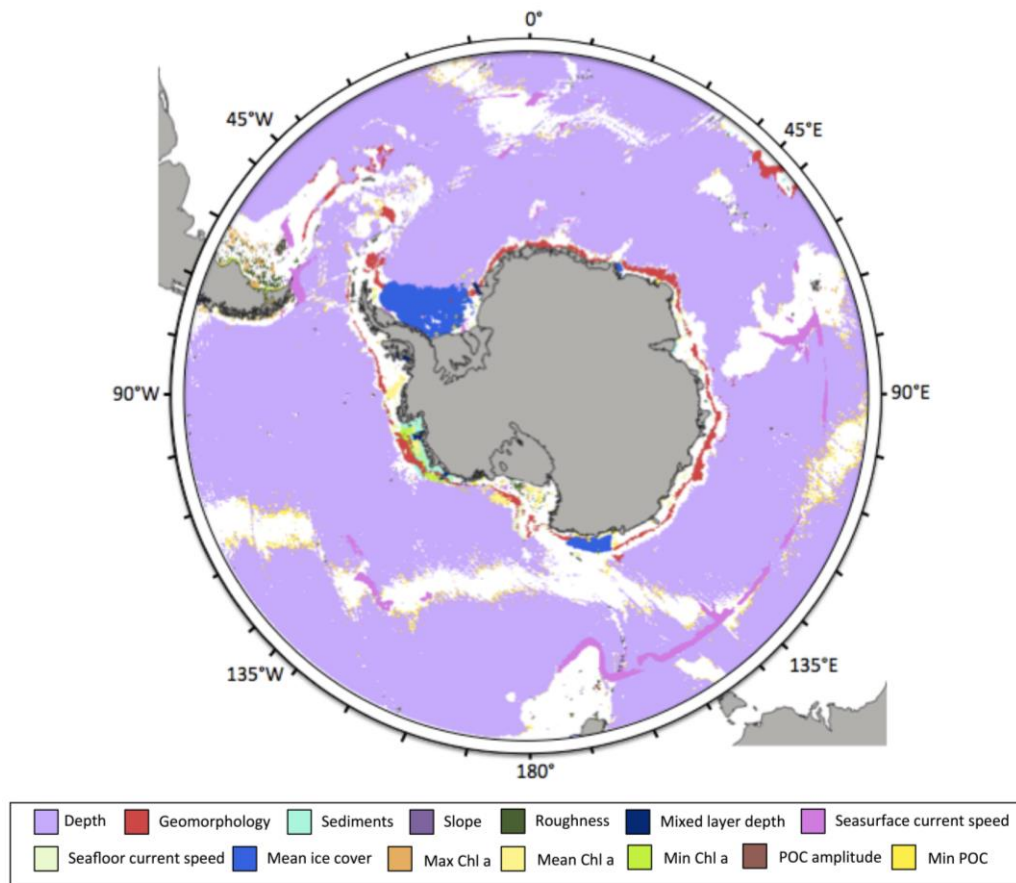


**Figure 6.** Extrapolation map of the SDM generated for the sea star *Acodontaster hodgsoni*, with all presence-only records available. Extrapolation corresponds here to the ensemble of environmental conditions that are outside of the boundaries of the calibration range. The extrapolation area is displayed in black and covers 78.6% of the entire projection area; coloured pixels (yellow-red color palette) show distribution probabilities (included between 0 and 1). Extracted from [Guillaumot et al. \(2020b\)](#).

In addition to quantifying the overall extrapolation area (Fig. 6), it is possible to fine-tune the analysis and define which environmental descriptors and areas are concerned with extrapolation ([Owens et al. 2013](#), [Guillaumot et al. 2021](#)) (Fig. 7). Such information could be used to resample the environmental descriptors implemented in the model.

In any case, it has been recommended to provide information on model extrapolation and more generally to other concepts of uncertainties (species detection, errors...) along with model predictions, because they are essential to accurate interpretation ([Beale and Lennon 2012](#), [Addison et al. 2013](#), [Guisan et al. 2013](#)).





**Figure 7.** Extrapolation map of the SDM generated for the sea star *Acodontaster hodgsoni* indicating environmental descriptors responsible for extrapolation (black pixels of Fig. 6 are here coloured according to the descriptor responsible for extrapolation; i.e. for each pixel, the predictor in question lies outside the calibration range). In this case study, 14 environmental descriptors are responsible for extrapolation, depth being the main contributor. White pixels correspond to areas where the model does not extrapolate (the corresponding model predictions are shown in Fig. 6). POC stands for 'Particulate Organic Carbon', Chla is the concentration in chlorophyll a in the seasurface. Generated from [Guillaumot et al. \(2021\)](#).

Limiting model projections to “realistic” depth ranges or some other environmental limiting factor based on a robust knowledge of species ecology (i.e. some expert-driven decision) was proved efficient to reduce extrapolation ([Kearney and Porter 2009](#), [Hare et al. 2012](#), [De Villiers et al. 2013](#), [Guillaumot et al. 2020b](#)). Such a strategy is transitional before complementary samples and more comprehensive occurrence datasets are made available to better define the species occupied space ([Guillaumot et al. 2020b](#)).

## 5. Model validation and accuracy of model predictions

### 5.1. Some usual metrics for the evaluation of model predictions

Once models are generated, the accuracy of predictions must be assessed to evaluate the validity of models with regards to scientific issues to address, compare different model outputs

and allow the formulation of reliable interpretations (Zurell et al. 2020). Several metrics were developed to evaluate the performance of models (Fielding and Bell 1997, Allouche et al. 2006). Most of them are based on the calculation of an error matrix (or confusion matrix) that displays the proportion of presence and absence records that are correctly predicted by the model (Allouche et al. 2006).

In most biological studies focused on the SO benthos, absence records are usually unavailable and SDMs are generated based on a set of presence/background records (see section 2.2). As a consequence, statistics that are commonly used for presence/absence datasets may not be appropriate for model evaluation (Wiley et al. 2003, Phillips et al. 2006, Braunish et al. 2013), such as the Kappa statistic (Allouche et al. 2006). In contrast, the AUC, the short for Area Under the Curve, or Area Under the Receiver Operating Curve, is one of the most used and appropriate metrics for measuring the performance of model predictions based on presence-background data (Hand 2009). The AUC is an objective measure that remains stable with low-prevalence datasets (i.e. low frequency of occurrences with regards to the projection space) and is not sensible to threshold effects (Manel et al. 2001, Hand 2009, van Proosdij et al. 2016). However, for presence/background models, specificity (the fraction of correctly predicted absences) might be overestimated when the number of background records is much higher than the number of presence-only records, or when background and presences are associated to very different environmental values. This incidentally inflates AUC scores (Phillips et al. 2006, Raes and ter Steege 2007, Lobo 2008, Jiménez-Valverde 2012) and invalidates the relevance of the AUC metrics (van Proosdij et al. 2016).

Even when properly employed, the AUC cannot be used to compare models when SDMs are generated for different species, based on different environmental descriptors, or projected on distinct regions, because values depend on the relative size of suitable areas, and prevalence scores may contrast (see section 3.3) (Wisz et al. 2008, Anderson and Gonzalez 2011). The AUC metrics must be used as a simple measure of the relative ranking of model predictions associated with a specific dataset (El-Gabbas and Dormann 2018). Overall, each statistics is characterised by specific advantages and potential biases, so that it is recommended to use several statistics for evaluating model predictions (Allouche et al. 2006).

The accuracy of model predictions can also be evaluated by testing the classification of independent test data, where the available occurrence dataset can be split into independent subsets to train or test the model (see Fielding and Bell 1997 for a review).

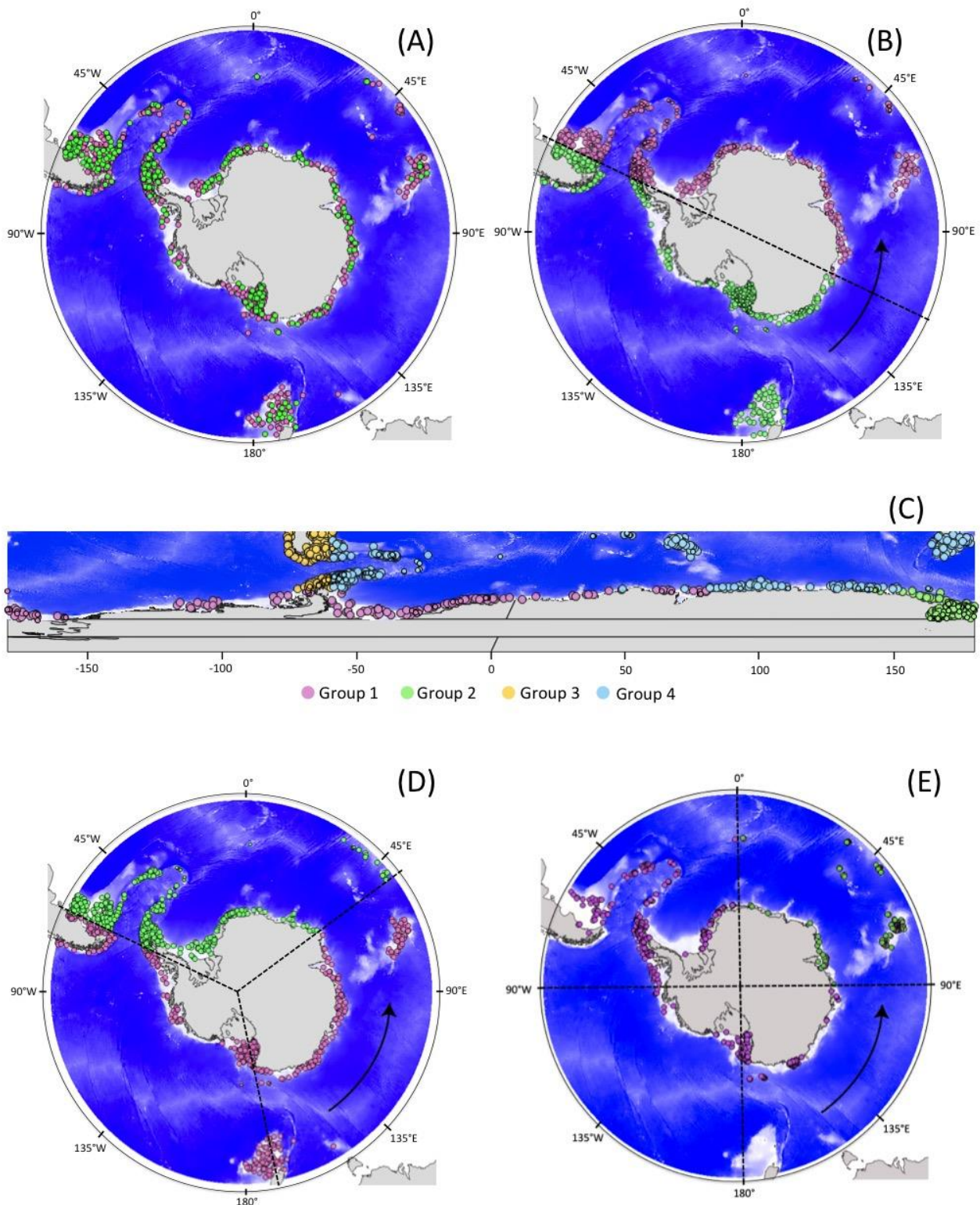
## *5.2. Cross-validation procedures*

Cross-validation procedures are aimed at evaluating model predictions using a subset of presence or absence records retrieved from the initial dataset used for modelling, in order to

assess how well test data match with modelled predictions (Bahn and McGill 2013). When working with presence-only datasets, two subsets of presence records are used: one subset is used to train the model (the training group), and the second subset is used to test the model (the test group). Test data and training data must be spatially independent from each other (Hijmans 2012, Bahn and McGill 2013). In most modelling exercises, standard cross-validation procedures are commonly used, in which the initial presence dataset is randomly split into a training and test subset. Frequently, as previously discussed, presence data are spatially aggregated in SO datasets and the necessary condition of independence between training and test data is seldom met, making the model accuracy evaluation over-optimistic (Telford and Birks 2009, Hijmans 2012, Radosavljevic and Anderson 2014). In contrast to random procedures, spatial cross-validation procedures improve the performance of the validation step by spatially segregating training and test subsets, ensuring the spatial independence between data even when they are spatially aggregated in initial datasets (Dhingra et al. 2016, Roberts et al. 2017, Guillaumot et al. 2019, see also

[http://cran.rapporter.net/web/packages/blockCV/vignettes/BlockCV\\_for\\_SDM.html](http://cran.rapporter.net/web/packages/blockCV/vignettes/BlockCV_for_SDM.html)).

Several spatial cross-validation procedures were proposed (Fig. 8), and the most appropriate one can be determined by comparing the different procedures, define the one that suits the most for the study (Muscarella et al. 2014, Radosavljevic and Anderson 2014, Valavi et al. 2018, Guillaumot et al. 2019), depending on the spatial scale of the analysis, on the number and spatial distribution of presence data, and on the selected algorithm (and its associated complexity) used for modelling (El-Gabbas and Dormann 2018, Hao et al. 2020).



**Figure 8.** Different cross-validation procedures based on the study of the sea star *Odontaster validus*, showing presence-only records and a random set of 1,000 background data selected according to a Kernel Density Estimation weighting scheme from the dataset of Griffiths et al. (2014) on sampling effort of the SO benthos. Data are split into training (pink) and test (green) subsets. The blue background corresponds to bathymetry and grey areas to emerged lands. (A) Random cross-validation procedure, with a random split into 75% training and 25% test data. (B) '2-fold CLOCK' clustering by random spatial



partition of the dataset into two groups (one training, one test). (C) 'BLOCK' splitting, generated according to median latitudinal and longitudinal values ([Muscarella et al. 2014](#)). After the generation of four groups (corresponding to the four colors), one group is randomly defined as the test subset, the other three groups as the training subsets. A different system of projection was used to represent this map to highlight the latitudinal and longitudinal definition of transects. (D) '3-fold CLOCK' clustering by random spatial partition of the dataset into three groups (2 training, 1 test). (E) '4-fold CLOCK' clustering by random spatial partition of the dataset into four groups (3 training, 1 test). Figure extracted from [Guillaumot et al. \(2019\)](#).

## 6. Conclusions and future prospects

This review summarizes some points and issues to be considered during SDM construction for modelling the distribution of SO species (Fig. 1). It shows that accurate and efficient SDMs can be produced for SO species when considering common potential biases and issues, and correcting for their side effects. Proposed corrections must be adjusted to each case study: no consensus method, nor implementation procedure always perform best, each case study requiring proper analyses to generate the most relevant and accurate predictions. This entails that, for each model, several procedures to implement the model should be tested to select the most suitable one, ideally giving priority to the availability of independent datasets to evaluate models. We showed that SDMs perform best when the species occupied space is accurately described, using extensive occurrence datasets with both presence and absence records, and when data are checked for positioning and georeferencing errors. A good knowledge of species' ecology, life history traits, and populational variations within the overall species distribution and environmental range help improve model quality ([Fois et al. 2018](#)). The compilation, examination, and preparation of datasets prior to modelling are essential steps to generate efficient models. Estimating and communicating uncertainties associated to model predictions is also an important task to be highlighted. This may include a "simple" interpretation of the ecological relevance of SDM outputs by experts ([Merow et al. 2017](#)) to the mapping of model extrapolation as illustrated here. Model uncertainties are part of model outputs and should not be omitted ([Guisan et al. 2013](#), [Grimm et al. 2014](#), [Grimm and Berger 2016](#)).

Remaining challenges to perform relevant SDMs for SO studies include more efforts in data collection outside of the main sampling hotspots and to fill knowledge gaps in SO species taxonomy. Some methodological perspectives, developed in other regions, address the integration of physiological information into SDMs. This completes the understanding of species environmental preferences and help better estimating species realised niche ([Kearney and Porter 2009](#), [Talluto et al. 2016](#), [Mathewson et al. 2017](#), [Rodríguez et al. 2019](#), [Gamliel et al. 2020](#)). Such studies have been recently developed for SO benthic species: in ([López-Farrán /](#)

[Guillaumot et al. in press](#)) the combination of physiological experimental results and SDM projections assessed the invading potential of the patagonian crab *Halicarcinus planatus* (Fabricius, 1775) on Antarctic coasts, as similarly done in [Byrne et al. \(2016\)](#) for the Arctic sea star *Asterias amurensis* Lutken, 1871. Hybrid modelling approaches constitute another exciting approach, where information of both SDM and physiological models are fully integrated, using the physiological information as a prior to inform the SDM ([Gamlie et al. 2020](#)). Recently applied to an endemic sea urchin of the Kerguelen Plateau ([Guillaumot et al. in prep.](#)), the method allows to more precisely predicting the effect of seasonal variations on species habitat suitability.

Other interesting methodological insights are the consideration of biotic interaction information, dispersal capacities estimates or population dynamics in complement to SDM predictions to generalize the understanding of the main drivers of species distribution ([Pellissier et al. 2010](#), [Meier et al. 2011](#), [Pagel and Schurr 2012](#), [Conlisk et al. 2013](#), [Pellissier et al. 2013](#), [Leach et al. 2016](#), [Anderson 2017](#)). This however necessitates a deep knowledge of the species ecology and of this surrounding environment, suggesting the first applications to be rather expected in local or regional scale studies.

A final take-home message is that model outputs should be interpreted carefully and model predictions always considered with a critical eye. Models are simple representations of complex systems and should be used in complement to other approaches to support conservation strategies or address fundamental research objectives ([Porfirio et al. 2014](#), [Kampichler and Sierdsema et al. 2017](#)).

## **Acknowledgements**

This work was supported by a “Fonds pour la formation à la Recherche dans l’Industrie et l’Agriculture” (FRIA) and “Bourse fondation de la mer” grants to C. Guillaumot.

This is contribution no. 47 to the vERSO project ([www.versoproject.be](http://www.versoproject.be)), funded by the Belgian Science Policy Office (BELSPO, contract n°BR/132/A1/vERSO). Research was also financed by the “Refugia and Ecosystem Tolerance in the Southern Ocean” project (RECTO; BR/154/A1/RECTO) funded by the Belgian Science Policy Office (BELSPO), this study being contribution number 24.

## References

- Addison, P.F., Rumpff, L., Bau, S.S., Carey, J.M., Chee, Y.E., Jarrad, F.C., ... & Burgman, M.A. (2013). Practical solutions for making models indispensable in conservation decision-making. *Diversity and Distributions*, 19(5-6), 490-502.
- Aguiar, L.M., da Rosa, R.O., Jones, G. & Machado, R.B. (2015). Effect of chronological addition of records to species distribution maps: The case of *Tonatia saurophila maresi* (Chiroptera, Phyllostomidae) in South America. *Austral Ecology*, 40(7), 836-844.
- Aguirre-Gutiérrez, J., Carvalheiro, L.G., Polce, C., van Loon, E.E., Raes, N., Reemer, M. & Biesmeijer, J.C. (2013). Fit-for-purpose: species distribution model performance depends on evaluation criteria—Dutch hoverflies as a case study. *PloS One*, 8(5), e63708.
- Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., & Anderson, R. P. (2015). spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, 38(5), 541-545.
- Allan, M.G. (2014). Remote sensing, numerical modelling and ground truthing for analysis of lake water quality and temperature (Doctoral dissertation, University of Waikato).
- Allouche, O., Tsoar, A. & Kadmon, R. (2006). Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of applied Ecology*, 43(6), 1223-1232.
- Anderson, R.P., & Raza, A. (2010). The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus *Nephelomys*) in Venezuela. *Journal of Biogeography*, 37(7), 1378-1393.
- Anderson, R.P. & Gonzalez Jr, I. (2011). Species-specific tuning increases robustness to sampling bias in models of species distributions: an implementation with Maxent. *Ecological Modelling*, 222(15), 2796-2811.
- Anderson, R.P. (2017). When and how should biotic interactions be considered in models of species niches and distributions? *Journal of Biogeography*, 44(1), 8-17.



Araújo, M.B., & Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, 33(10), 1677-1688.

Araújo, M.B., & New, M. (2007). Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*, 22(1), 42-47.

Araújo, M.B., Anderson, R.P., Barbosa, A.M., Beale, C.M., Dormann, C.F., Early, R. ... & O'Hara, R.B. (2019). Standards for distribution models in biodiversity assessments. *Science Advances*, 5(1), eaat4858.

Arthur, B., Hindell, M., Bester, M., De Bruyn, P.N., Trathan, P., Goebel, M. & Lea, M.A. (2017). Winter habitat predictions of a key Southern Ocean predator, the Antarctic fur seal (*Arctocephalus gazella*). *Deep Sea Research Part II: Topical Studies in Oceanography*, 140, 171-181.

Ashton, G.V., Morley, S.A., Barnes, D.K., Clark, M.S. & Peck, L.S. (2017). Warming by 1 C drives species and assemblage level responses in Antarctica's marine shallows. *Current Biology*, 27(17), 2698-2705.

Assis, J., Tyberghein, L., Bosch, S., Verbruggen, H., Serrão, E.A. & De Clerck, O. (2018). BioORACLE v2. 0: Extending marine data layers for bioclimatic modelling. *Global Ecology and Biogeography*, 27(3), 277-284.

Aubry, K.B., Raley, C.M., & McKelvey, K.S. (2017). The importance of data quality for generating reliable distribution models for rare, elusive, and cryptic species. *PloS one*, 12(6), e0179152.

Austin, M.P. (2002). Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, 157(2-3), 101-118.

Austin, M.P., & Van Niel, K.P. (2011). Improving species distribution models for climate change studies: variable selection and scale. *Journal of Biogeography*, 38(1), 1-8.

Bahn, V. & McGill, B.J. (2007). Can niche-based distribution models outperform spatial interpolation? *Global Ecology and Biogeography*, 16(6), 733-742.

Bahn, V., & McGill, B.J. (2013). Testing the predictive performance of distribution models. *Oikos*, 122(3), 321-331.

Barbet-Massin, M., Jiguet, F., Albert, C.H. & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, 3(2), 327-338.

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S.P., Peterson, A.T. ... & Villalobos, F. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, 222(11), 1810-1819.

Basher, Z. & Costello, M.J. (2016). The past, present and future distribution of a deep-sea shrimp in the Southern Ocean. *PeerJ*, 4, e1713.

Beale, C.M. & Lennon, J.J. (2012). Incorporating uncertainty in predictive species distribution modelling. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1586), 247-258.

Bean, W.T., Stafford, R. & Brashares, J.S. (2012). The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. *Ecography*, 35(3), 250-258.

Beaumont, L.J., Hughes, L. & Pitman, A.J. (2008). Why is the choice of future climate scenarios for species distribution modelling important? *Ecology Letters*, 11(11), 1135-1146.

Beaumont, L.J., Graham, E., Duursma, D.E., Wilson, P.D., Cabrelli, A., Baumgartner, J.B. ... & Laffan, S.W. (2016). Which species distribution models are more (or less) likely to project broad-scale, climate-induced shifts in species ranges? *Ecological Modelling*, 342, 135-146.

Bertrand, A., Habasque, J., Hattab, T., Hintzen, N.T., Oliveros-Ramos, R., Gutiérrez, M. ... & Gerlotto, F. (2016). 3-D habitat suitability of jack mackerel *Trachurus murphyi* in the Southeastern Pacific, a comprehensive study. *Progress in Oceanography*, 146, 199-211.

biodiversity.aq (2020) Antarctic Biodiversity Portal. Scientific Committee for Antarctic Research. [www.biodiversity.aq](http://www.biodiversity.aq)

Bloom, T.D., Flower, A. & DeChaine, E.G. (2018). Why georeferencing matters: introducing a practical protocol to prepare species occurrence records for spatial analysis. *Ecology and Evolution*, 8(1), 765-777.

Boria, R.A., Olson, L.E., Goodman, S.M. & Anderson, R.P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, 275, 73-77.

Boulanger, Y., Parisien, M.A. & Wang, X. (2018). Model-specification uncertainty in future area burned by wildfires in Canada. *International Journal of wildland fire*, 27(3), 164-175.

Bradie, J. & Leung, B. (2017). A quantitative synthesis of the importance of variables used in MaxEnt species distribution models. *Journal of Biogeography*, 44(6), 1344-1361.

Braunisch, V., Coppes, J., Arlettaz, R., Suchant, R., Schmid, H. & Bollmann, K. (2013). Selecting from correlated climate variables: a major source of uncertainty for predicting species distributions under climate change. *Ecography*, 36(9), 971-983.

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984). Classification and regression trees. Wadsworth International Group, Belmont, CA.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

Breiner, F.T., Guisan, A., Bergamini, A. & Nobis, M.P. (2015). Overcoming limitations of modelling rare species by using ensembles of small models. *Methods in Ecology and Evolution*, 6(10), 1210-1218.

Breiner, F.T., Nobis, M.P., Bergamini, A. & Guisan, A. (2018). Optimizing ensembles of small models for predicting the distribution of species with few occurrences. *Methods in Ecology and Evolution*, 9(4), 802-808.

Breitzke, M. (2014). Overview of seismic research activities in the Southern Ocean-quantifying the environmental impact. *Antarctic Science*, 26(1), 80.

Briscoe, N.J., Kearney, M.R., Taylor, C.A. & Wintle, B.A. (2016). Unpacking the mechanisms captured by a correlative species distribution model to improve predictions of climate refugia. *Global Change Biology*, 22(7), 2425-2439.

Brotans, L., Thuiller, W., Araújo, M.B. & Hirzel, A.H. (2004). Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, 27(4), 437-448.

Buckley, L.B., Urban, M.C., Angilletta, M.J., Crozier, L.G., Rissler, L.J. & Sears, M.W. (2010). Can mechanism inform species' distribution models?. *Ecology Letters*, 13(8), 1041-1054.

Bucklin, D.N., Basille, M., Benschoter, A.M., Brandt, L.A., Mazzotti, F.J., Romanach, S.S. ... & Watling, J.I. (2015). Comparing species distribution models constructed with different subsets of environmental predictors. *Diversity and Distributions*, 21(1), 23-35.

Buisson, L., Thuiller, W., Casajus, N., Lek, S. & Grenouillet, G. (2010). Uncertainty in ensemble forecasting of species distribution. *Global Change Biology*, 16(4), 1145-1157.

Byrne, M., Gall, M., Wolfe, K., & Agüera, A. (2016). From pole to pole: the potential for the Arctic seastar *Asterias amurensis* to invade a warming Southern Ocean. *Global Change Biology*, 22(12), 3874-3887.

Cacciapaglia, C. & van Woesik, R. (2017). Marine species distribution modelling and the effects of genetic isolation under climate change. *Journal of Biogeography*, 45(1), 154-163.

Cavanagh, R.D., Murphy, E.J., Bracegirdle, T.J., Turner, J., Knowland, C.A., Corney, S.P. ... & Constable, A.J. (2017). A synergistic approach for evaluating climate model output for ecological applications. *Frontiers in Marine Science*, 4, 308.

Carvalho, B.M., Rangel, E.F., Ready, P.D. & Vale, M.M. (2015). Ecological niche modelling predicts southward expansion of *Lutzomyia (Nyssomyia) flaviscutellata* (Diptera: Psychodidae: Phlebotominae), vector of *Leishmania (Leishmania) amazonensis* in South America, under climate change. *PLoS One*, 10(11), e0143282.

Chefaoui, R.M. & Lobo, J.M. (2008). Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological Modelling*, 210(4), 478-486.

Chevin, L.M., Lande, R. & Mace, G.M. (2010). Adaptation, plasticity, and extinction in a changing environment: towards a predictive theory. *PLoS Biology*, 8(4).

Clark, M.S., Nieva, L.V., Hoffman, J.I., Davies, A.J., Trivedi, U.H., Turner, F., ... & Peck, L.S. (2019). Lack of long-term acclimation in Antarctic encrusting species suggests vulnerability to warming. *Nature Communications*, 10(1), 1-10.

Clarke, A., Griffiths, H.J., Linse, K., Barnes, D.K. & Crame, J.A. (2007). How well do we know the Antarctic marine fauna? A preliminary study of macroecological and biogeographical patterns in Southern Ocean gastropod and bivalve molluscs. *Diversity and Distributions*, 13(5), 620-632.

Conlisk, E., Syphard, A.D., Franklin, J., Flint, L., Flint, A. & Regan, H. (2013). Uncertainty in assessing the impacts of global change with coupled dynamic species distribution and population models. *Global Change Biology*, 19(3), 858-869.

Connor, T., Hull, V., Viña, A., Shortridge, A., Tang, Y., Zhang, J. ... & Liu, J. (2018). Effects of grain size and niche breadth on species distribution modeling. *Ecography*, 41(8), 1270-1282.

Costa, H., Foody, G.M., Jiménez, S. & Silva, L. (2015). Impacts of species misidentification on species distribution modeling with presence-only data. *ISPRS International Journal of Geo-Information*, 4(4), 2496-2518.

Cruse, B., Liedloff, A.C. & Wintle, B.A. (2012). A new method for dealing with residual spatial autocorrelation in species distribution models. *Ecography*, 35(10), 879-888.

Crimmins, S.M., Dobrowski, S.Z. & Mynsberge, A.R. (2013). Evaluating ensemble forecasts of plant species distributions under climate change. *Ecological Modelling*, 266, 126-130.

Cryer, M. (2015). Progress on predicting the distribution of Vulnerable Marine Ecosystems and options for designing spatial management areas for bottom fisheries within the SPRFMO Convention Area. In *3rd Meeting of the SPRFMO Scientific Committee*.

Currie, D.J. (2007). Disentangling the roles of environment and space in ecology. *Journal of Biogeography*, 34(12), 2009-2011.

Dhingra, M.S., Artois, J., Robinson, T.P., Linard, C., Chaiban, C., Xenarios, I. ... & Von Dobschuetz, S. (2016). Global mapping of highly pathogenic avian influenza H5N1 and H5Nx clade 2.3. 4.4 viruses with spatial cross-validation. *Elife*, 5, e19571.

De'ath, G. & Fabricius, K.E. (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, 81, 3178–3192.

De Broyer, C. & Danis, B. (2011). How many species in the Southern Ocean? Towards a dynamic inventory of the Antarctic marine species. *Deep sea research Part II: Topical studies in oceanography*, 58(1-2), 5-17.

De Broyer, C., Koubbi, P., Griffiths, H.J., Raymond, B., d'Udekem d'Acoz, C., Van de Putte, A.P. ... Ropert-Coudert, Y. (2014). *Biogeographic atlas of the Southern Ocean* (p. 498). C. De Broyer & P. Koubbi (Eds.). Cambridge: Scientific Committee on Antarctic Research.

de la Hoz, C.F., Ramos, E., Acevedo, A., Puente, A., Losada, Í.J. & Juanes, J.A. (2018). OCLE: A European open access database on climate change effects on littoral and oceanic ecosystems. *Progress in Oceanography*, 168, 222-231.

de la Hoz, C.F., Ramos, E., Puente, A. & Juanes, J.A. (2019). Temporal transferability of marine distribution models: The role of algorithm selection. *Ecological Indicators*, 106, 105499.

De Villiers, M., Hattingh, V. & Kriticos, D.J. (2013). Combining field phenological observations with distribution data to model the potential distribution of the fruit fly *Ceratitis rosa* Karsch (Diptera: Tephritidae). *Bulletin of Entomological Research*, 103(1), 60.

Deleersnijder, E., Van Ypersele, J. P., & Campin, J. M. (1993). An orthogonal curvilinear coordinate system for a world ocean model. *Ocean Modelling (SCOR WG Newsletter)*.

Diniz-Filho, J.A., Bini, L.M. & Hawkins, B.A. (2003). Spatial autocorrelation and red herrings in geographical ecology. *Global Ecology and Biogeography*, 12(1), 53-64.

Diniz-Filho, J.A., Mauricio Bini, L., Fernando Rangel, T., Loyola, R.D., Hof, C., Nogués-Bravo, D. & Araújo, M.B. (2009). Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. *Ecography*, 32(6), 897-906.

Dixon, G.B., Davies, S.W., Aglyamova, G.V., Meyer, E., Bay, L.K. & Matz, M.V. (2015). Genomic determinants of coral heat tolerance across latitudes. *Science*, 348(6242), 1460-1462.

Dormann, C.F. (2007). Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology and Biogeography*, 16(2), 129-138.

Dormann, C.F., Purschke, O., Márquez, J.R., Lautenbach, S. & Schröder, B. (2008). Components of uncertainty in species distribution analysis: a case study of the great grey shrike. *Ecology*, 89(12), 3371-3386.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G. ... & Münkemüller, T. (2012). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1), 27-46.

Eby, M., & Holloway, G. (1994). Grid transformation for incorporating the Arctic in a global ocean model. *Climate Dynamics*, 10(4-5), 241-247.

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

Elith, J., Leathwick, J.R. & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4), 802-813.

Elith, J. & Graham, C.H. (2009). Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. *Ecography*, 32(1), 66-77.

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

Elith, J., Kearney, M. & Phillips, S. (2010). The art of modelling range-shifting species. *Methods in Ecology and Evolution*, 1(4), 330-342.

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

El-Gabbas, A. & Dormann, C.F. (2018). Wrong, but useful: regional species distribution models may not be improved by range-wide data under biased sampling. *Ecology and Evolution*, 8(4), 2196-2206.



- El Mahrad, B., Newton, A., Icely, J.D., Kacimi, I., Abalansa, S. & Snoussi, M. (2020). Contribution of remote sensing technologies to a holistic coastal and marine environmental management framework: A review. *Remote Sensing*, 12(14), 2313.
- Ensing, D.J., Moffat, C.E. & Pither, J. (2012). Taxonomic identification errors generate misleading ecological niche model predictions of an invasive hawkweed. *Botany*, 91(3), 137-147.
- Fabri-Ruiz, S., Saucède, T., Danis, B. & David, B. (2017a). Southern Ocean Echinoids database—An updated version of Antarctic, Sub-Antarctic and cold temperate echinoid database. *ZooKeys*, (697), 1.
- Fabri-Ruiz, S., Danis, B., David, B. & Saucedo, T. (2017b) Environmental data of the Southern Ocean, 1955-2012, Ver. 1, *Australian Antarctic Data Centre* - [doi:10.4225/15/5a30800470009](https://doi.org/10.4225/15/5a30800470009), Accessed: 2020-10-26
- Fabri-Ruiz, S. (2018). *Modèles de distribution et changements environnementaux: Application aux faunes d'échinides de l'océan Austral et écorégionalisation* (Doctoral dissertation).
- Fabri-Ruiz, S., Danis, B., David, B. & Saucède, T. (2019). Can we generate robust species distribution models at the scale of the Southern Ocean? *Diversity and Distributions*, 25(1), 21-37.
- Feng, X. & Papeş, M. (2017). Can incomplete knowledge of species' physiology facilitate ecological niche modelling? A case study with virtual species. *Diversity and Distributions*, 23(10), 1157-1168.
- Fielding, A.H. & Bell, J.F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24(1), 38-49.
- Fitzpatrick, M.C. & Hargrove, W.W. (2009). The projection of species distribution models and the problem of non-analog climate. *Biodiversity and Conservation*, 18(8), 2255.
- Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S.C., Collins, W. ... & Forest, C. (2014). Evaluation of climate models. In *Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 741-866). Cambridge University Press.

Fois, M., Cuenca-Lombraña, A., Fenu G. & Bacchetta, G. (2018). Using species distribution models at local scale to guide the search of poorly known species: Review, methodological issues and future directions. *Ecological Modelling*, 385, 124-132.

Fordham, D.A., Mellin, C., Russell, B.D., Akçakaya, R.H., Bradshaw, C.J., Aiello-Lammens, M.E., ... & Brook, B.W. (2013). Population dynamics can be more important than physiological limits for determining range shifts under climate change. *Global Change Biology*, 19(10), 3224-3237.

Fourcade, Y., Engler, J.O., Rödder, D. & Secondi, J. (2014). Mapping species distributions with MAXENT using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. *PloS One*, 9(5).

Franklin, J. (2009). *Mapping species distributions: spatial inference and prediction*. Cambridge University Press.

Franklin, J. (2010). Moving beyond static species distribution models in support of conservation biogeography. *Diversity and Distributions*, 16(3), 321-330.

Fraser, C.I., Morrison, A.K., Hogg, A.M., Macaya, E.C., van Sebille, E., Ryan, P.G., ... & Waters, J.M. (2018). Antarctica's ecological isolation will be broken by storm-driven dispersal and warming. *Nature Climate Change*, 8(8), 704.

Freer, J.J., Partridge, J.C., Tarling, G.A., Collins, M.A. & Genner, M.J. (2018). Predicting ecological responses in a changing ocean: the effects of future climate uncertainty. *Marine Biology*, 165(1), 7.

Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189-1232.

Frölicher, T.L., Rodgers, K.B., Stock, C.A. & Cheung, W.W. (2016). Sources of uncertainties in 21st century projections of potential ocean ecosystem stressors. *Global Biogeochemical Cycles*, 30(8), 1224-1243.

- Gage, J.D. (2004). Diversity in deep-sea benthic macrofauna: the importance of local ecology, the larger scale, history and the Antarctic. *Deep Sea Research Part II: Topical Studies in Oceanography*, 51(14-16), 1689-1708.
- Galante, P.J., Alade, B., Muscarella, R., Jansa, S.A., Goodman, S.M. & Anderson, R.P. (2018). The challenge of modeling niches and distributions for data-poor species: a comprehensive approach to model complexity. *Ecography*, 41(5), 726-736.
- Gall, J. (1885). Use of cylindrical projections for geographical astronomical, and scientific purposes. *Scottish Geographical Magazine*, 1(4), 119-123.
- Gallego, R., Dennis, T.E., Basher, Z., Lavery, S. & Sewell, M.A. (2017). On the need to consider multiphasic sensitivity of marine organisms to climate change: A case study of the Antarctic acorn barnacle. *Journal of Biogeography*, 44, 2165–2175.
- Gamliel, I., Buba, Y., Guy-Haim, T., Garval, T., Willette, D., Rilov, G. & Belmaker, J. (2020). Incorporating physiology into species distribution models moderates the projected impact of warming on selected Mediterranean marine species. *Ecography*, 43(7), 1090-1106.
- García-Callejas, D. & Araújo, M.B. (2016). The effects of model and data complexity on predictions from species distributions models. *Ecological Modelling*, 326, 4-12.
- González-Wevar, C.A., Hüne, M., Rosenfeld, S., Nakano, T., Saucède, T., Spencer, H. & Poulin, E. (2019). Systematic revision of *Nacella* (Patellogastropoda: Nacellidae) based on a complete phylogeny of the genus, with the description of a new species from the southern tip of South America. *Zoological Journal of the Linnean Society*, 186(2), 303-336.
- Gotelli, N.J. & Stanton-Geddes, J. (2015). Climate change, genetic markers and species distribution modelling. *Journal of Biogeography*, 42(9), 1577-1585.
- Graham, C.H., Elith, J., Hijmans, R.J., Guisan, A., Townsend Peterson, A., Loiselle, B.A. & NCEAS Predicting Species Distributions Working Group. (2008). The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology*, 45(1), 239-247.
- Griffiths, H.J., Van de Putte, A.P. & Danis, B. (2014). CHAPTER 2.2. Data distribution: Patterns and implications. In: De Broyer, C., Koubbi, P., Griffiths, H.J., Raymond, B., Udekem d'Acoz, C.

d', et al. (eds.). Biogeographic Atlas of the Southern Ocean. Scientific Committee on Antarctic Research, Cambridge, pp. 16-26.

Grimm, V., Augusiak, J., Focks, A., Frank, B.M., Gabsi, F., Johnston, A.S. ... & Thorbek, P. (2014). Towards better modelling and decision support: documenting model development, testing, and analysis using TRACE. *Ecological Modelling*, 280, 129-139.

Grimm, V. & Berger, U. (2016). Robustness analysis: Deconstructing computational models for ecological theory and applications. *Ecological Modelling*, 326, 162-167.

Guillaumot, C., Martin, A., Fabri-Ruiz, S., Eléaume, M. & Saucède, T. (2016). Echinoids of the Kerguelen Plateau—occurrence data and environmental setting for past, present, and future species distribution modelling. *ZooKeys*, (630), 1.

Guillaumot, C., Martin, A., Eléaume, M. & Saucède, T. (2018a). Methods for improving species distribution models in data-poor areas: example of sub-Antarctic benthic species on the Kerguelen Plateau. *Marine Ecology Progress Series*, 594, 149-164.

Guillaumot, C., Fabri-Ruiz, S., Martin, A., Eléaume, M., Danis, B., Féral, J.P. & Saucède, T. (2018b). Benthic species of the Kerguelen Plateau show contrasting distribution shifts in response to environmental changes. *Ecology and Evolution*, 8(12), 6210-6225.

Guillaumot, C., Raymond, B. & Danis, B. (2018c) Marine environmental data layers for Southern Ocean species distribution modelling, Ver. 1, *Australian Antarctic Data Centre* - [doi:10.26179/5b8f30e30d4f3](https://doi.org/10.26179/5b8f30e30d4f3)

Guillaumot, C., Artois, J., Saucède, T., Demoustier, L., Moreau, C., Eléaume, M. ... & Danis, B. (2019). Broad-scale species distribution models applied to data-poor areas. *Progress in Oceanography*, 175, 198-207.

Guillaumot, C., Danis, B. & Saucède, T. (2020a). Selecting environmental descriptors is critical for modelling the distribution of Antarctic benthic species. *Polar Biology*, 1-19.

Guillaumot C., Moreau C., Danis B. & Saucède T. (2020b). Extrapolation in species distribution modelling. Application to Southern Ocean marine species. *Progress in Oceanography*, 188, 102438

Guillaumot, C., Martin, A., Eléaume, M., Danis, B. & Saucède, T. (2021). SDMPlay. Species Distribution Modelling Playground. <https://CRAN.R-project.org/package=SDMPlay>

Guillera-Aroita, G. (2016). Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. *Ecography*, 40(2), 281-295.

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

Guisan, A., Graham, C. H., Elith, J., Huettmann, F. & NCEAS Species Distribution Modelling Group. (2007a). Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions*, 13(3), 332-340.

Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I. ... & Martin, T.G. (2013). Predicting species distributions for conservation decisions. *Ecology Letters*, 16(12), 1424-1435.

Guisan, A., Thuiller, W. & Zimmermann, N.E. (2017). *Habitat suitability and distribution models: with applications in R*. Cambridge University Press.

Gutt, J., Zurell, D., Bracegirdle, T., Cheung, W., Clark, M., Convey, P. ... & Griffiths, H. (2012). Correlative and dynamic species distribution modelling for ecological predictions in the Antarctic: a cross-disciplinary concept. *Polar Research*, 31(1), 11091.

Gutt, J., Isla, E., Bertler, A.N., Bodeker, G.E., Bracegirdle, T., Cavanagh, R.D., ... Xavier, J.C. (2017). Cross-disciplinarity in the advance of Antarctic ecosystem research. *Marine Genomics*, 37, 1–17.

Halvorsen, R., Mazzoni, S., Dirksen, J.W., Næsset, E., Gobakken, T. & Ohlson, M. (2016). How important are choice of model selection method and spatial autocorrelation of presence data for distribution modelling by MaxEnt? *Ecological Modelling*, 328, 108-118.

Hand, D.J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1), 103-123.

Hansen, L.K. & Salamon, P. (1990). Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10), 993-1001.

Hao, T., Elith, J., Guillera-Aroita, G. & Lahoz-Monfort, J.J. (2019). A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD. *Diversity and Distributions*, 25(5), 839-852.

Hao, T., Elith, J., Lahoz-Monfort, J.J. & Guillera-Aroita, G. (2020). Testing whether ensemble modelling is advantageous for maximising predictive performance of species distribution models. *Ecography*, 43(4), 549-558.

Hare, J.A., Wuenschel, M.J. & Kimball, M.E. (2012). Projecting range limits with coupled thermal tolerance-climate change models: an example based on gray snapper (*Lutjanus griseus*) along the US east coast. *PLoS One*, 7(12).

Hayhoe, K., Edmonds, J., Kopp, R., LeGrande, A., Sanderson, B., Wehner, M. & Wuebbles, D. (2017). Climate models, scenarios, and projections. In *Climate Science Special Report: A Sustained Assessment Activity of the U.S. Global Change Research Program* [Wuebbles, D.J., D.W. Fahey, K.A. Hibbard, D.J. Dokken, B.C. Stewart, and T.K. Maycock (eds.)]. U.S. Global Change Research Program, Washington, DC, USA (2017), pp. 186-227.

Hengl, T., Sierdsema, H., Radović, A. & Dilo, A. (2009). Spatial prediction of species' distributions from occurrence-only records: combining point pattern analysis, ENFA and regression-kriging. *Ecological Modelling*, 220(24), 3499-3511.

Henson, S.A., Yool, A. & Sanders, R. (2015). Variability in efficiency of particulate organic carbon export: A model study. *Global Biogeochemical Cycles*, 29(1), 33-45.

Hernandez, P.A., Graham, C.H., Master, L.L., & Albert, D.L. (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, 29(5), 773-785.

Hibberd, T. (2016). *Describing and predicting the spatial distribution of benthic biodiversity in the sub-Antarctic and Antarctic* (Doctoral dissertation, University of Tasmania).

Hijmans, R.J. (2012). Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. *Ecology*, 93(3), 679-688.

Hindell, M.A., Reisinger, R.R., Robert-Coudert, Y., Hückstädt, L.A., Trathan, P.N., Bornemann, H., ... & Lea, M.A. (2020). Tracking of marine predators to protect Southern Ocean ecosystems. *Nature*, 580(7801), 87-92.

Hortal, J., Lobo, J.M. & Jiménez-Valverde, A. (2007). Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife, Canary Islands. *Conservation Biology*, 21(3), 853-863.

Hortal, J., Jiménez-Valverde, A., Gómez, J.F., Lobo, J.M. & Baselga, A. (2008). Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos*, 117(6), 847-858.

Hutchinson, G.E. (1957). Concluding remarks Cold Spring Harbor Symposia on Quantitative Biology, 22: 415–427. *GS SEARCH*.

IPCC (2000). IPCC Special Report. Emissions Scenarios, summary for Policymakers, 27pp. <https://www.ipcc.ch/site/assets/uploads/2018/03/sres-en.pdf>

Ikeda, D.H., Max, T.L., Allan, G.J., Lau, M.K., Shuster, S.M. & Whitham, T.G. (2017). Genetically informed ecological niche models improve climate change predictions. *Global Change Biology*, 23(1), 164-176.

Iturbide, M., Bedia, J., Herrera, S., del Hierro, O., Pinto, M. & Gutiérrez, J.M. (2015). A framework for species distribution modelling with improved pseudo-absence generation. *Ecological Modelling*, 312, 166-174.

Iturbide, M., Bedia, J. & Gutiérrez, J.M. (2018). Background sampling and transferability of species distribution model ensembles under climate change. *Global and Planetary Change*, 166, 19-29.

Jarnevich, C.S., Stohlgren, T.J., Kumar, S., Morissette, J.T. & Holcombe, T.R. (2015). Caveats for correlative species distribution modeling. *Ecological Informatics*, 29, 6-15.

Jarnevich, C.S., Talbert, M., Morissette, J., Aldridge, C., Brown, C.S., Kumar, S. ... & Holcombe, T. (2017). Minimizing effects of methodological decisions on interpretation and prediction in species distribution studies: An example with background selection. *Ecological Modelling*, 363, 48-56.



Jerosch, K., Scharf, F.K., Deregibus, D., Campana, G.L., Zacher, K., Pehlke, H. ... & Abele, D. (2019). Ensemble modelling of Antarctic macroalgal habitats exposed to glacial melt in a polar fjord. *Frontiers in Ecology and Evolution*, 7, 207.

Jiménez-Valverde, A. & Lobo, J.M. (2007). Threshold criteria for conversion of probability of species presence to either–or presence–absence. *Acta oecologica*, 31(3), 361-369.

Jiménez-Valverde, A. (2012). Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography*, 21(4), 498-507.

Jiménez-Valverde, A., Rodríguez-Rey, M. & Peña-Aguilera, P. (2020). Climate data source matters in species distribution modelling: the case of the Iberian Peninsula. *Biodiversity and Conservation*, 1-18.

Kaiser, S., Brandão, S.N., Brix, S., Barnes, D.K.A., Bowden, D.A., Ingels, J. ...& Yasuhara, M. (2013). Patterns, processes and vulnerability of Southern Ocean benthos: A decadal leap in knowledge and understanding. *Marine Biology*, 160, 2295–2317.

Kamgar-Parsi, B. & Sander, W.A. (1989). Quantization error in spatial sampling: comparison between square and hexagonal pixels. In *Proceedings CVPR'89: IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 604-611). IEEE.

Kampichler, C. & Sierdsema, H. (2018). On the usefulness of prediction intervals for local species distribution model forecasts. *Ecological Informatics*, 47, 67-72.

Kearney, M. & Porter, W. (2009). Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters*, 12(4), 334-350.

Kearney, M.R., Wintle, B.A. & Porter, W.P. (2010). Correlative and mechanistic models of species distribution provide congruent forecasts under climate change. *Conservation Letters*, 3(3), 203-213.

Kennicutt, M.C., Chown, S.L., Cassano, J.J., Liggett, D., Massom, R., Peck, L.S., ... & Sutherland, W.J. (2014). Six priorities for Antarctic science. *Nature*, 512, 23–25.

Kindt, R. (2017). R Package "BiodiversityR": Package for Community Ecology and Suitability Analysis. Version 2.8-0. <https://cran.r-project.org/web/packages/BiodiversityR/index.html>

Knowles, L.L., Carstens, B.C. & Keat, M.L. (2007). Coupling genetic and ecological-niche models to examine how past population distributions contribute to divergence. *Current Biology*, 17(11), 940-946.

Knutti, R. (2010). The end of model democracy? *Climatic Change*, 102(3-4), 395-404.

Kramer-Schadt, S., Niedballa, J., Pilgrim, J.D., Schröder, B., Lindenborn, J., Reinfelder, V. ... & Cheyne, S.M. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, 19(11), 1366-1379.

Kühn, I. (2007). Incorporating spatial autocorrelation may invert observed patterns. *Diversity and Distributions*, 13(1), 66-69.

Lahoz-Monfort, J.J., Guillera-Aroita, G. & Wintle, B.A. (2014). Imperfect detection impacts the performance of species distribution models. *Global Ecology and Biogeography*, 23(4), 504-515.

Lauzeral, C., Grenouillet, G. & Brosse, S. (2013). Spatial range shape drives the grain size effects in species distribution models. *Ecography*, 36(7), 778-787.

Leach, K., Montgomery, W.I. & Reid, N. (2016). Modelling the influence of biotic factors on species distribution patterns. *Ecological Modelling*, 337, 96-106.

Lieske, D.J., Schmid, M.S. & Mahoney, M. (2018). Ensembles of Ensembles: Combining the Predictions from Multiple Machine Learning Methods. In *Machine Learning for Ecology and Sustainable Natural Resource Management* (pp. 109-121). Springer, Cham.

Liu, C., Newell, G. & White, M. (2019). The effect of sample size on the accuracy of species distribution models: considering both presences and pseudo-absences or background sites. *Ecography*, 42(3), 535-548.

Lobo, J.M. (2008). More complex distribution models or more representative data?. *Biodiversity Informatics*, 5.

Lobo, J.M., Jiménez-Valverde, A. & Hortal, J. (2010). The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, 33(1), 103-114.

Lomba, A., Pellissier, L., Randin, C., Vicente, J., Moreira, F., Honrado, J. & Guisan, A. (2010). Overcoming the rare species modelling paradox: a novel hierarchical framework applied to an Iberian endemic plant. *Biological Conservation*, 143(11), 2647-2657.

Loots, C., Koubbi, P. & Duhamel, G. (2007). Habitat modelling of *Electrona antarctica* (Myctophidae, Pisces) in Kerguelen by generalized additive models and geographic information systems. *Polar Biology*, 30, 951– 959.

López-Farrán, Z., Guillaumot, C., (co-first authorship), Vargas-Chacoff, L., Paschke, K., Dulière, V., Danis, B., Poulin, E., Saucède, T. & Gerard, K. (in press.). Current and predicted invasive capacity of *Halicarcinus planatus* (Fabricius, 1775) in the Antarctic Peninsula. *Global Change Biology*.

Lorena, A.C., Jacintho, L.F., Siqueira, M.F., De Giovanni, R., Lohmann, L.G., De Carvalho, A. C. & Yamamoto, M. (2011). Comparing machine learning classifiers in potential distribution modelling. *Expert Systems with Applications*, 38(5), 5268-5275.

Lozier, J.D., Aniello, P. & Hickerson, M.J. (2009). Predicting the distribution of Sasquatch in western North America: anything goes with ecological niche modelling. *Journal of Biogeography*, 36(9), 1623-1627.

Luedeling, E., Kindt, R., Huth, N.I. & Koenig, K. (2014). Agroforestry systems in a changing climate—challenges in projecting future performance. *Current Opinion in Environmental Sustainability*, 6, 1-7.

Luoto, M., Pöyry, J., Heikkinen, R.K. & Saarinen, K. (2005). Uncertainty of bioclimate envelope models based on the geographical distribution of species. *Global Ecology and Biogeography*, 14(6), 575-584.

Maldonado, C., Molina, C.I., Zizka, A., Persson, C., Taylor, C.M., Albán, J., ... & Antonelli, A. (2015). Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? *Global Ecology and Biogeography*, 24(8), 973-984.

Manel, S., Williams, H.C. & Ormerod, S.J. (2001). Evaluating presence–absence models in ecology: the need to account for prevalence. *Journal of applied Ecology*, 38(5), 921-931.

Marcer, A., Méndez-Vigo, B., Alonso-Blanco, C. & Picó, F.X. (2016). Tackling intraspecific genetic structure in distribution models better reflects species geographical range. *Ecology and Evolution*, 6(7), 2084-2097.

Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R.K. & Thuiller, W. (2009). Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions*, 15(1), 59-69.

Massada, A.B., Syphard, A.D., Stewart, S.I. & Radeloff, V.C. (2013). Wildfire ignition-distribution modelling: a comparative study in the Huron–Manistee National Forest, Michigan, USA. *International Journal of wildland fire*, 22(2), 174-183.

Mathewson, P.D., Moyer-Horner, L., Beever, E.A., Briscoe, N.J., Kearney, M., Yahn, J.M. & Porter, W.P. (2017). Mechanistic variables can enhance predictive models of endotherm distributions: the American pika under current, past, and future climates. *Global Change Biology*, 23(3), 1048-1064.

McPherson, J.M., Jetz, W. & Rogers, D.J. (2004). The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of applied Ecology*, 41(5), 811-823.

Mearns, L.O., Hulme, M., Carter, T.R., Leemans, R., Lal, M., Whetton, P. ... & Wilby, R. (2001). Climate scenario development. In *Climate change 2001: the science of climate change* (pp. 739-768). Cambridge University Press. <https://www.ipcc.ch/site/assets/uploads/2018/03/TAR-13.pdf>

Meier, E.S., Edwards Jr, T.C., Kienast, F., Dobbertin, M. & Zimmermann, N.E. (2011). Co-occurrence patterns of trees along macro-climatic gradients and their potential influence on the present and future distribution of *Fagus sylvatica* L. *Journal of Biogeography*, 38(2), 371-382.

Merow, C., Smith, M.J. & Silander Jr, J.A. (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, 36(10), 1058-1069.

Merow, C., Smith, M.J., Edwards Jr, T.C., Guisan, A., McMahon, S.M., Normand, S. ... & Elith, J. (2014). What do we gain from simplicity versus complexity in species distribution models? *Ecography*, 37(12), 1267-1281.

Merow, C., Wilson, A.M. & Jetz, W. (2017). Integrating occurrence data and expert maps for improved species range predictions. *Global Ecology and Biogeography*, 26(2), 243-258.

Mesgaran B.M., Cousens R.D. & Webber B.L. (2014). Here be dragons: a tool for quantifying novelty due to covariate range and correlation change when projecting species distribution models. *Diversity and Distributions*, 20, 1147-1159.

Meynard, C.N. & Quinn, J.F. (2007). Predicting species distributions: a critical comparison of the most common statistical models using artificial species. *Journal of Biogeography*, 34(8), 1455-1469.

Mitchell, P.J., Monk, J. & Laurenson, L. (2017). Sensitivity of fine-scale species distribution models to locational uncertainty in occurrence data across multiple sample sizes. *Methods in Ecology and Evolution*, 8(1), 12-21.

Molloy, S.W., Davis, R.A., Dunlop, J.A. & van Etten, E. (2017). Applying surrogate species presences to correct sample bias in species distribution models: a case study using the Pilbara population of the Northern Quoll. *Nature Conservation*, 18, 27-46.

Monk, J. (2014). How long should we ignore imperfect detection of species in the marine environment when modelling their distribution? *Fish and Fisheries*, 15(2), 352-358.

Moore, J.M., Carvajal, J.I., Rouse, G.W. & Wilson, N.G. (2018). The Antarctic Circumpolar Current isolates and connects: Structured circumpolarity in the sea star *Glabraster antarctica*. *Ecology and Evolution*, 8(21), 10621-10633.

Morales, N.S., Fernández, I.C. & Baca-González, V. (2017). MaxEnt's parameter configuration and small samples: are we paying attention to recommendations? A systematic review. *PeerJ*, 5, e3093.

Moreau, C., Saucède, T., Jossart, Q., Agüera, A., Brayard, A. & Danis, B. (2017). Reproductive strategy as a piece of the biogeographic puzzle: a case study using Antarctic sea stars (Echinodermata, Asteroidea). *Journal of Biogeography*, 44(4), 848-860.

Moreau, C., Mah, C., Agüera, A., Améziane, N., Barnes, D., Crokaert, G. ... & Jażdżewska, A. (2018). Antarctic and sub-Antarctic Asteroidea database. *ZooKeys*, (747), 141.

Moreau, C., Jossart, Q., Danis, B., Eléaume, M., Christiansen, H., Guillaumot, C., Downey, R., Saucède, T. (2020). The overlooked diversity of Southern Ocean sea stars (Asteroidea) reveals original evolutionary pathways. *Progress in Oceanography*. 102472.

Moreno-Amat, E., Mateo, R.G., Nieto-Lugilde, D., Morueta-Holme, N., Svenning, J.C. & García-Amorena, I. (2015). Impact of model complexity on cross-temporal transferability in Maxent species distribution models: An assessment using paleobotanical data. *Ecological Modelling*, 312, 308-317.

Mormède, S., Irisson, J.O. & Raymond, B. (2014). Chapter 2.3. Distribution modelling. In: De Broyer C., Koubbi P., Griffiths H.J., Raymond B., Udekem d'Acoz C. d', et al. (eds.). *Biogeographic Atlas of the Southern Ocean*. Scientific Committee on Antarctic Research, Cambridge, pp. 27-29.

Mulcahy, K.A. & Clarke, K.C. (2001). Symbolization of map projection distortion: a review. *Cartography and Geographic Information Science*, 28(3), 167-182.

Murphey, P.C., Guralnick, R P., Glaubitz, R., Neufeld, D., & Ryan, J.A. (2004). Georeferencing of museum collections: A review of problems and automated tools, and the methodology developed by the Mountain and Plains Spatio-Temporal Database-Informatics Initiative (Mapstedi). *PhyloInformatics*, 3, 1-29.

Murray, R.J. (1996). Explicit generation of orthogonal grids for ocean models. *Journal of Computational Physics*, 126(2), 251-273.

Muscarella, R., Galante, P.J., Soley-Guardia, M., Boria, R.A., Kass, J.M., Uriarte, M. & Anderson, R.P. (2014). ENM eval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. *Methods in Ecology and Evolution*, 5(11), 1198-1205.

Nachtsheim, D.A., Jerosch, K., Hagen, W., Plötz, J. & Bornemann, H. (2017). Habitat modelling of crabeater seals (*Lobodon carcinophaga*) in the Weddell Sea using the multivariate approach Maxent. *Polar Biology*, 40(5), 961-976.

Naimi, B., Skidmore, A.K., Groen, T.A. & Hamm, N.A. (2011). Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling. *Journal of Biogeography*, 38(8), 1497-1509.

Naimi, B., Araujo, M.B., Naimi, M.B., Naimi, B. & Araujo, M.B. (2019). Package 'sdm'. <https://cran.r-project.org/web/packages/sdm/index.html>

Newbold, T. (2010). Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in Physical Geography*, 34(1), 3-22.

Ocaranza-Barrera, P., González-Wevar, C.A., Guillemin, M.L., Rosenfeld, S. & Mansilla, A. (2019). Molecular divergence between *Iridaea cordata* (Turner) Bory de Saint-Vincent from the Antarctic Peninsula and the Magellan Region. *Journal of applied Phycology*, 31(2), 939-949.

Olden J.D., Lawler J.J. & Poff N.L. (2008) Machine learning methods without tears: a primer for ecologists. *The Quarterly Review of Biology*, 83(2), 171–193

Osborne, P.E. & Leitão, P.J. (2009). Effects of species and habitat positional errors on the performance and interpretation of species distribution models. *Diversity and Distributions*, 15(4), 671-681.

Owens, H.L., Campbell, L.P., Dornak, L.L., Saupe, E.E., Barve, N., Soberón, J. ... & Peterson, A.T. (2013). Constraints on interpretation of ecological niche models by limited environmental ranges on calibration areas. *Ecological Modelling*, 263, 10-18.

Pagel, J. & Schurr, F.M. (2012). Forecasting species ranges by statistical estimation of ecological niches and spatial population dynamics. *Global Ecology and Biogeography*, 21(2), 293-304.

Pardo-Gandarillas, M.C., Ibáñez, C.M., Torres, F.I., Sanhueza, V., Fabres, A., Escobar-Dodero, J. ... & Méndez, M.A. (2018). Phylogeography and species distribution modelling reveal the effects of the Pleistocene ice ages on an intertidal limpet from the south-eastern Pacific. *Journal of Biogeography*, 45(8), 1751-1767.



Pearson, R.G., Thuiller, W., Araújo, M.B., Martinez-Meyer, E., Brotons, L., McClean, C. ... & Lees, D.C. (2006). Model-based uncertainty in species range prediction. *Journal of Biogeography*, 33(10), 1704-1711.

Pearson, R.G. (2007). Species' distribution modeling for conservation educators and practitioners. *Synthesis. American Museum of Natural History*, 50, 54-89.

Pearson, R.G., Raxworthy, C.J., Nakamura, M. & Townsend Peterson, A. (2007). Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography*, 34(1), 102-117.

Perrault-Hébert M., Girard F. Boucher Y. Fournier R., Mansuy N., Valeria O. (2019). Evaluation of spatiotemporal transferability of wildfire probability across eastern boreal forest of North America. Chapter 1, PhD dissertation. University Montréal.

Peel, S.L., Hill, N.A., Foster, S.D., Wotherspoon, S.J., Ghiglione, C. & Schiaparelli, S. (2019) Reliable species distributions are obtainable with sparse, patchy and biased data by leveraging over species and data types. *Methods in Ecology and Evolution*. 10(7), 1002-1014.

Pellissier, L., Anne Bråthen, K., Pottier, J., Randin, C.F., Vittoz, P., Dubuis, A. ... & Guisan, A. (2010). Species distribution models reveal apparent competitive and facilitative effects of a dominant species on the distribution of tundra plants. *Ecography*, 33(6), 1004-1014.

Pellissier, L., Rohr, R.P., Ndiribe, C., Pradervand, J.N., Salamin, N., Guisan, A. & Wisz, M. (2013). Combining food web and species distribution models for improved community projections. *Ecology and Evolution*, 3(13), 4572-4583.

Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R.P., Martínez-Meyer, E., Nakamura, M. & Araújo, M.B. (2011). *Ecological niches and geographic distributions (MPB-49)* (Vol. 56). Princeton University Press.

Petitpierre, B., Broennimann, O., Kueffer, C., Daehler, C. & Guisan, A. (2017). Selecting predictors to maximize the transferability of species distribution models: lessons from cross-continental plant invasions. *Global Ecology and Biogeography*, 26(3), 275-287.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3-4), 231-259.

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

Phillips, N.D., Reid, N., Thys, T., Harrod, C., Payne, N. L., Morgan, C.A., ... & Houghton, J.D. (2017). Applying species distribution modelling to a data poor, pelagic fish complex: The ocean sunfishes. *Journal of Biogeography*, 44(10), 2176-2187.

Pierrat, B., Saucède, T., Laffont, R., De Ridder, C., Festeau, A. & David, B. (2012). Large-scale distribution analysis of Antarctic echinoids using ecological niche modelling. *Marine Ecology Progress Series*, 463, 215-230.

Pinkerton, M.H., Smith, A.N., Raymond, B., Hosie, G.W., Sharp, B., Leathwick, J.R. & Bradford-Grieve, J.M. (2010). Spatial and seasonal distribution of adult *Oithona similis* in the Southern Ocean: Predictions using boosted regression trees. *Deep Sea Research Part I: Oceanographic Research Papers*, 57, 469–485.

Pittman, S.J. (Ed.). (2017). *Seascape ecology*. John Wiley & Sons. 501pp

Pope, A., Wagner, P., Johnson, R., Shutler, J.D., Baeseman, J. & Newman, L. (2017). Community review of Southern Ocean satellite data needs. *Antarctic Science*, 29(2), 97-138.

Porfirio, L.L., Harris, R.M., Lefroy, E.C., Hugh, S., Gould, S.F., Lee, G. ... & Mackey, B. (2014). Improving the use of species distribution models in conservation planning and management under climate change. *PLoS One*, 9(11), e113749.

Qiao, H., Soberón, J. & Peterson, A.T. (2015). No silver bullets in correlative ecological niche modelling: insights from testing among many potential algorithms for niche estimation. *Methods in Ecology and Evolution*, 6(10), 1126-1136.

Radosavljevic, A. & Anderson, R.P. (2014). Making better Maxent models of species distributions: complexity, overfitting and evaluation. *Journal of Biogeography*, 41(4), 629-643.

Raes, N. & ter Steege, H. (2007). A null-model for significance testing of presence-only species distribution models. *Ecography*, 30(5), 727-736.

Raes, N. (2012). Partial versus full species distribution models. *Natureza & Conservação*, 10(2), 127-138.

Ranc, N., Santini, L., Rondinini, C., Boitani, L., Poitevin, F., Angerbjörn, A. & Maiorano, L. (2017). Performance tradeoffs in target-group bias correction for species distribution models. *Ecography*, 40(9), 1076-1087.

Randin, C.F., Dirnböck, T., Dullinger, S., Zimmermann, N.E., Zappa, M. & Guisan, A. (2006). Are niche-based species distribution models transferable in space?. *Journal of Biogeography*, 33(10), 1689-1703.

Raymond, B. (2012) Polar Environmental Data Layers, Ver. 3, *Australian Antarctic Data Centre* - [https://data.aad.gov.au/metadata/records/Polar\\_Environmental\\_Data](https://data.aad.gov.au/metadata/records/Polar_Environmental_Data), Accessed: 2020-10-26

Raymond, B., Lea, M.A., Patterson, T., Andrews-Goff, V., Sharples, R., Charrassin, J.B., ... & Goldsworthy, S.D. (2015). Important marine habitat off east Antarctica revealed by two decades of multi-species predator tracking. *Ecography*, 38(2), 121-129.

Ready, J., Kaschner, K., South, A.B., Eastwood, P.D., Rees, T., Rius, J. ... & Froese, R. (2010). Predicting the distributions of marine organisms at the global scale. *Ecological Modelling*, 221(3), 467-478.

Reiss, H., Cunze, S., König, K., Neumann, H. & Kröncke, I. (2011). Species distribution modelling of marine benthos: a North Sea case study. *Marine Ecology Progress Series*, 442, 71-86.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Aroita, G., ... & Warton, D.I. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8), 913-929.

Robinson, L.M., Elith, J., Hobday, A.J., Pearson, R.G., Kendall, B.E., Possingham, H.P. & Richardson, A.J. (2011). Pushing the limits in marine species distribution modelling: lessons from the land present challenges and opportunities. *Global Ecology and Biogeography*, 20(6), 789-802.

Robinson, N.M., Nelson, W.A., Costello, M.J., Sutherland, J.E. & Lundquist, C.J. (2017). A systematic review of marine-based Species Distribution Models (SDMs) with recommendations for best practice. *Frontiers in Marine Science*, 4, 421.

Rocchini, D., Hortal, J., Lengyel, S., Lobo, J.M., Jimenez-Valverde, A., Ricotta, C., ... & Chiarucci, A. (2011). Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. *Progress in Physical Geography*, 35(2), 211-226.

Rodda, G.H., Jarnevich, C.S. & Reed, R.N. (2011). Challenges in identifying sites climatically matched to the native ranges of animal invaders. *PloS one*, 6(2), e14670.

Rodríguez, L., García, J.J., Carreño, F. & Martínez, B. (2019). Integration of physiological knowledge into hybrid species distribution modelling to improve forecast of distributional shifts of tropical corals. *Diversity and Distributions*, 25(5), 715-728.

Ropert-Coudert, Y., Van de Putte, A. P., Reisinger, R. R., Bornemann, H., Charrassin, J. B., Costa, D. P., ... & Hindell, M. A. (2020). The retrospective analysis of Antarctic tracking data project. *Scientific data*, 7(1), 1-11.

Sánchez-Fernández, D., Lobo, J.M. & Hernández-Manrique, O.L. (2011). Species distribution models that do not incorporate global data misrepresent potential distributions: a case study using Iberian diving beetles. *Diversity and Distributions*, 17(1), 163-171.

Scales, K.L., Miller, P.I., Ingram, S.N., Hazen, E.L., Bograd, S.J. & Phillips, R.A. (2016). Identifying predictable foraging habitats for a wide-ranging marine predator using ensemble ecological niche models. *Diversity and Distributions*, 22(2), 212-224.

Segurado, P., Araujo, M.B. & Kunin, W.E. (2006). Consequences of spatial autocorrelation for niche-based models. *Journal of Applied Ecology*, 43(3), 433-444.

Seo, C., Thorne, J.H., Hannah, L. & Thuiller, W. (2009). Scale effects in species distribution models: implications for conservation planning under climate change. *Biology Letters*, 5(1), 39-43.

Schapire. (1990). The strength of weak learnability. *Machine Learning*, 5(2): 197–227.

Sillero, N. (2011). What does ecological modelling model? A proposed classification of ecological niche models based on their underlying methods. *Ecological Modelling*, 222(8), 1343-1346.

Sillero, N., & Barbosa, A. M. (2021). Common mistakes in ecological niche models. *International Journal of Geographical Information Science*, 35(2), 213-226.

Smith, A.B. (2013). On evaluating species distribution models with random background sites in place of absences when test presences disproportionately sample suitable habitat. *Diversity and Distributions*, 19(7), 867-872.

Snickars, M., Gullström, M., Sundblad, G., Bergström, U., Downie, A.L., Lindegarth, M. & Mattila, J. (2014). Species–environment relationships and potential for distribution modelling in coastal waters. *Journal of Sea Research*, 85, 116-125.

Staveley, T.A., Perry, D., Lindborg, R. & Gullström, M. (2017). Seascape structure and complexity influence temperate seagrass fish assemblage composition. *Ecography*, 40(8), 936-946.

Støa, B., Halvorsen, R., Mazzoni, S. & Gusarov, V.I. (2018). Sampling bias in presence-only data used for species distribution modelling: theory and methods for detecting sample bias and its effects on models. *Sommerfeltia*, 38(1), 1-53.

Stock, A., Subramaniam, A., Van Dijken, G.L., Wedding, L.M., Arrigo, K.R., Mills, M.M. ... & Micheli, F. (2020). Comparison of Cloud-Filling Algorithms for Marine Satellite Data. *Remote Sensing*, 12(20), 3313.

Stockwell, D.R. & Peterson, A.T. (2002). Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, 148(1), 1-13.

Syfert, M.M., Smith, M.J. & Coomes, D.A. (2013). The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. *PloS one*, 8(2).

Talluto, M.V., Boulangeat, I., Ameztegui, A., Aubin, I., Berteaux, D., Butler, A. ... & Liénard, J. (2016). Cross-scale integration of knowledge for predicting species ranges: a metamodelling framework. *Global Ecology and Biogeography*, 25(2), 238-249.

Telford, R.J. & Birks, H.J. (2009). Evaluation of transfer functions in spatially structured environments. *Quaternary Science Reviews*, 28(13-14), 1309-1316.

Tessarolo, G., Ladle, R., Rangel, T. & Hortal, J. (2017). Temporal degradation of data limits biodiversity research. *Ecology and Evolution*, 7(17), 6863-6870.

Thatje, S. (2012). Effects of capability for dispersal on the evolution of diversity in Antarctic benthos. *Integrative and Comparative Ecology*, 52(4): 470-482

Thiers, L., Delord, K., Bost, C.-A., Guinet, C. & Weimerskirch, H. (2017). Important marine sectors for the top predator community around Kerguelen Archipelago. *Polar Biology*, 40, 365–378.

Thuiller, W., Vayreda, J., Pino, J., Sabate, S., Lavorel, S. & Gracia, C. (2003). Large-scale environmental correlates of forest tree distributions in Catalonia (NE Spain). *Global Ecology and Biogeography*, 12(4), 313-325.

Thuiller, W., Brotons, L., Araújo, M.B. & Lavorel, S. (2004). Effects of restricting environmental range of data to project current and future species distributions. *Ecography*, 27(2), 165-172.

Thuiller, W., Lafourcade, B., Engler, R. & Araújo, M.B. (2009). BIOMOD—a platform for ensemble forecasting of species distributions. *Ecography*, 32(3), 369-373.

Thuiller, W., Georges, D. & Engler, R. (2018). Ensemble platform for species distribution modeling package biomod2 version3.3 <https://CRAN.Rproject.org/package=biomod2>.

Thyrring, J., Bundgaard, A. & Sejr, M.K. (2017). Seasonal acclimation and latitudinal adaptation are of the same magnitude in *Mytilus edulis* and *Mytilus trossulus* mitochondrial respiration. *Polar Biology*, 40(9), 1885-1891.

Tirunelveli, G., Gordon, R. & Pistorius, S. (2002). Comparison of square-pixel and hexagonal-pixel resolution in image processing. In *IEEE CCECE2002. Canadian Conference on Electrical and Computer Engineering. Conference Proceedings (Cat. No. 02CH37373)* (Vol. 2, pp. 867-872). IEEE.

Titeux, N., Maes, D., Van Daele, T., Onkelinx, T., Heikkinen, R.K., Romo, H. ... & Schweiger, O. (2017). The need for large-scale distribution data to estimate regional changes in species richness under future climate change. *Diversity and Distributions*, 23(12), 1393-1407.

Trolle, D., Elliott, J.A., Mooij, W.M., Janse, J.H., Bolding, K., Hamilton, D.P. & Jeppesen, E. (2014). Advancing projections of phytoplankton responses to climate change through ensemble modelling. *Environmental Modelling & software*, 61, 371-379.

Trull, T.W., Passmore, A., Davies, D.M., Smit, T., Berry, K. & Tilbrook, B. (2018). Distribution of planktonic biogenic carbonate organisms in the Southern Ocean south of Australia: a baseline for ocean acidification impact assessment. *Biogeosciences*, 15(1), 31.

Tsoar, A., Allouche, O., Steinitz, O., Rotem, D. & Kadmon, R. (2007). A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions*, 13(4), 397-405.

Turner, J., Barrand, N., Bracegirdle, T., Convey, P., Hodgson, D.A., Jarvis, M., ... *et al.* (2014). Antarctic climate change and the environment: an update. *Polar Record*, 50 (3), 237-259.

Valavi, R., Elith, J., Lahoz-Monfort, J.J. & Guillerá-Arroita, G. (2018). blockCV: an R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *bioRxiv*, 357798.

Vale, C.G., Tarroso, P. & Brito, J.C. (2014). Predicting species distribution at range margins: testing the effects of study area extent, resolution and threshold selection in the Sahara–Sahel transition zone. *Diversity and Distributions*, 20(1), 20-33.

Vandersteen, W. (2011). Detecting gene expression profiles associated with environmental stressors within an ecological context. *Molecular Ecology*, 20(7), 1322-1323.

Vanden Berghe, E. (2013). Report of the project 'Turning OBIS data into information'; project funded by the Census of Marine Life International Cosmos Prize Fund. Rutgers University. 62 pp.

van Proosdij, A.S., Sosef, M.S., Wieringa, J.J. & Raes, N. (2016). Minimum required number of specimen records to develop accurate species distribution models. *Ecography*, 39(6), 542-552.



Vapnik, V. (1998). *Statistical learning theory* Wiley. New York, 1, 624.

Warren, D.L. & Seifert, S.N. (2011). Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. *Ecological Applications*, 21(2), 335-342.

Watling, J.I., Brandt, L.A., Bucklin, D.N., Fujisaki, I., Mazzotti, F.J., Romañach, S.S. & Speroterra, C. (2015). Performance metrics and variance partitioning reveal sources of uncertainty in species distribution models. *Ecological Modelling*, 309, 48-59.

Wenger, S.J. & Olden, J.D. (2012). Assessing transferability of ecological models: an underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, 3(2), 260-267.

Wiley, E.O., McNyset, K.M., Peterson, A.T., Robins, C.R. & Stewart, A.M. (2003). Niche modeling perspective on geographic range predictions in the marine environment using a machine-learning algorithm.

Williams, J.W. & Jackson, S.T. (2007). Novel climates, no-analog communities, and ecological surprises. *Frontiers in Ecology and the Environment*, 5(9), 475-482.

Williams, J.W., Jackson, S.T. & Kutzbach, J.E. (2007). Projected distributions of novel and disappearing climates by 2100 AD. *Proceedings of the National Academy of Sciences*, 104(14), 5738-5742.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A. & NCEAS Predicting Species Distributions Working Group. (2008). Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, 14(5), 763-773.

Wisz, M.S. & Guisan, A. (2009). Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. *BMC Ecology*, 9(1), 8.

White-Newsome, J.L., Brines, S.J., Brown, D.G., Dvonch, J.T., Gronlund, C.J., Zhang, K., ... & O'Neill, M.S. (2013). Validating satellite-derived land surface temperature with in situ measurements: A public health perspective. *Environmental health perspectives*, 121(8), 925-931.

Whittingham, M.J., Stephens, P.A., Bradbury, R.B. & Freckleton, R.P. (2006). Why do we still use stepwise modelling in ecology and behaviour?. *Journal of Animal Ecology*, 75(5), 1182-1189.

WOCE (2013). <https://www.nodc.noaa.gov/OC5/woa13/woa13data.html>

Xavier, J.C., Raymond, B., Jones, D.C. & Griffiths, H. (2016). Biogeography of Cephalopods in the Southern Ocean using habitat suitability prediction models. *Ecosystems*, 19, 220–247.

Yackulic, C.B., Chandler, R., Zipkin, E.F., Royle, J.A., Nichols, J.D., Campbell Grant, E.H. & Veran, S. (2013). Presence-only modelling using MAXENT: when can we trust the inferences?. *Methods in Ecology and Evolution*, 4(3), 236-243.

Zaniewski, A.E., Lehmann, A. & Overton, J.M. (2002). Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, 157(2-3), 261-280.

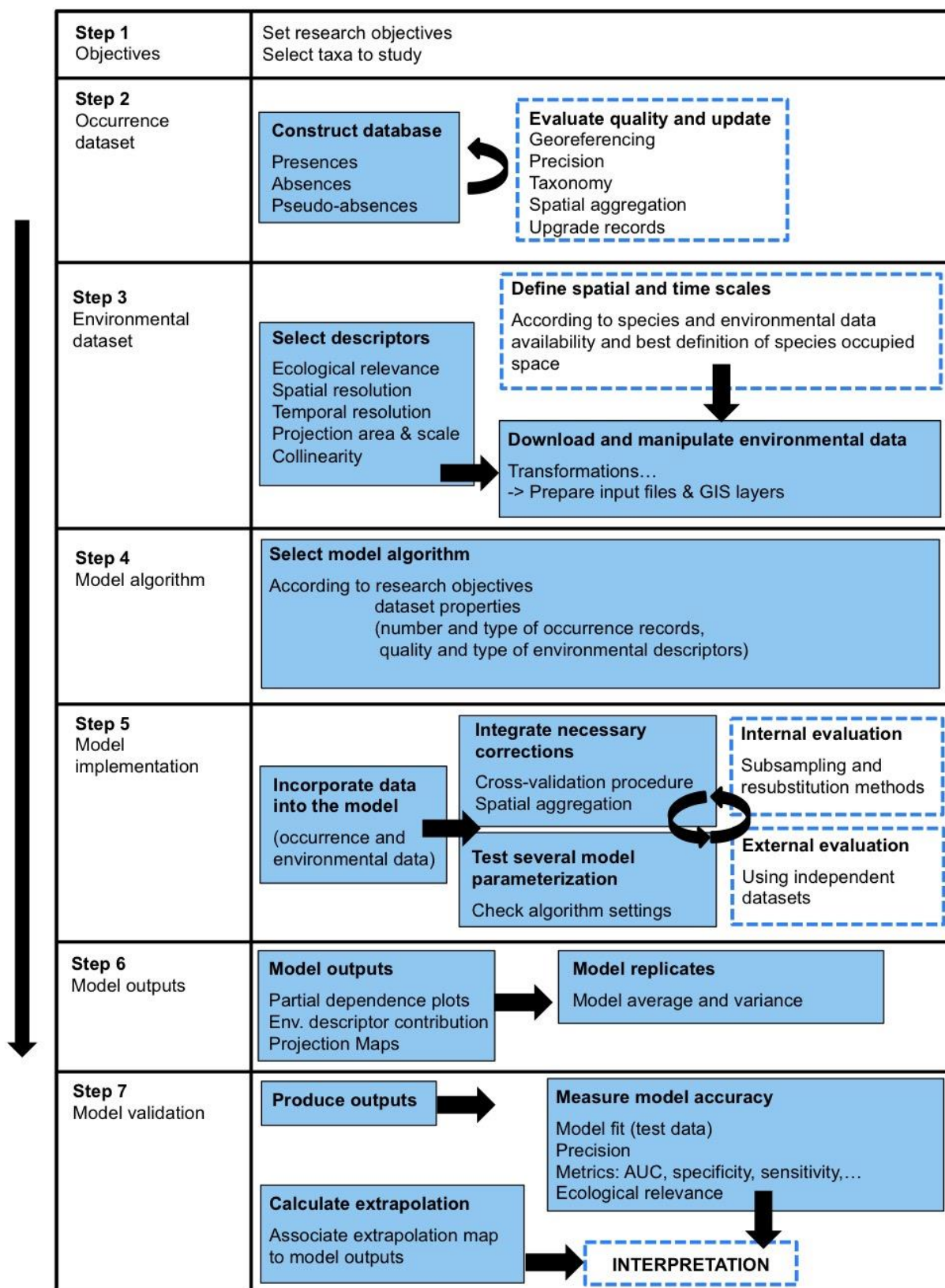
Zhang, X., & Mahadevan, S. (2019). Ensemble machine learning models for aviation incident risk prediction. *Decision Support Systems*, 116, 48-63.

Zhou, Z.H. (2012). *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC.

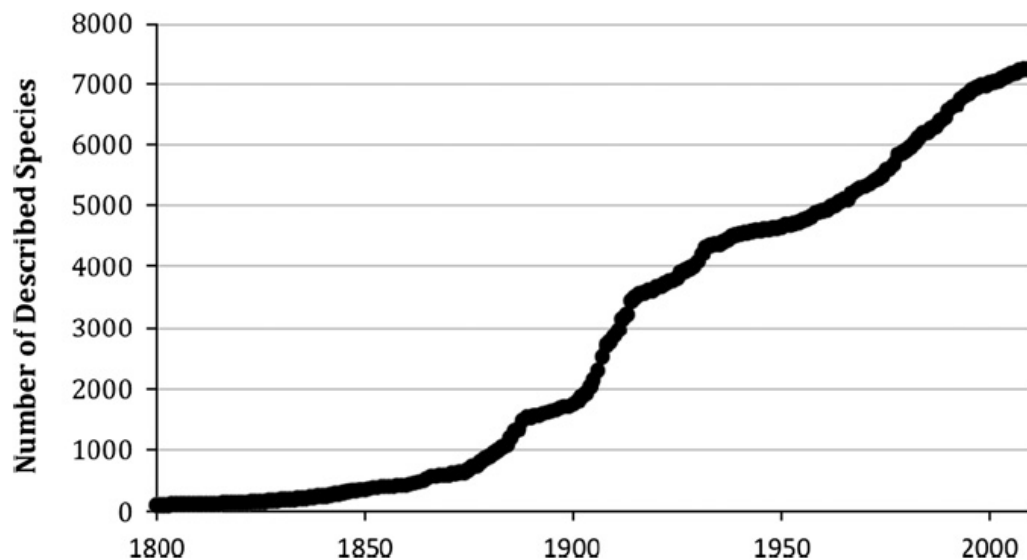
Zhu, G.P. & Peterson, A.T. (2017). Do consensus models outperform individual models? Transferability evaluations of diverse modeling approaches for an invasive moth. *Biological Invasions*, 19(9), 2519-2532.

Zurell, D., Franklin, J., König, C., Bouchet, P.J., Dormann, C.F., Elith, J. ... & Lahoz-Monfort, J.J. (2020). A standard protocol for reporting species distribution models. *Ecography*.

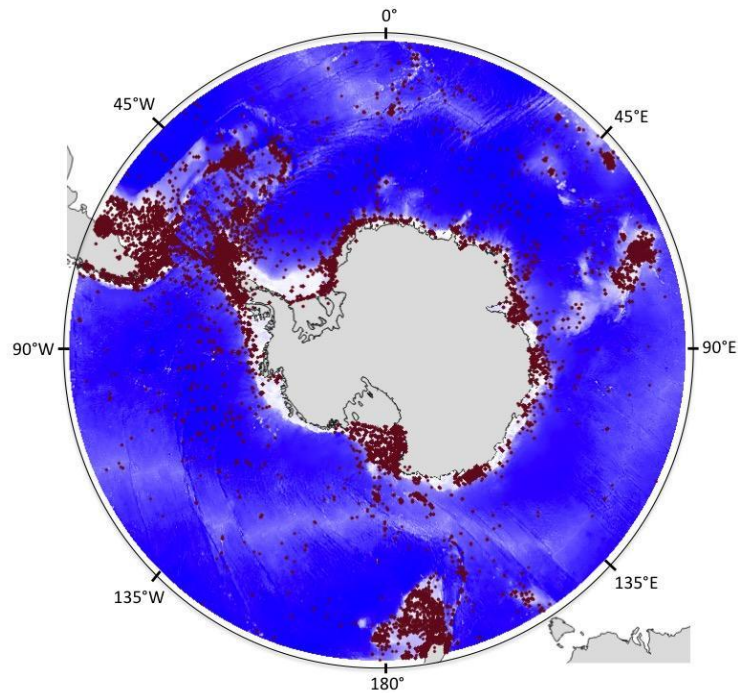




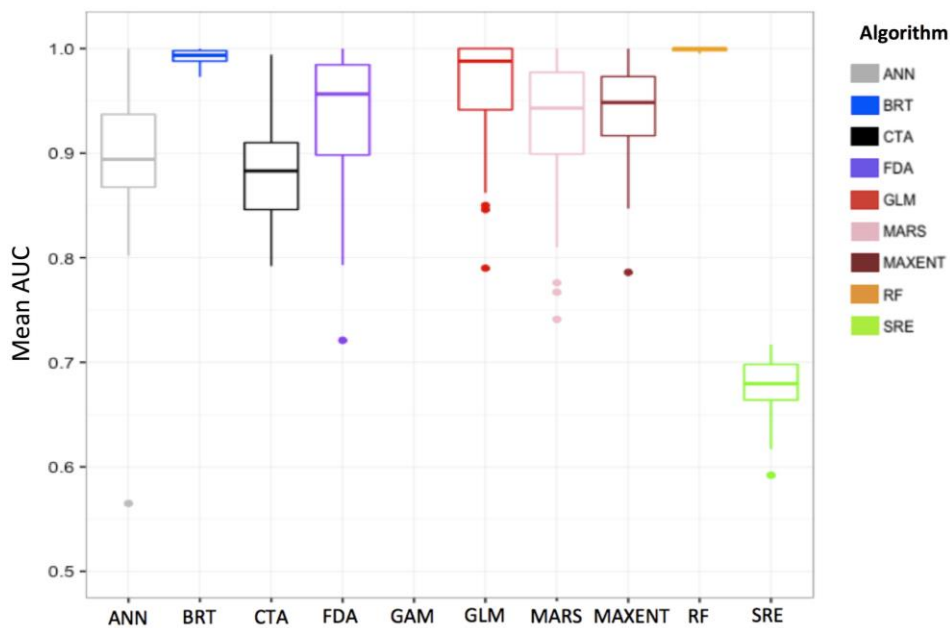
**Figure 1.** Flow chart of the SDM construction process. Steps 1 to 4 concern data collection, and treatment. Steps 5 to 7 integrate procedures for model implementation and validation. Dashed rectangles allow for a possible step backward when assessing model uncertainties or evaluating model performance.



**Figure 2.** Cumulative number of Antarctic species described over time, according to data available in the Register of Antarctic Marine species (until March 2010). From De Broyer and Danis (2011).

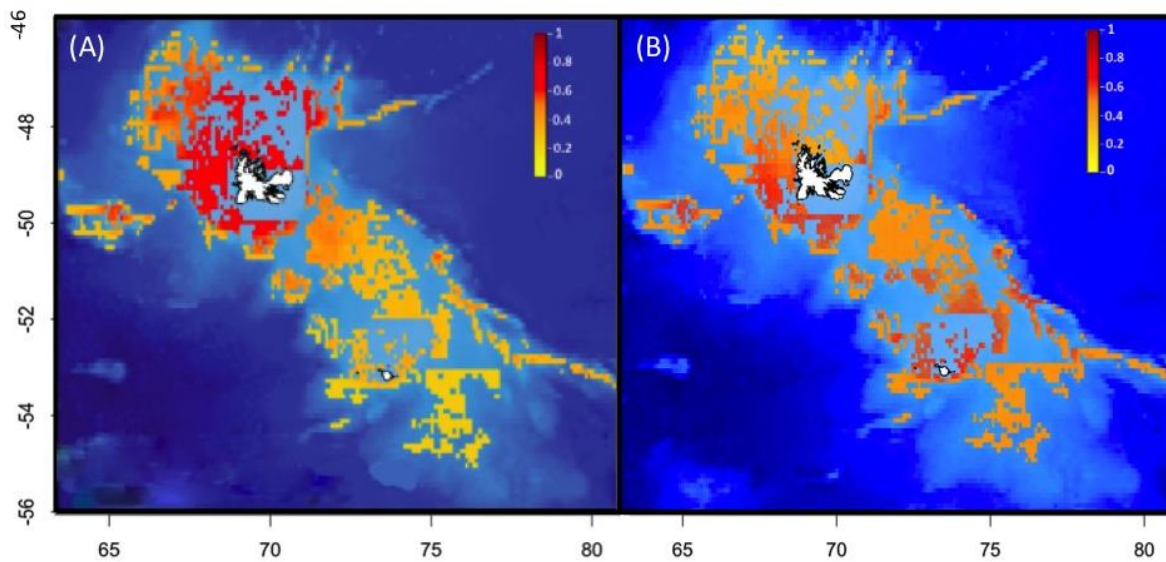


**Figure 3.** Distribution of benthos sampling sites (red dots) in the Southern Ocean (SO, <math><45^{\circ}\text{S}</math>). Sampling sites are not evenly distributed in the SO, showing important spatial aggregation in the Scotia Arc region and Western Antarctic Peninsula with several clusters along the Antarctic shelf, and over the Kerguelen and Campbell plateaus. In contrast, deep-sea regions and remote areas of the Antarctic shelf are under-sampled. From [Guillaumot et al. \(2019\)](#), updated from [Griffiths et al. \(2014\)](#).



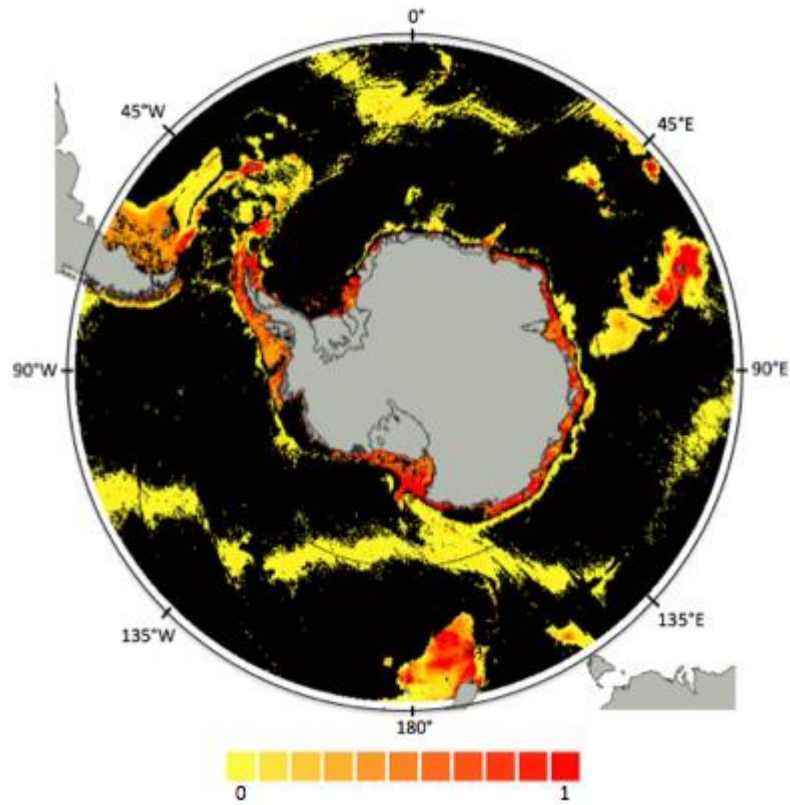
**Figure 4.** Compared Area Under the Curve (AUC) performances of SDMs generated with different algorithms (ANN=Artificial Neural Network, BRT=Boosted Regression Trees, CTA=Classification Tree Analysis, FDA=Flexible Discriminant Analysis, GAM=Generalised Additive Model, GLM=Generalised

Linear Model, MARS=Multivariate Adaptive Regression Splines, MAXENT=Maximum Entropy, RF=Random Forest, SRE=Surface Range Envelope) to predict the distribution of the sea urchin *Sterechinus diadema* in the Southern Ocean. Results show a good performance for BRT and RF, adapted to small, historically compiled datasets (temporally heterogeneous) and spatially aggregated presence-only data. Models were calibrated with presence-only data and 200 background data randomly sampled in the study area. Average scores of 100 model replicates. See [Guillaumot et al. \(2018b\)](#) for details.

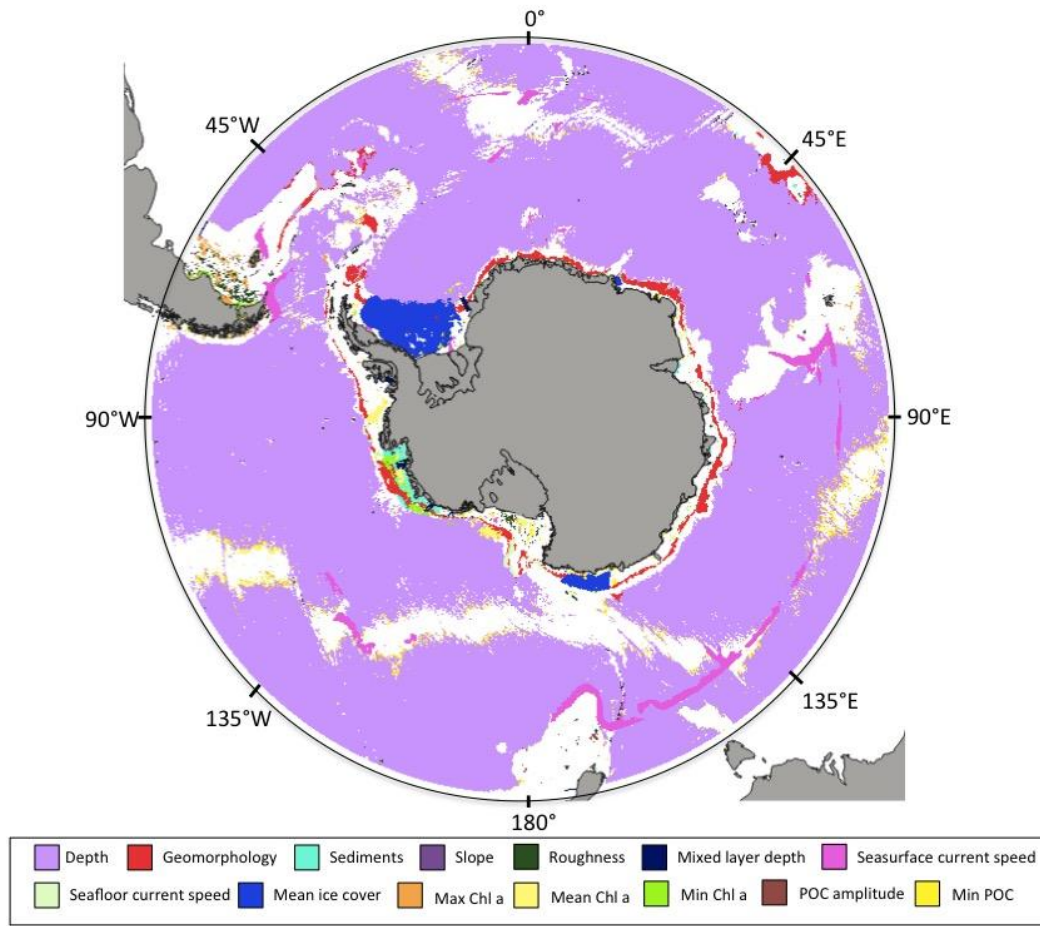


**Figure 5.** Comparison of predicted distribution probabilities (between 0 and 1) of the sea urchin *Ctenocidaris nutrix* on the Kerguelen Plateau: (A) without compensating for sampling bias; (B) with a kernel density estimator (KDE) correction: more background data are sampled in highly sampled areas. The spatial aggregation of presence-only records near the shoreline of the Kerguelen Islands strongly biases model predictions. The KDE correction was proved efficient to correct for such a bias and provides more relevant predictions. From [Guillaumot et al. \(2018a\)](#).

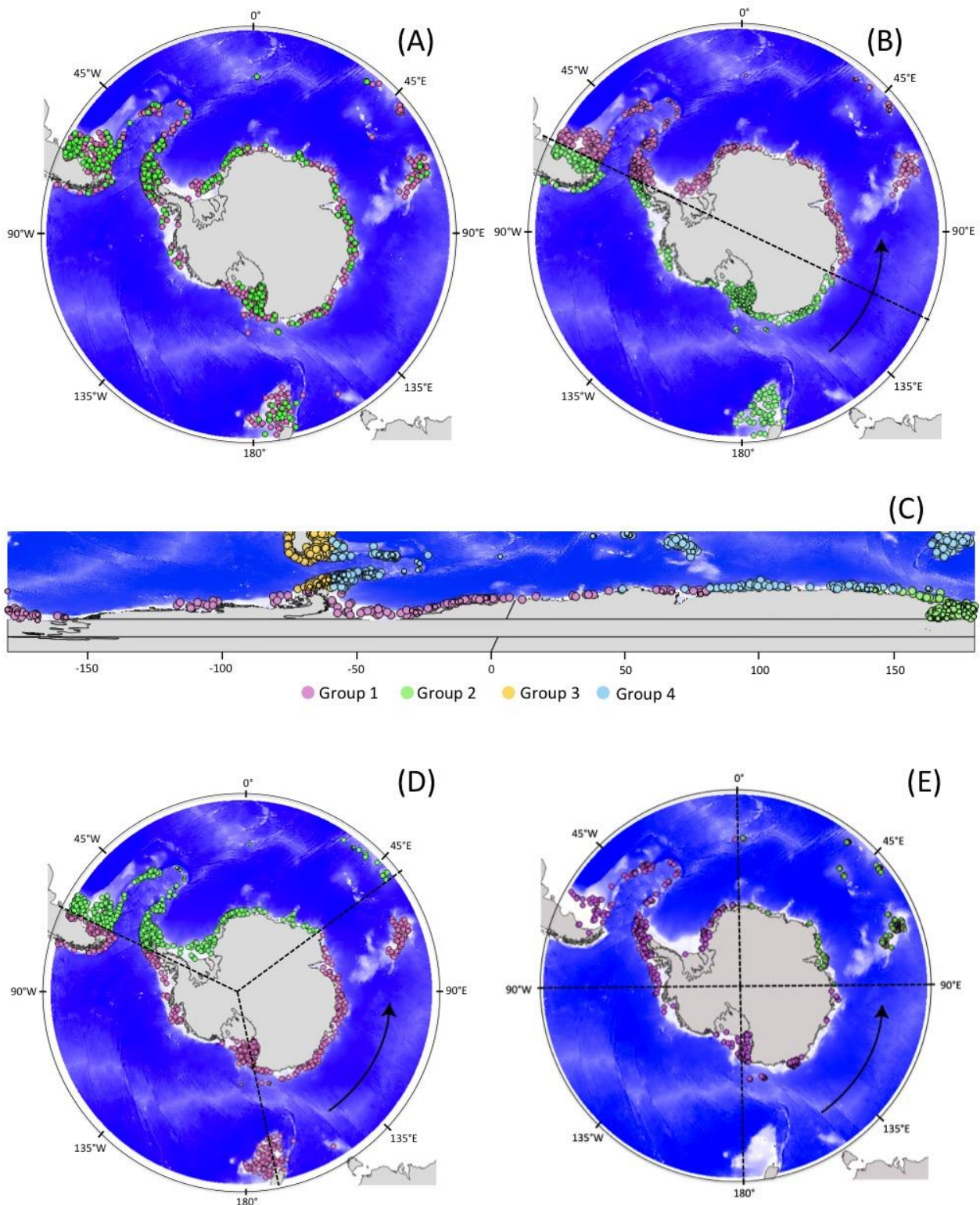




**Figure 6.** Extrapolation map of the SDM generated for the sea star *Acodontaster hodgsoni*, with all presence-only records available. Extrapolation corresponds here to the ensemble of environmental conditions that are outside of the boundaries of the calibration range. The extrapolation area is displayed in black and covers 78.6% of the entire projection area; coloured pixels (yellow-red color palette) show distribution probabilities (included between 0 and 1). Extracted from [Guillaumot et al. \(2020b\)](#).



**Figure 7.** Extrapolation map of the SDM generated for the sea star *Acodontaster hodgsoni* indicating environmental descriptors responsible for extrapolation (black pixels of Fig. 6 are here coloured according to the descriptor responsible for extrapolation; i.e. for each pixel, the predictor in question lies outside the calibration range). In this case study, 14 environmental descriptors are responsible for extrapolation, depth being the main contributor. White pixels correspond to areas where the model does not extrapolate (the corresponding model predictions are shown in Fig. 6). POC stands for 'Particulate Organic Carbon', Chla is the concentration in chlorophyll a in the seafloor. Generated from Guillaumot et al. (2021).



**Figure 8.** Different cross-validation procedures based on the study of the sea star *Odontaster validus*, showing presence-only records and a random set of 1,000 background data selected according to a Kernel Density Estimation weighting scheme from the dataset of Griffiths et al. (2014) on sampling effort of the SO benthos. Data are split into training (pink) and test (green) subsets. The blue background corresponds to bathymetry and grey areas to emerged lands. (A) Random cross-validation procedure, with a random split into 75% training and 25% test data. (B) '2-fold CLOCK' clustering by random spatial

partition of the dataset into two groups (one training, one test). (C) 'BLOCK' splitting, generated according to median latitudinal and longitudinal values ([Muscarella et al. 2014](#)). After the generation of four groups (corresponding to the four colors), one group is randomly defined as the test subset, the other three groups as the training subsets. A different system of projection was used to represent this map to highlight the latitudinal and longitudinal definition of transects. (D) '3-fold CLOCK' clustering by random spatial partition of the dataset into three groups (2 training, 1 test). (E) '4-fold CLOCK' clustering by random spatial partition of the dataset into four groups (3 training, 1 test). Figure extracted from [Guillaumot et al. \(2019\)](#).