Ecological Modelling 323 (2016) 61–76

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Estimating absence locations of marine species from data of scientific surveys in OBIS

Gianpaolo Coro[a,*], Chiara Magliozzi[a], Edward Vanden Berghe[b], Nicolas Bailly[c,d], Anton Ellenbroek[e], Pasquale Pagano[a]

[a] Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo" – CNR, via Moruzzi 1, 56124 Pisa, Italy
[b] Vrije Universiteit Brussel (VUB), Pleinlaan 2, 1050 Elsene, Belgium
[c] LifeWatchGreece, Hellenic Centre for Marine Research (HCMR), Gouves, 71500 Heraklion, Greece
[d] FishBase Information and Research Group (FIN), 4031 Los Baños, Laguna, Philippines
[e] Food and Agriculture Organization of the United Nations (FAO), Viale delle Terme di Caracalla, 00153 Rome, Italy

ARTICLE INFO

Article history:
Received 16 July 2015
Received in revised form
14 December 2015
Accepted 16 December 2015

Keywords:
Absence locations
Species distribution maps
Occurrence data
Ecological niche modelling
Marine biodiversity
Scientific surveys

ABSTRACT

Estimating absence locations of a species is important in conservation biology and conservation planning. For instance, using reliable absence as much as presence information, species distribution models can enhance their performance and produce more accurate predictions of the distribution of a species. Unfortunately, estimating reliable absence locations is difficult and often requires a deep knowledge of the species' distribution and of its abiotic and biotic environmental preferences and tolerance. In this paper, we propose a methodology to reconstruct reliable absence information from presence-only information, and the conditions that those presence-only data have to meet to make this possible.

Large species occurrence data collections (otherwise called occurrence datasets) contain high quality and expert-reviewed species observation records from scientific surveys. These surveys can be used to retrieve species presence locations, but they also record places where the species in their target list were not observed. Although these absences could be simply due to sampling variation, it is possible to intersect many of these reports to estimate true absence locations, i.e. those due to habitat unsuitability or geographical hindrances. In this paper, we present a method to generate reliable absence locations of this type for marine species, using scientific surveys reports contained in the Ocean Biogeographic Information System (OBIS), an authoritative species occurrence dataset. Our method spatially aggregates information from surveys focussing on the same target species. It detects absence locations for a given species as those locations in which repeated surveys (that included the species of interest in their target list) reported information only on other species. We qualitatively demonstrate the reliability of our method using distribution records of the Atlantic cod as a case study. Additionally, we quantitatively estimate its performance using another authoritative large species occurrence dataset, the Global Biodiversity Information Facility (GBIF). We also demonstrate that our approach has higher accuracy and presents complementary behaviour with respect to another method using environmental envelopes. Our process can support species distribution models (as well as other types of models, e.g. climate change models) by providing reliable data to presence/absence approaches. It can manage regional as well as global scale scenarios and runs within a collaborative e-Infrastructure (D4Science) that publishes it as-a-Service, allowing biologists to reproduce, repeat and share experimental results.

1. Introduction

Species distribution models (SDMs) estimate species distributions at global or local scale, by relating species occurrence records to a set of environmental parameters. SDMs have high potential in conservation biology and conservation planning, because they give hints to understand the relationship between a species and its abiotic and biotic environment, and to test ecological or biogeographical hypotheses about species distributions and ranges.

* Corresponding author. Tel.: +39 050 315 2978; fax: +39 050 621 3464.
E-mail addresses: coro@isti.cnr.it (G. Coro), chiara.magliozzi@isti.cnr.it (C. Magliozzi), evberghe@gmail.com (E. Vanden Berghe), nbailly@hcmr.gr (N. Bailly), Anton.Ellenbroek@fao.org (A. Ellenbroek), pagano@isti.cnr.it (P. Pagano).

http://dx.doi.org/10.1016/j.ecolmodel.2015.12.008
0304-3800/

They generalise the distribution that can be inferred from the observed locations, and possibly account for bias due to non-uniform observations sampling. Several technologies are used to build SDMs, ranging from explicit modelling of physiological limits and tolerances (Pearson, 2012), to the automatic correlation between species presence and environmental characteristics (Elith and Leathwick, 2009). Often, the output of an SDM is a probability distribution map reporting locations, at a certain resolution, where habitat is suitable for a species. SDMs usually use habitat information on recorded species observations and some models use also *habitat-related* absence locations, where habitat is unsuitable for species subsistence. Estimating these absence locations is a necessary step in these SDMs and requires separate modelling effort, e.g. envelope models based on species preferences to abiotic and biotic factors (Barbet-Massin et al., 2012). In this paper, we will distinguish these locations from the absences reported by scientific surveys (*sampling* absences). *Sampling* absences could refer either to complete absence of a species in a certain location, or to "undetected" presence, which could be due to intrinsic issues in species detectability and seasonality, or just to random sampling variation. Thus, the difference between *habitat-related* absences and *sampling* absences is in the fact that the former type is estimated from abiotic and biotic parameters, and the latter type is estimated from surveyed locations without presence data. Both the types are *pseudo*-absences because they use partial information about the species to estimate absence locations. Apart from *pseudo*-absences, in this work we will use the expression *absence locations* (or *true* absences) to indicate locations where the species is absent due to real habitat unsuitability or geographical hindrances. Based on this nomenclature, we define *reliable pseudo*-absences as those *pseudo*-absences that well approximate real absence locations.

Some SDMs rely on presence information only, for example Genetic Algorithm for Rule-set Production (GARP) and Ecological Niche Factor Analysis (ENFA) are based on simulated *pseudo*-absences (Stockwell, 1999; Engler et al., 2004), but models relying on both presence and absence information may reach higher accuracy and are especially better when modelling rare species distributions (Guisan and Thuiller, 2005; Ferrier, 2002; Gibson et al., 2007). Unfortunately, this requires estimating reliable pseudo-absence information (Guisan and Zimmermann, 2000; Coro et al., 2013c), which is not always possible. Today, large species occurrence data collections, sometimes referred to as species occurrence datasets (Jones et al., 2012; Casal et al., 2013) (SODs), expose high quality, expert-reviewed species observation records. For each record, these datasets usually provide information about (i) the recording time, (ii) the scientist who recorded the occurrence, (iii) the revision of the database record and (iv) the scientific survey this record belongs to. These surveys are the main source of information of large SODs, but they record species occurrences only along their routes (OBIS, 2015a; Vanden Berghe et al., 2010b,a; Tsontos and Kiefer, 2002; Ricard et al., 2010; Zeller et al., 2005; Halpin, 2009). These datasets usually store only observation records, and few examples of SODs storing also routes trajectories are available (Halpin et al., 2006), which would be useful when assessing absence locations.

Surveys usually focus on a limited taxonomic scope, for which the research vessel's scientific crew has expertise in identification. Large SODs usually do not report sampling absence information from surveys for a given species, but it is possible to reconstruct this information from locations where only other species' presence was reported. This reconstructed pseudo-absence information could be just due to the random geo-temporal sampling variation, possibly causing missed observations (e.g. individual undetected due to its behaviour or poor survey conditions), and could not reliably indicate habitat unsuitability or geographical absence in general. Further processing, in fact, is required to separate true absence from absence due to random sampling variation.

This paper presents a method to estimate absence locations for marine species, based on sampling-absences. In particular, we present a process to generate reliable pseudo-absence locations, i.e. absences that well approximate true absences. This process uses scientific survey data from an authoritative SOD containing a large amount of marine species observation records, i.e. the Ocean Biogeographic Information System (OBIS, Grassle, 2000; Vanden Berghe et al., 2010b; OBIS, 2015c). For each analysed species, our method (i) collects information from surveys that had the species in their target list, (ii) intersects and processes surveys' report locations to produce presence locations and sampling-absences, (iii) selects sampling-absence locations as those that are well separated from presence locations, i.e. not overlapping with presence locations according to a user-defined distance threshold.

In the paper, we take the Atlantic cod (*Gadus morhua* Linnaeus, 1758; Gadiformes: Gadidae) as a case study to demonstrate the reliability of our method. Additionally, we compare the performance of our process with another approach based on environmental envelopes. We use benchmark data from another authoritative SOD, the Global Biodiversity Information Facility (GBIF, Edwards et al., 2000; GBIF, 2014) for this comparison. Our process runs within a collaborative e-Infrastructure that publishes it as-a-Service (D4Science, 2015; Candela et al., 2015a; Coro et al., 2013a). The D4Science e-Infrastructure hosts this algorithm within a free-to-use platform using Cloud computing to execute processes (Coro et al., 2014a). This platform allows for (i) producing reliable pseudo-absence records, (ii) enriching them with environmental information, (iii) filtering on environmental values and (iv) using them in ecological niche models.

This paper is organized as follows: Section 2 reports about modelling methods to estimate species pseudo-absences. Section 3 gives the details of our approach along with its limitations. Section 4 reports a case study for the Atlantic cod and a statistical analysis on 550 aquatic species to quantitative estimate of the performance of our process. It also evaluates the sensitivity of our method to the values of two crucial input parameters. Finally, Section 5 contains summary considerations, including possible usages of our method in other models.

2. Overview

In this section, we briefly introduce species distribution models and describe their dependency on species presence and absence information. Then, we report methods to generate pseudo-absences. Finally, we discuss about the dependency of presence-only methods on the quality of data.

A variety of methods are currently used to build predictive species distribution models, which can be classified as those relying on presence-only versus presence/absence data (Pearson, 2012). Presence-only methods usually search for correlations between environmental parameters and observation records, whereas presence/absence approaches also use information about locations where the species of interest was not found. Presence/absence models have proven to improve their performance when reliable pseudo-absence information is available (Brotons et al., 2004; Guisan and Zimmermann, 2000).

Several methods are available to automatically estimate *pseudo*-absence locations, e.g. randomly taking locations (named "background points") in the area under analysis (Stockwell, 1999) to maximize relative differences with respect to known presence points (e.g. in the MaxEnt model, Elith et al., 2011), or using weighting criteria based on environmental information (Engler et al., 2004; Zaniewski et al., 2002). However, producing realistic

Table 1
Report of several methods to estimate species pseudo-absence locations, for further details see Pearson (2012), Thomas et al. (2002) and Franklin (2010).

Modeling method	Software	Response type	Response function
Generalized Additive Models (GAMs)	BIOMOD, GRASP, ECOSPAT	Scores, Categories, Boolean	Smoothing function
Multivariate adaptive regression splines (MARSs)	MARS	Score, Categories	Adaptive piecewise, Linear regression
Classification and regression trees	STATMOD ZONE, MART	Score, Categories	Divisive, monothetic decision rules from binary recursive partitioning
Density Surface Modelling (DSMs)	Distance 6.0	Score, Categories	Detection function
Artificial Neural Networks (ANNs)	NNETW	Categories	Non-linear decision function
AquaMaps	gCube	Boolean	Rule-based expert system

pseudo-absence locations requires much powerful models. Table 1 summarizes some powerful techniques to produce pseudo-absence locations.

In particular, Generalised Additive Models (GAMs) are used in both marine and terrestrial species distribution studies (Barlow et al., 2009; Leathwick, 1998; Brown, 1994; Bio et al., 1998; Frescino et al., 2001; Guisan et al., 2002; Lehmann et al., 2002; Platts et al., 2008; Lassalle et al., 2008) and have demonstrated to gain good performance. Their main drawback is that they are computationally intensive and their complexity depends on the number of explanatory variables included. When this number is high, GAMs require a rich dataset that adequately covers the variables ranges (Yee and Mitchell, 1991; Leathwick et al., 2006).

Multivariate Adaptive Regression Splines (MARSs) combine regression trees and spline fitting (Friedman, 2001) to model the non-linear relationship between species and environmental parameters (Elith and Leathwick, 2007; Hastie and Tibshirani, 1990). The advantage of MARSs is their speed and the possibility to manage a large number of model parameters (Leathwick et al., 2005). Their performance is comparable to GAMs' on freshwater species, but MARSs are more suited to model species with low prevalence (Muñoz and Felicísimo, 2004; Leathwick et al., 2006; Elith et al., 2006).

Another technique to produce pseudo-absences is Multiple Additive Regression Trees (MARTs), which enhances the combination between regression and classification trees and has been used in epidemiology (Friedman and Meulman, 2003) as well as in freshwater species distribution modelling (Cappo et al., 2005). The advantage of using classification trees is that their output allows reconstructing the role/weight of each parameter in the classification. The drawback is that they strongly factorise the parameters, which implies the sometimes weak assumption that they are independent of each other.

Distance sampling (Thomas et al., 2002; Buckland et al., 2001) is a widely used methodology for estimating animal abundance. Distance sampling methods infer species distribution based on the distances between a species' presence and survey lines or points. In particular, perpendicular distance from individuals to survey lines and radial distance to survey points are processed to estimate abundance in a certain area. These technique explicitly model the distance between an observer and the species, with the underlying hypothesis that the probability of detecting an animal decreases with its distance from the observer. Distance sampling models involve different functions (*detection functions*) to model the probability of detecting an animal given its distance from the survey routes. Based on this technique, Density Surface Modelling (DSM) techniques can estimate a population density in a certain area (Buckland, 2004; Schweder, 2007; Forney et al., 2015). DMSs have been used in combination with GAMs (Katsanevakis, 2007) and generalised linear models (a sub-case of GAMs, McCullagh, 1984) to enhance the modelling of spatial variation in animal density (Corkeron et al., 2011). The advantages and drawbacks of DMSs are the same of the distance sampling techniques (Barraclough, 2000). In particular, the advantages include: (i) the estimation of the absolute density of a population, (ii) accounting for missing subjects detection by different observers, (iii) they do not need the size of the sample area to be known. The drawbacks include: (i) the minimum number of required species detections is high (typically over 60), thus small populations are unlikely to be properly modelled, (ii) the method is inappropriate when the observer presence interferes with the animal presence, e.g. the North Island robin is attracted by observers, which may invalidate the assumptions made by the detection function. For this last reason, the data collection survey should be aware that this technique will be used to model the species distribution (Barraclough, 2000).

Other promising approaches employ Artificial Neural Networks (ANNs), which have been used in ecology (Olden et al., 2008) with applications to species distribution modelling (Özesmi and Özesmi, 1999; Hilbert and Ostendorf, 2001; Pearson et al., 2002, 2004; Dedecker et al., 2004; Garzon et al., 2006; Lippitt et al., 2008). These models may gain higher performance than others, but are difficult to train (Hastie et al., 2005) and it is not easy to reconstruct the role of each variable in the pseudo-absence location classification (Moisen and Frescino, 2002).

The AquaMaps model is an expert model that has been used to produce habitat-related pseudo-absence locations (Coro et al., 2013c). AquaMaps combines expert information with environmental envelopes (representing a species' habitat preference) to estimate a species distribution. In the same way as other environmental envelope models (Guisan and Thuiller, 2005; Pearson and Dawson, 2003), AquaMaps correlates the observed geographical distribution of a species to climatic variables. Low probability zones are associated to unsuitability areas and thus can be interpreted as habitat-related absence locations (Coro et al., 2015b, 2013c). Although AquaMaps has demonstrated good performance when many observations are available (Ready et al., 2010), its independent modelling of the environmental envelopes is a weak assumption and often requires maps to be manually reviewed (Coro et al., 2015b, 2013b).

Unlike presence/absence methods, presence-only approaches require collecting occurrence record sets that are as complete as possible (Berger, 1996; Phillips and Dudik, 2008; Elith et al., 2011). Records are usually taken from SODs (e.g. GBIF or OBIS) and are subject to quality control procedures also to ensure reliable coordinates of the observations. However, these SODs may contain biases because the surveys that provide observation data do not cover the ocean surfaces uniformly. This phenomenon is evident by looking at the global distribution map of the GBIF occurrence records (Fig. 1), where it is notable that the observations sampling concentrates along specific routes. It is also notable that this bias exists on land, along roads, but also in open ocean, along predefined oceanographic campaigns and shipping routes in some cases. Presence-only methods are very sensitive to these biases (Fromentin et al., 1993; Ibanez, 1982), and the most difficult task in estimating reliable presences/absences is to calculate the effort that was required to produce an observation/absence record. Sometimes this calculation is possible for reports belonging to scientific surveys, whereas it is hard for records coming from datasets that are the result of a less-structured sampling strategy, such as museum collections. In addition, a common problem with SODs is that the

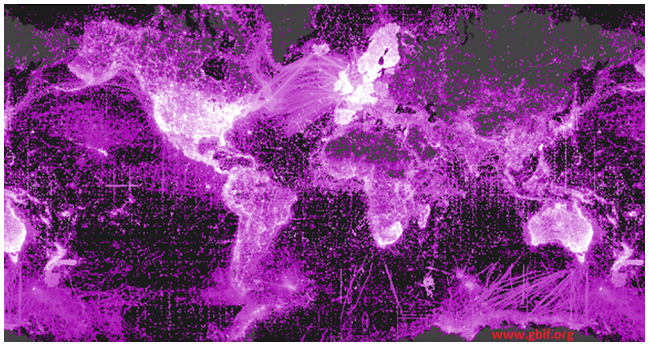

Fig. 1. Map of the occurrence records contained in GBIF, highlighting non-uniform survey sampling (taken from GBIF, 2015). Violet colours indicate occurrence records. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

observation platforms used during surveys are not equally well suited for all the target species, which could generate false absence reports. Statistical analysis on a large number of overlapping collections and surveys can partially overcome this issue, but cannot completely solve it. This kind of analysis is part of our method and will be presented in the next section.

3. Method

In this section, we describe the parameters, the input and the details of our process and we clarify its aims and limitations. Our process currently uses the OBIS SOD, the world's largest database on the occurrences of marine species. OBIS currently provides free access to 40 million observations of 115,000 marine species, integrated from more than 1600 data provided by nearly 500 institutions worldwide (OBIS, 2015b). The OBIS information system is based on a PostGIS database (Holl and Plum, 2009; Fujioka et al., 2012) and stores many individual data from scientific surveys, research projects, national monitoring programmes, museum collections and so on. OBIS hosts expert-reviewed data, where information about the observation (location, scientific name, person who identified the species etc.) has been validated by an expert of the species and the scientific name has been linked to its synonyms indexed on the database. Our algorithm accesses OBIS data through direct queries to the OBIS PostGIS database hosting observation records. Access is gained through the D4Science e-Infrastructure (Candela et al., 2015a). OBIS is connected to this e-Infrastructure and D4Science grants read-only access to the tables for authenticated users only.

3.1. The algorithm

Our algorithm generates pseudo-absences by evaluating a regularized grid built over existing surveys. It is reported in Fig. 2 and depicted in Fig. 3. It is developed using the R scripting language (R Core Team, 2015; Vanden Berghe, 2015) and is freely available for download through the D4Science e-Infrastructure as reported in Coro and Magliozzi (2015a). Even if the source code is open, the OBIS database requires authentication, thus the script can be executed only if OBIS grants access to its database,[1] otherwise it should be used as-a-Service through D4Science after free Web authentication.[2]

The main input of the script is a list of marine species scientific names in a text file. The user is prompted to specify a number of configuration parameters summarized in Table 2. In particular, the "Time Start" and "Time End" parameters define the time period for which the analysis will produce absence information. These parameters may allow investigating seasonal (e.g. filtering by month) and climatic (e.g. filtering by year) changes in species distributions, if the user is aware of the quality and the systematic methods used by the surveys. The "resolution" parameter defines the minimum size of the square cells that will be used to report pseudo-absence locations. This parameter is also used to separate non-overlapping absence and presence locations: it defines the minimum distance between a pseudo-absence produced by the algorithm and the presence locations selected from OBIS. The "geographical extent" parameter is the geographical bounding box on which the analysis will focus. The "observation frequency threshold" is a filtering threshold on the percentage of presence records reported by a survey for the target species: a survey is selected by the algorithm only if this percentage is over the threshold. The default value is 10%, which is suitable for global scale analyses, based on statistics on the OBIS database also investigated in a previous work (Coro et al., 2015c). Our procedure also requires connection parameters to a read-only instance of the OBIS database. We clarify the role of the above input parameters in the following algorithm description.

All data sources, including surveys, have a unique identifier in OBIS (resource id). In its first step, our algorithm prepares a dataset containing, for each survey, the overall number of reported presence locations. This dataset is identical for all the input species and does not need to be created on-the-fly. In fact, it can be prepared (cached) before the algorithm starts and only needs to be recreated whenever the backend OBIS database is updated.

For each species, the algorithm retrieves the unique OBIS-id representing its scientific name and all available synonyms. In the next step, the process retrieves all the identifiers of the surveys that reported the species at least once in the selected time period, along with the number of related distinct observation records. Each of these are records in OBIS may also refer to groups of individuals, but only the number of records is considered. At this point, the algorithm uses the "observation frequency threshold" to withhold only those surveys in which the species represented a sufficiently large fraction (i.e. set as 10% by default). The goal of this step is to minimize the inclusion of non-target species, e.g. chance observations, and possibly to consider locations where the species was actually searched for. This step is crucial for scientists who are not familiar with the original objective of the surveys and thus are unable to select data based on their relevance. With this step we implement an automated filter to identify these relevant surveys.

The algorithm then processes the selected surveys data in two steps: (i) it lists all the reported observations on a geo-referenced grid independently of the date of sampling and (ii) it merges all the locations having exactly the same coordinates. After this step, the algorithm classifies presence locations as locations with at least one record of the analysed species. In addition, this procedure classifies as absence locations those locations reported only for other species (i.e. not the species of interest), and then it produces a dense map with the distribution of absence and presence locations (Fig. 4).

The main issue with the pseudo-absence points in Fig. 4, is that they are too close to presence points. Nevertheless, our algorithm tries to estimate absence locations related to habitat unsuitability or geographical hindrance based on these points. Even if presence points come from different surveys or have been sampled in other ways, they possibly indicate areas which have suitable habitat for the species or that are geographically reachable by the species. Thus, in order to find reliable pseudo-absence locations, the algorithm searches for sampling absences which are *far* from presence locations, i.e. it keeps pseudo-absence locations well separated from presence locations. Equivalently, it applies a buffer of a given distance around all the presence locations. The appropriate

[1] Contacting them through the website http://www.iobis.org/node/179.
[2] http://services.d4science.org.

1. Open the input species TXT file
2. Prepare variables for the input parameters (res,extent,occ_percentage)
3. Open a database connection towards OBIS
4. Extract the total number of observations reported by each OBIS contributors (surveys)
5. For each species (sp) in the input file (one for each line)
   (a) Prepare an equirectangular-projected grid of square cells (grid) on the specified geographical bounding box (extent), having "res" resolution
   (b) Extract all the resources ids (res_ids) of the surveys containing observations for sp
   (c) Calculate the number of observations reported by each res_ids for the species
   (d) Reduce this dataset by taking only the res_ids having a relative number of occurrence for sp higher than occ_percentage (viable_res_ids)
   (e) Extract all the 0.1 degrees cells from OBIS, reported by the viable_res_ids (all_cells)
   (f) Extract all the 0.1 degrees cells from OBIS, reported by the viable_res_ids only for sp (presence_cells)
   (g) Merge the two datasets, summing the number of overlapping presence_cells and reporting 0 for the other cells (pres_abs_cells)
   (h) Aggregate information using the grid matrix: report 1 in a grid cell containing only presence records, and report -1 if it contains only absences. Report -2 for the cells containing both absences and presences and 0 for the ones reporting no point.
   (i) Select only the -1 cells from the grid (absence_cells)
   (j) Calculate the coordinates of the centres of the absence_cells
   (k) Delete the points falling on land
   (l) Produce an image of the distribution (Absences_<speciesname>_<res>.png)
   (m) Attach survey metadata to the absence records (for overlapping surveys report the most recent)
   (n) Write the absence_coordinates and the related metadata in an output CSV file (Absences_<speciesname>_<res>.csv)
6. Close the database connection

Fig. 2. Our algorithm, with steps described in natural language.

Table 2
The input parameters of our algorithm.

Parameter name	Description	Default value
Input list	A list of species scientific names in a text file	species.txt (e.g. http://goo.gl/XUKJVa)
Time start	The initial date of the analysis (e.g. 01-01-1980)	No time filter
Time end	The final date of the analysis (e.g. 01-01-2015)	No time filter
Resolution	The size of square areas that will be reported as absence locations	0.5°
Geograph. extent	The geographical bounding box of the analysis	Long [-180;180] Lat [-90;90]
Species occ. threshold	Minimum percentage of species records defining a viable biological survey	10%
Database connection parameters	Parameters to connect to the OBIS database	Manually configured in the script or provided by D4Science at runtime

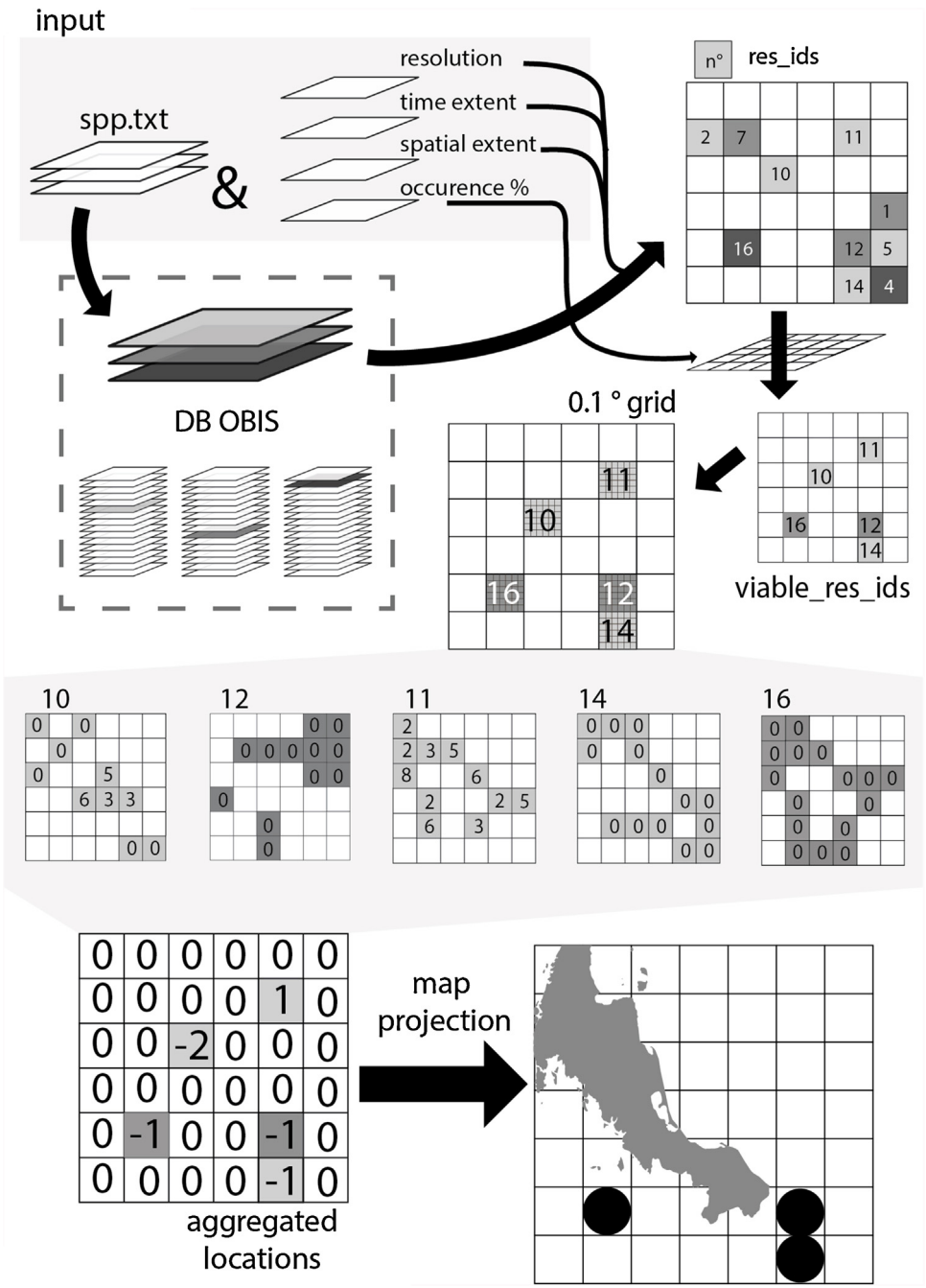

Fig. 3. Our algorithm represented as a diagram: "spp.txt" is the input file, represented along with the input parameter. The arrows represent the workflow of the algorithm and the enlargements report the processing in each 0.1° cell. The grey scale of the cells represents the percentage of occurrence, with darker colours representing higher percentage.

distance between presences and pseudo-absences depends on the geographic scale at which the user wants to produce pseudo-absence information. For example, distance from presence points goes as low as 0.001 degrees in the benchmark dataset we prepared for our algorithm evaluation, a distance that is negligible for global scale analyses. Thus, the best distance to separate presences and pseudo-absences depends on the scope of the user: for example, 0.5° (or 0.1°) distances can be sufficient for global scale niche models, whereas lower resolution would be better for regional scale models. In order to calculate the distance between pseudo-absence and presence locations we used an equirectangular projection for the points and calculated Euclidean distances in the R code instead of using the native PostGIS functions. This reduces the dependency of the algorithm from the OBIS database and allows easier modification of the code should another SOD be used.

The final step of our algorithm aggregates both presence and pseudo-absence points to the resolution requested by the user (the "resolution" parameter). During this process, the algorithm excludes the following types of aggregated locations: (i) those containing only presence points, (ii) those containing at least one presence and pseudo-absence point at the same time and (iii) locations falling on land. Thus, the remaining locations are those including pseudo-absence points only.

Fig. 5 shows an example of the output of our procedure, which consists of a CSV file reporting latitude and longitude points for the pseudo-absences along with their survey information and an image

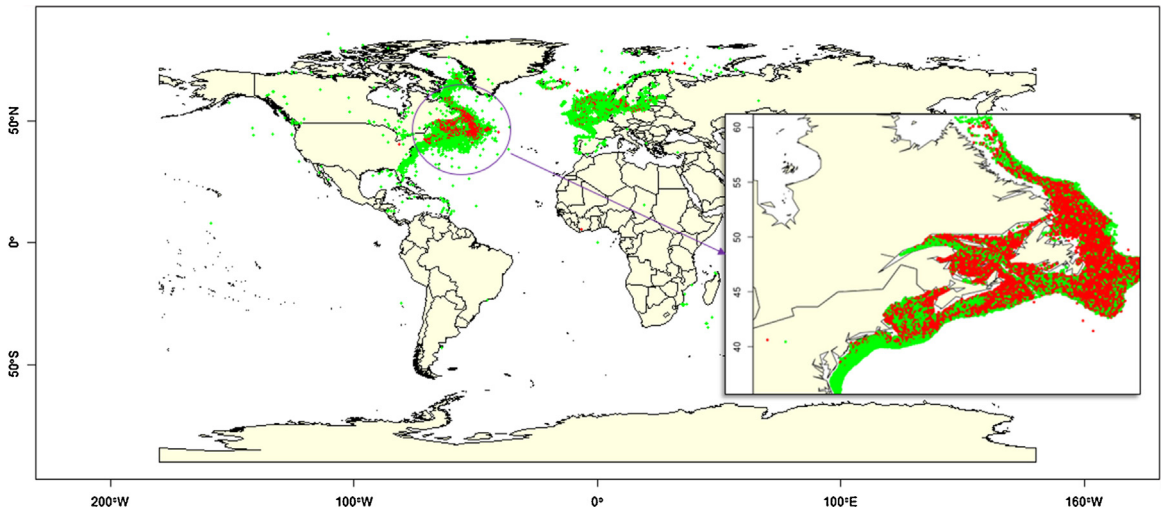

Fig. 4. Intermediate output of our process to estimate pseudo-absence locations. Darker/red colours indicate presence locations, brighter/green colours indicate pseudo-absences. The enlarged frame highlights the density of presence and pseudo-absence locations in the most abundant presence area.

file displaying the points at global scale. These files are produced for each species, for which at least one pseudo-absence location was found. The algorithm requires almost 1 hour to process 500 species on a CentOS 5.7 × 86 64 machine with 4 CPU cores and 8 GB of RAM. At the end of the process, the originating survey for each pseudo-absence location is listed, so that these can be cited in any future publication. The surveys can be less than those selected by the "observation frequency threshold" at the beginning of the process. Some surveys, in fact, could be discarded by the algorithm especially during the final aggregating-discarding step. Indicating the survey originating pseudo-absences allows users to select only those surveys of which they know reliability or relevance. This information is important especially when working at regional scales, because scientists who work at these scales possibly know the quality and the focus of the surveys. Thus, the output produced by our algorithm allows making *a posteriori* reasoning without overriding the input parameters of the experiments. However, the "observation frequency threshold" can be tuned to highlight the surveys that possibly involved experts of the species, and would reinforce a user's choice if it selected surveys the user deemed reliable.

3.2. Limitations

Our algorithm maximises the reliability of sampling-absences but overall it underestimates true absence locations. This is not a very strong limitation, since data reliability is more important in distribution models than data quantity (see Section 2). Underestimation is mainly due to the separation process that excludes high as well as low presence density locations without distinction, and to the tuning of the "observation frequency threshold" that may exclude relevant surveys. However, the algorithm can also take very old records into account, if the focus time period is large. This has the effect of overestimating presence points, but also means that our pseudo-absence locations will indicate places where the species has never been observed for the selected surveys, which enhances their reliability.

Changing the "observation frequency threshold" affects the number of produced pseudo-absence locations. This parameter needs fine tuning and we noticed that it is correlated to the required scale of the analysis, i.e. to the entire spatial domain of the study. Setting this threshold to a low value, produces more absences and may be useful when data resources can be manually checked after the processing. Generally, given a set of input parameters, our algorithm may not produce pseudo-absence locations for all the input species. Model's parameters can be indeed tuned to force the algorithm to be more specific in producing pseudo-absence locations for species. In the next section, we evaluate the sensitivity of the model to the parameters values with a case study. However, it is not easy to establish the *a priori* minimum number of presence locations that are sufficient to ensure high reliability of the pseudo-absence records. The performance of our method increases if the input parameters are tuned to select many surveys focussing on the same locations. In fact, the number of pseudo-absence locations extracted in the first part of the algorithm increases with the number of surveys available in the selected time frame. Additionally, the aggregation and filtering processes in the second part of the algorithm depend on the geographical extent of the selected surveys: these processes check if the analysis of different surveys produce contrasting results (i.e. if some survey never observed the species in a certain location whereas others did), thus they depend on the number of extracted surveys. Assessing these contrasting results in other ways would have required to evaluate the reliability of the surveys, which is difficult to automatise. This issue is also due to the lack of effort report by some data collections and surveys. For this reason, our algorithm discards locations which contain pseudo-absences and at least one presence record, and in this way it gives the same weight to all the surveys. Thus, our idea is also that, if many surveys are selected, the number of overlapping reports can account for missing effort information.

4. Evaluation

In this section, we first report a qualitative evaluation of our method using the Atlantic cod (*Gadus morhua* Linnaeus, 1758) as a case study. Then, we compare the performance of our model with another method by using a large benchmark survey SOD. Finally, we assess the sensitivity of our model to two of the input parameters.

4.1. Qualitative evaluation

We focussed on the Atlantic cod to evaluate our approach on a specific case study. *Gadus morhua* has a North Atlantic and Arctic distribution with concentrated populations. Schools have been observed both offshore and inshore (Cohen et al., 1990). OBIS contains 1000 surveys for this species, 101 of which satisfy a 10% "observation frequency threshold", although a maximum of 29 surveys originate pseudo-absence locations according to our

a.

Longitude	Latitude	Resource id	Resource name	Resource code	Abstract	Temporal scope	Last harvest date
-106	69	14687583	Atlantic Reference Centre Museum of Canadian Atlantic Organisms - Invertebrates and Fishes Data	arc	This is the OBIS version of the Atlantic Reference Centre museum database for Canadian Atlantic marine organisms. Specimens represent invertebrates from sponges to tunicates, and fishes. The ichthyoplankton collection is the most extensive, with complete holdings from many scientific broad-scale surveys. Geographic coverage is the Arctic to Cape Cod and the coast to the slope water. Temporally, most specimens were collected from the 1960s to the present.	1910 to 2013	28/10/2014 20:58
-59	59	14986587	DFO Maritimes - Groundfish and Small Pelagic Tagging database. 1953-1999	dfo_mar_fishtagging	Contains mark and recapture data for almost every Canadian fish tagging study conducted west of Newfoundland between 1953 and 1999. This dataset is the OBIS version of the fish tagging database records. The dataset contains the release and the recapture locations for the tagged fish.	1953 to 1999	05/11/2014 21:25
4	62	10555503	ICES Fish stomach contents dataset	ices_stomach_content_dat	The Fish predator/prey database (stomach database) contains data covering 11 countries. A total of 217 399 stomachs are reported in the North Sea data. Eight predator species were analysed and 854 NODC prey codes have been reported	1980 to 1991	13/12/2014 02:18
-10	48	11773180	ICES French Southern Atlantic Bottom Trawl Survey for commercial fish species	ices_datras_evhoe	The dataset includes age- and length-based catch per unit effort data for commercial fish species collected by the French trawl survey EVHOE.	1997 to 2010	13/12/2014 02:15
0	63	14989729	IMR Juvenile fish monitoring	imr_juveline_fish_monitori	The dataset contains juvenile fish (postlarvae) from the Barents Sea sampled during June-July 1977-91 (1980-91 in database, 1977-79 on files).	Could not be determined	13/12/2014 09:47
-1	65	14991522	IMR Macroplankton surveys	imr_macroplankton_survey	The Macroplankton dataset in the IMR plankton database contains macroplankton from the Norwegian and Barents Seas. The dataset includes small vertically migratory fish.	2007 to 2013	13/12/2014 01:58
-12	54	11781726	Irish Ground Fish Survey for commercial fish species	ices_datras_ie_igfs	The dataset includes age- and length-based catch per unit effort data for commercial fish species collected during the Irish Ground Fish trawl survey.	2003 to 2008	13/12/2014 05:00

b.

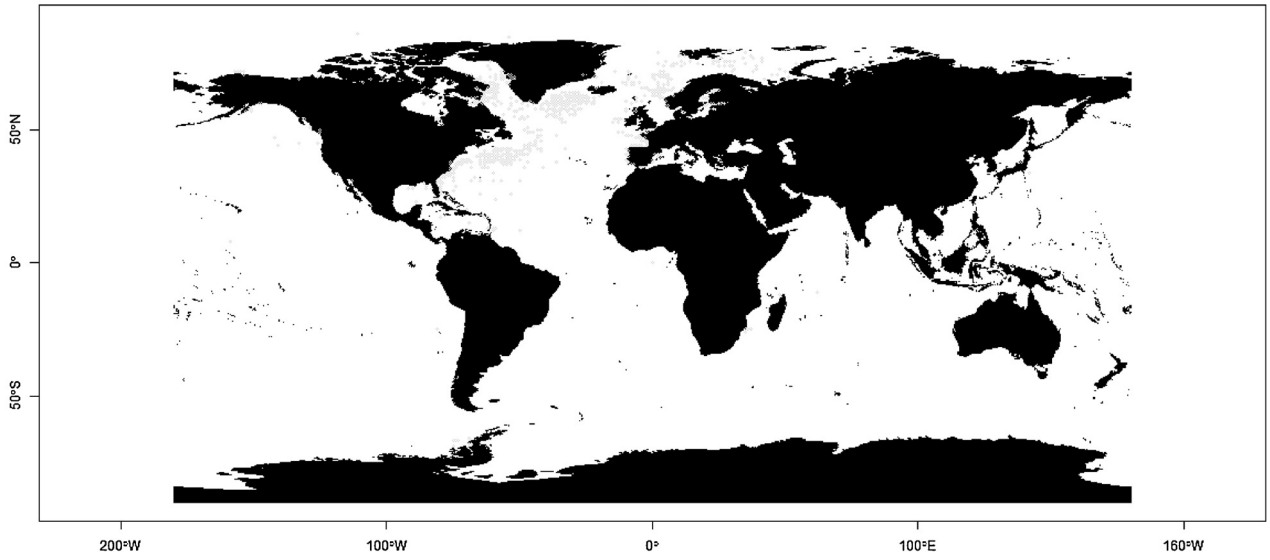

Fig. 5. Output of our algorithm: (a) a table reporting the coordinates of pseudo-absence locations along with metadata about their related surveys, (b) an image representing the global geographical distribution of the pseudo-absence locations (in grey).

process (see next section). Overall OBIS surveys for the Atlantic cod correspond to 128,435 observation records with almost uniform geographical distribution, with latitude ranging between −69 and 86 decimal degrees and longitude ranging between −159 and 146 decimal degrees. Using this large amount of data, we want to demonstrate that our produced pseudo-absences do not clash with presence locations reported in literature and are also close to some expert-reported absence locations. We selected this case as one example of output based on a large and widely distributed presence dataset. For this species, we built a raster file using Quantum GIS (Quantum GIS, 2011) (Fig. 6), which included (i) our generated pseudo-absence records (at 0.5° resolution, without time filtering and using an "observation frequency threshold" equal to 10%), (ii) presence areas extracted from several studies falling in the

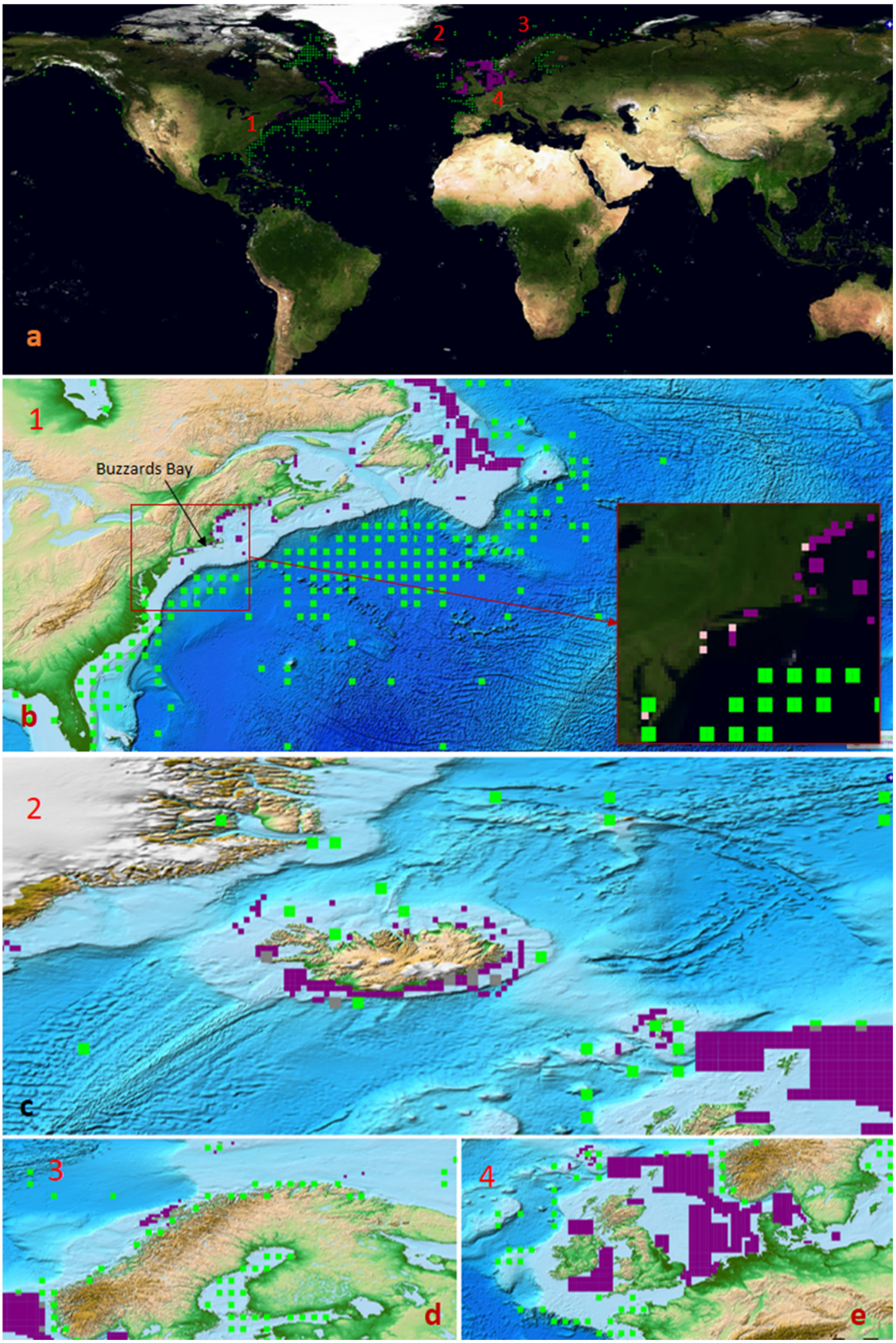

Fig. 6. Distribution of the pseudo-absence locations produced by our algorithm (green), presence areas from several literature studies (violet), pseudo-absence areas from one literature study (pink) and overlap between pseudo-absences and presences (grey) for *Gadus morhua*, with focus frames on (a) overall distribution of the locations, (b) East USA/Canadian coasts, (c) Iceland, (d) Norway and (e) North Sea areas. The red box indicates the region where the few pink points are located. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

geographical range of the OBIS presence records (Drinkwater, 2005; Righton et al., 2010; Ruzzante et al., 2001; Gregory et al., 1997; Neat et al., 2006; DeYoung and Rose, 1993; Knijn et al., 1993; Daan et al., 1990; Fox et al., 2008; Hedger et al., 2004; Svedäng et al., 2007; Olsen et al., 2004; Neuenfeldt et al., 2013), (iii) absence locations (derived from Fahay et al., 1999) and (iv) locations where presence areas overlap with our estimated absence locations. There is no work reporting absence information for the Atlantic cod at global scale, thus we used absence locations from Fahay et al. (1999) south of the Buzzards Bay to evaluate the proximity of our pseudo-absences to that area. Evaluating the exact overlap with these points is not meaningful because our pseudo-absences set is not meant to be complete.

A number of studies validate our pseudo-absence locations: for example, on the East coast of the United States of America (USA) and Canada (Fig. 6(b)) we report inshore presence locations from literature, which depend on Atlantic cod's preference for the salinity level in near-freshwater and oceanic waters (Smith and Page, 1996). Our pseudo-absence locations fall outside this area and are close to literature absence areas (see the magnified region in Fig. 6(b)). In particular, most of our pseudo-absence locations fall outside shallower continental slopes, which confirms that depth plays an important role for the distribution of *Gadus morhua* (Olsen et al., 2004). It is notable that our pseudo-absence points do not overlap with presence areas in shallower waters (e.g. from Florida to Maine). Indeed, the surveys used by our process include early life stage observations and no early life stage observation has been ever reported south of Buzzards Bay (41° 25′N, Fig. 6(b)). Furthermore, the species is uncommon at this stage in low salinity areas (Fahay et al., 1999).

Gadus morhua has been recorded along the South and South-East coasts of Iceland (Pálsson and Thorsteinsson, 2003), migration patterns for feeding are regularly detected among North Iceland, Greenland and Faroe waters (Neuenfeldt et al., 2013; Jónsson, 1996). South of Iceland, our pseudo-absence locations overlap with presence areas derived from literature (Fig. 6(c)). Indeed, there are no reports for the Atlantic cod in the OBIS surveys for this region.

In the Baltic Sea (Fig. 6(d)), the species has been reported only in inshore South-West coasts (Pihl and Ulmestrand, 1993; Knutsen et al., 2004), which agrees with our estimates. In the Norwegian Sea, our pseudo-absence locations fall mostly in fjords. Even if the surveys in OBIS cover this area, they do not report observations for the Atlantic cod, and this can be due to the ecological response by the Atlantic cod to the Norwegian ecosystem, as well as to the effect of local fishing pressure (Olsen et al., 2004). In the North Sea (Fig. 6(e)), the majority of our pseudo-absence locations do not overlap with presence locations indicated by other studies (Fox et al., 2008; Daan et al., 1990; Knijn et al., 1993; Hedger et al., 2004). In this area, Atlantic cod seems to prefer the temperature range of its southern presence locations (O'Brien et al., 2000). By looking at pseudo-absence locations north and south of United Kingdom (Fig. 6(e)), a correlation between the break of continental slopes and absences is notable, which is confirmed by other studies (Wheeler and Du Heaume, 1969).

4.1.1. Comparison with the AquaMaps distribution

Fig. 7 reports the AquaMaps distribution of *Gadus morhua* (The AquaMaps Consortium, 2014; Kaschner et al., 2008), where darker/red locations represent areas with higher probability of habitat suitability. The figure also shows the superposition between the AquaMaps distribution and our distribution of Atlantic cod's presence/absence locations. Brighter/yellow colours indicate low suitability and non-coloured areas indicate complete unsuitability. In this section we want to highlight the general overlap between the AquaMaps probability distribution and our pseudo-absence locations. Next section reports a quantitative evaluation of this similarity, which requires absence information to be properly inferred from AquaMaps. Instead, from a qualitative point of view, it is notable that our pseudo-absence records substantially overlap with unsuitability areas from AquaMaps in several locations, e.g. between Canada and Greenland (Labrador Sea), off the East coast of the USA and in the Baltic Sea. Disagreement is especially visible in the Norway fjords, where pseudo-absence locations are close to low-suitability areas but do not overlap with them. Indeed, AquaMaps calculates correlation between species observations and selected environmental characteristics and does not take other factors such as fishing pressure into account.

4.2. Quantitative evaluation

In order to quantitatively evaluate the overall performance of our approach, we compared our estimated pseudo-absence locations based on OBIS with a large number of presence locations from another source. We preferred using real occurrence records instead of simulated data because we could access a large amount of data through D4Science (see the work in Candela et al., 2015b) and produce results on data possibly containing real biases, which are difficult to simulate and categorize. Thus, our aim in this section, is to show the performance of our algorithm on a large amount of real data despite the unavoidable biases, rather than classifying these biases. This should be considered in the aim of publishing this process as-a-Service for real-world applications. As a benchmark species list, we selected the species contained in the Food and Agriculture Organization of the United Nations (FAO) Fact Sheets (FAO, 2015). FAO Fact Sheets provide key information on fisheries and aquaculture-related subjects on 550 aquatic species of particular commercial, biological and human interest. The species include mammals, fishes and crustaceans, and the catalogue is available freely online (FAO, 2015).

4.2.1. Comparison with GBIF data

Starting from the OBIS presence information, our process was able to produce 53,815 pseudo-absence locations for 280 species from the FAO list (Coro and Magliozzi, 2015b), using an "observation frequency threshold" equal to 10% and no time filtering. We also estimated pseudo-absence locations using the AquaMaps model (Kaschner et al., 2008). According to the indications in Coro et al. (2014b) and in Coro et al. (2013b), AquaMaps can produce pseudo-absence locations as 0.5° cells having either less than 0.2 or less than 0.5 (when used as a binary map, in which only 1 or 0 values are reported) probability of occurrence. Thus, we produced two sets of pseudo-absence records from AquaMaps using these two thresholds. With our benchmark species, these resulted in 1,000,613 global scale 0.5° locations using a 0.2 probability threshold, and in 2,224,542 locations using a 0.5 probability threshold.

We used locations from GBIF to test the performance of our method. In particular, we selected (equirectangularly projected) locations that were not included in OBIS, since OBIS data are ingested into GBIF (Kot et al., 2010). In order to assess the performance, we calculated how many pseudo-absence locations did not overlap (agree) with presence locations at a certain resolution. Our algorithm takes pseudo-absence locations distant from OBIS presence records, in other words it uses OBIS records as a training set. On the 208 benchmark species, our process used 353,531 presence records from OBIS survey data as training set, covering latitudes between −85 and 85 decimal degrees and longitudes between −180 and 180 decimal degrees. For the same 208 test species, GBIF hosts 1,738,366 survey data. However, we selected GBIF presence records that did not include these locations in the same spatial range of the OBIS surveys, using different distance thresholds to identify equivalent locations.

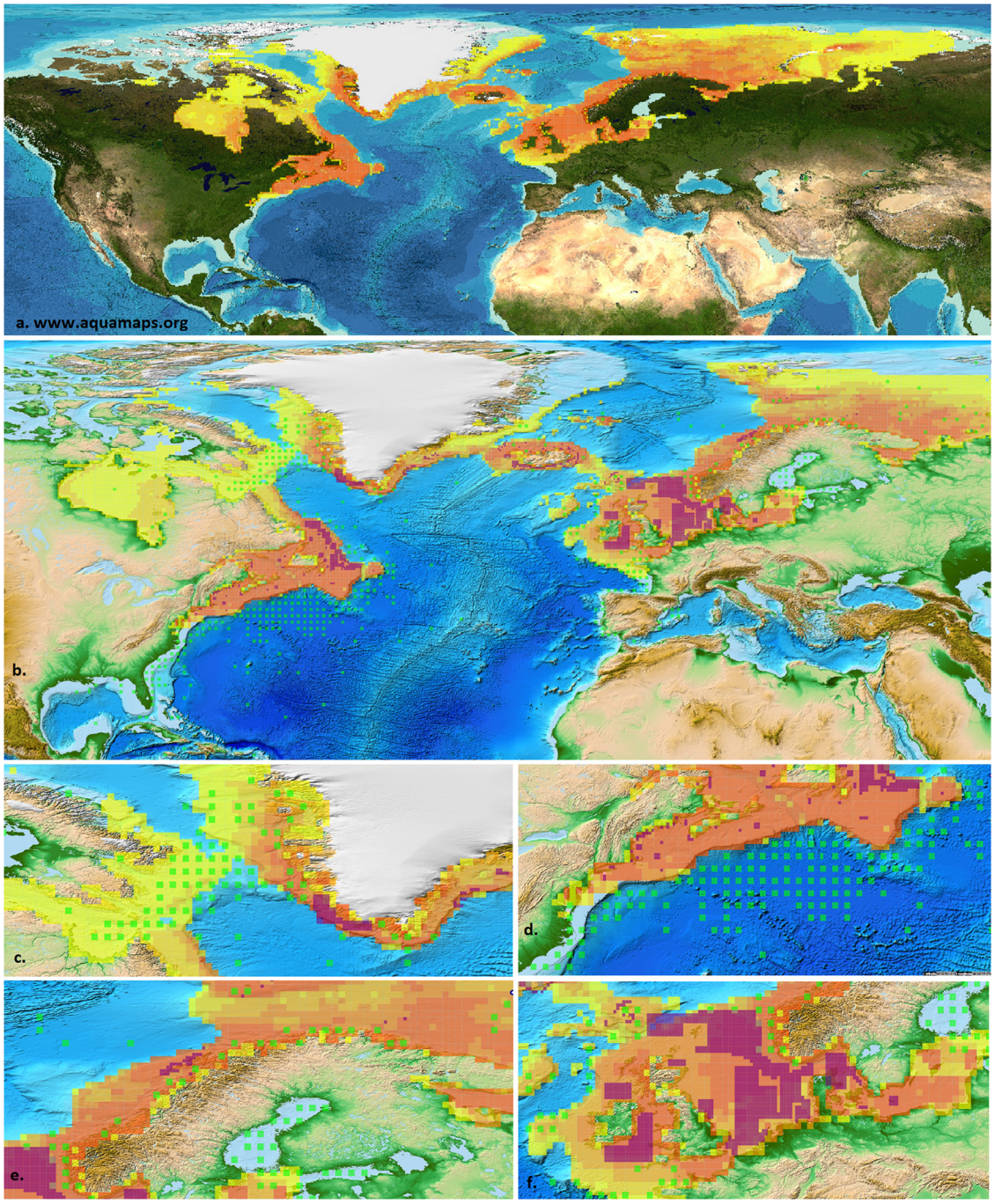

Fig. 7. Comparison with overlay between the AquaMaps distribution (a) and a dataset with pseudo-absence/presence locations for the Atlantic cod (*Gadus morhua*), with focus on different areas: North Atlantic ocean (b), Greenland (c), East USA/Canadian coasts (d), Norway and Baltic Sea (e) and North Sea (f). Green dots represent pseudo-absences automatically generated by our model and violet polygons represent presence locations. Red colours in the AquaMaps distribution refer to higher probability locations. In particular, red coloured areas in the AquaMaps distribution are higher probability zones (0.8–1) and yellow areas are lower probability (0.2–0.4) zones. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3 reports the intersection between several datasets. In particular, the first two sections compare the OBIS and the GBIF datasets and highlight that OBIS records are largely included in GBIF, also using different thresholds. This is also due to the fact that in our algorithm we used only a selection of the surveys available in OBIS. We used complementary GBIF records as our test set, using both the 0.1° and the 0.5° complementary points (1,635,802 and 1,401,122 points respectively).

Our pseudo-absences have less overlap with GBIF (96.1–99.6%) with respect to AquaMaps (93.6–94.6%). For AquaMaps comparisons we used only a 0.5° resolution, because this is the only possible resolution of the model. The fifth section in Table 3, indicates that simply merging these two datasets does not result in higher performance than the single AquaMaps-based method (94.6%).

Since AquaMaps uses an envelope-based approach as opposed to our survey-based process, in the last section of Table 3 we report the overlap between these two systems to evaluate their complementarity. This table section reports how many of our pseudo-absence locations are close to AquaMaps pseudo-absence

Table 3
The first two sections of this table report the overlap between presence locations of OBIS used by our algorithm to estimate pseudo-absences (training presence points) and presence locations from GBIF. They represent the degree of complementary information contained in GBIF. The other sections report the performance of our method and of pseudo-absences produced from AquaMaps. Automatic pseudo-absences are the ones generated by our process. The "AquaMaps probability threshold" is the probability threshold under which AquaMaps is forced to define 0.5° squared locations as pseudo-absences. Agreement is calculated as the percentage of non-overlapping points at the given resolution.

Performance comparison	
OBIS training presence points in GBIF	
Comparison resolution	Overlap
0.1°	75.2%
0.5°	95.3%
GBIF presence points in the OBIS training set	
Comparison resolution	Overlap
0.1°	5.9%
0.5°	19.4%
Automatic pseudo-absences vs GBIF test set	
Comparison resolution	Agreement
0.1°	96.1%
0.5°	99.6%
AquaMaps pseudo-absences vs GBIF test set (0.5° comp. res.)	
AquaMaps prob. threshold	Agreement
0.2	94.6%
0.5	93.6%
Merged AquaMaps and autom. pseudo-absences vs GBIF test set	
AquaMaps prob. threshold	Agreement
0.2	94.6%
Automatic pseudo-absences close to AquaMaps pseudo-absences (0.5° comp. res.)	
AquaMaps prob. threshold	Overlap
0.2	64.9%
0.5	69.7%

locations. The overlap percentage is either 64.9% or 69.7%, depending on different AquaMaps probability thresholds. This means that there is a certain degree of complementarity between the two systems; a substantial fraction of our pseudo-absence locations were not captured by AquaMaps, and vice-versa.

A confusion matrix in Table 4 summarises the overlaps between the datasets we have considered in this section. In particular, it reports the number of pseudo-absence locations shared by our model and by the AquaMaps process, as well as the number of presence locations in GBIF that were wrongly classified as pseudo-absences by AquaMaps and by our process. The number of GBIF data in the first row is higher than the one of our pseudo-absence locations, because we counted all the GBIF points that were at a distance lower than 0.5° from the pseudo-absences. In fact, Several points can be close to the same pseudo-absence location. If all the GBIF points were aggregated at 0.5° resolution (i.e. having a maximum of one GBIF point per 0.5° cell), that number became as low as 22,418.

The comparison between AquaMaps and our process is interesting because AquaMaps is supposed to produce a wider set of pseudo-absence locations. In fact, different from our process, AquaMaps does not discard its extracted pseudo-absence locations: the difference between our pseudo-absences and the AquaMaps pseudo-absences is that our algorithm maximizes the reliability of the pseudo-absences at the expense of completeness, whereas AquaMaps reports all the possible locations whose habitat is on average unsuitable (on annual basis) for the species to subsist. This indirectly means that the species is unlikely to be observed in the AquaMaps-produced pseudo-absence locations on annual average. However, the AquaMaps pseudo-absences that are complementary to ours, could involve locations where the species could indeed be observed in one single time instant. Instead, our pseudo-absences directly indicate low observation probability in those locations at every time, because surveys never found the species there and the locations are far from presence points. As a final remark, we want to highlight that the number and the reliability of the surveys for some species could be low. In these cases, AquaMaps is expected to perform better especially if the species presence is highly dependent on environmental conditions (Coro et al., 2015a). However, the reported statistical comparison on a quite large number of species should compensate these specific cases.

4.2.2. Evaluation of the sensitivity of the model to input parameters values

In order to evaluate the dependency of our model on the "resolution" and the "observation frequency threshold" input parameters, we selected one case study in which the algorithm produces many pseudo-absences belonging to several surveys. OBIS data for *Gadus morhua* have these characteristics, and Fig. 8 reports the change of the produced information at the variation of the parameters. In particular, Fig. 8(1) shows that the number of distinct surveys attached to the pseudo-absences strongly decreases at the increase of the "observation frequency threshold". With a 0.5° resolution, this number undergoes a 93.1% decrease, because it passes from 29 to 2. The number of surveys producing pseudo-absence records strongly changes using different resolution parameters, in particular it passes from 29 (at 0.5° resolution) to 3 (at 10° resolution). Fig. 8(2) shows that the number of produced pseudo-absences is still remarkable (308) when only pseudo-absences distant from presence points are selected (at 10° resolution), although this number undergoes a 96.4% decrease (from 8512 to 308) when resolution increases. The trend of the number of pseudo-absences strongly decreases when the "observation frequency threshold" increases. For example, when the value of this parameter passes from 0.5 to 0.8 at 0.5° resolution, the results is a 95.5% decrease of the number of pseudo-absences produced, whereas this decrease is 81.3% when passing from 0.05 to 0.1. Thus, the decrease is more than linear between 0.5 and 0.8, which demonstrates higher sensitivity in this range. In particular, sensitivity is higher in this range also when resolution changes, because passing from 1° to 5° results in a 97.6% decrease (from 123 to 3) of the number of pseudo-absences produced. However, sensitivity is lower for values of the threshold higher than 0.8.

Table 4
Performance comparison as a confusion matrix. In the first row, the table reports the number of pseudo-absence locations produced by our algorithm that are shared with the AquaMaps process (using two different probability thresholds) and those overlapping with GBIF presence locations. Note that several GBIF locations can be at less than 0.5° from one pseudo-absence location, thus overlapping GBIF points can be more than absence points. In the second row, the table reports the number of test set points in GBIF that are wrongly classified as pseudo-absences by the AquaMaps process (using two different probability thresholds).

	Total num. of locations	Num. of locations in common with		
		AquaMaps absence loc. (prob. thr. 0.2)	AquaMaps absence loc. (prob. thr. 0.5)	GBIF (0.5° test set)
Automatic absences (0.5° res.)	53,815 (absences)	34,926	37,509	56,045
GBIF (0.5° test set)	1,401,122 (presences)	75,661	89,672	–

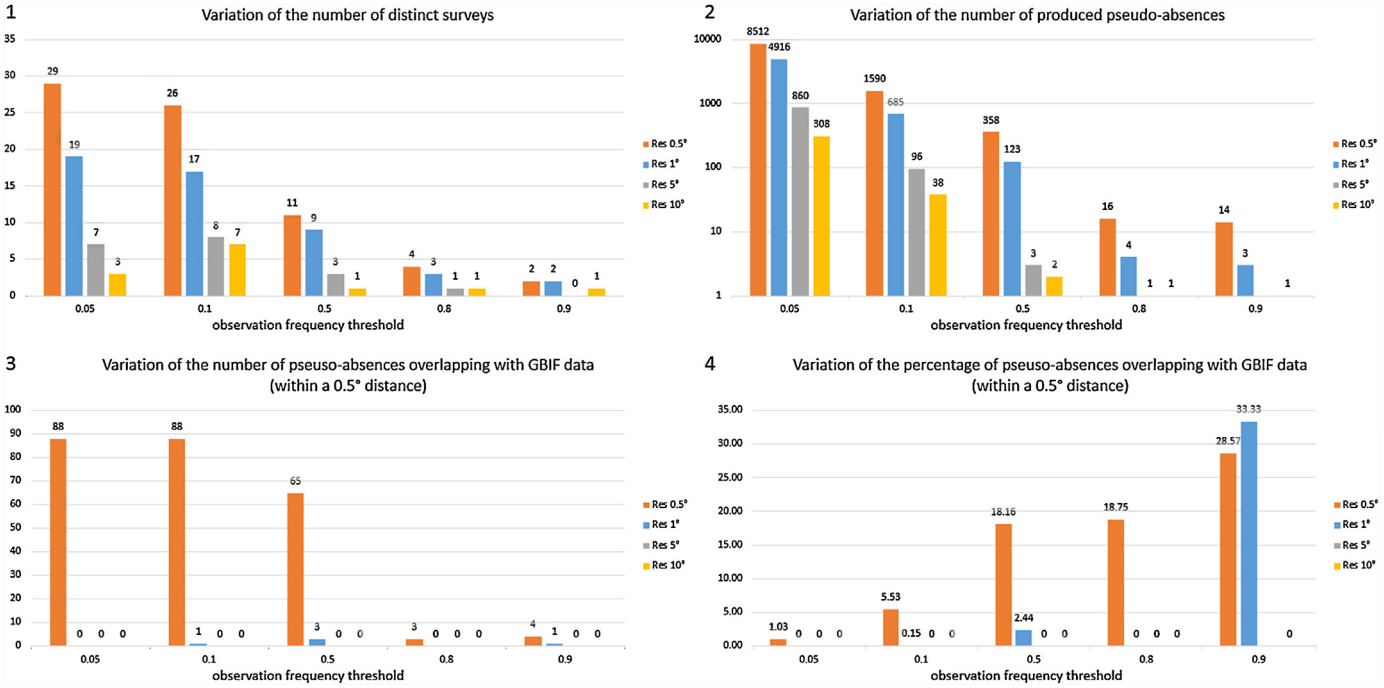

Fig. 8. Variation of quantities produced by our model, at the change of the "resolution" (Res) and of the "observation frequency threshold" input parameters. The following quantities are reported: number of distinct surveys attached to pseudo-absence records (1), number of pseudo-absences (2), number (3) and percentage (4) of pseudo-absences overlapping with GBIF observations. Chart number 2 is in logarithmic scale.

Fig. 8(3) reports the number of overlapping locations between the produced pseudo-absences and GBIF records for *Gadus morhua*. In particular, we selected GBIF points in the Atlantic Ocean that were at least 0.5° distant from OBIS records. We used this threshold as a reasonable tolerance distance for an analysis in the Atlantic Ocean. Most of the overlap is found when a 0.5° resolution is used, i.e. when pseudo-absences are close to OBIS presence points. Fig. 8(4) shows the same overlap with respect to the number of produced pseudo-absences, i.e. as a percentage of the number of pseudo-absences estimated. This chart highlights that using lower values for "resolution" and "observation frequency threshold" results in better performance in this case, even if the absolute number of overlaps is higher. Generally, percentages are all reasonably low, thus our algorithm has good performance (i.e. poor overlap with GBIF presence data).

This example shows that the algorithm is particularly sensible to the "resolution" and the "observation frequency threshold", although this sensitivity is higher in a certain range of the "observation frequency threshold" parameter. These parameters can be tuned, for example, to increase the percentage of reliable pseudo-absences at the expense of the number of selected surveys and vice-versa.

5. Conclusions

In this paper, we have described a process to reliably produce true absence locations for marine species. Our process uses observation reports from OBIS (e.g. scientific survey data, museum collections etc.) and we have confirmed the good performance of our method using GBIF data. Additionally, we have shown its sensitivity to two crucial input parameters that users can change to fine tune the output. We offer effective support for species distribution modelling by expanding the data available to presence/absence approaches. Although the set of pseudo-absence locations produced by our model is not complete, it aims at increasing the quality/reliability of the produced data, which is a crucial aspect for presence/absence species distribution models (see Section 2).

As further enhancement of our process, we plan to produce a ranked list of pseudo-absence locations, where a weight is associated to each location based on factors like frequency of survey coverage, fishing pressure, habitat suitability and other information that could be available from the data provider. Indeed, we have demonstrated that our pseudo-absence records can contain complementary information with respect to habitat suitability (see Section 4.2), thus adding habitat indication would effectively add more information to our output. Furthermore, we also plan to extend the method to make it applicable also to other SODs than OBIS. This will require understanding which information is shared among different providers of occurrence records, possibly relying on our previous investigations in this domain, e.g. Candela et al. (2015b). In particular, we will make the information extraction process more "modular" in our implementation, for example we will build separate connectors to extract specific information objects from the data providers (e.g. species synonyms, observations locations, data collections etc.) and later assembling them.

Our algorithm is parametric and can be tuned to manage regional as well as global scale scenarios. It also runs within the D4Science e-Infrastructure, which publishes it as-a-Service (Candela et al., 2015a; Coro et al., 2013a), reduces calculation time and promotes experiments re-usability and reproducibility. The benefits brought by this platform and the modality to integrate an algorithm with it are described in Coro et al. (2014a) and in Candela et al. (2013).

Apart from species distribution models, our pseudo-absence locations could be relevant also to climate change analyses, where species distribution models have an important role (Araújo et al., 2005; Cheung et al., 2009, 2010). Change in time of species absence can be also used to extract biodiversity indicators, for example to evaluate species commonness in a certain area (Coro et al., 2015c). In these applications, producing reliable output for a certain time frame is critical, since shifts in spatial patterns can be caused by changes in survey sampling patterns rather than effective shifts in species distributions. Thus, users should rely on the information attached to the output to *a posteriori* select reliable surveys,

or should *a priori* exclude surveys by tuning the "observation frequency threshold".

Species absence in a certain area could be also used by policy makers, for example in stock assessment models estimating safe biological limits for fishes (Froese et al., 2014). Stock assessment models, in fact, rely on metrics of fishing catch, biomass, on life history traits (e.g. growth, length-at-age etc.) and presence reports. Including pseudo-absence information estimated from surveys may enrich the information available to these systems and enhance their performance. Our future work will further explore this possibility.

Acknowledgments

The reported work has been partially supported by the i-Marine project (FP7 of the European Commission, INFRASTRUCTURES-2011-2, Contract No. 283644) and by the Giovanisi project of the Presidency of the Tuscan Regional Government. The authors thank Dr. Paula Moreno and Prof. Carlo Cerrano for their precious comments that helped improving the paper readability.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ecolmodel.2015.12.008.

References

Araújo, M.B., Pearson, R.G., Thuiller, W., Erhard, M., 2005. Validation of species-climate impact models under climate change. Glob. Change Biol. 11 (9), 1504–1513.

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

Barlow, J., Ferguson, M.C., Becker, E.A., Redfern, J.V., Forney, K.A., Vilchis, I.L., Fiedler, P.C., Gerrodette, T., Ballance, L.T., 2009. Predictive modeling of cetacean densities in the eastern Pacific Ocean. US Department of Commerce, National Oceanic and Atmospheric Administration, National Marine Fisheries Service, Southwest Fisheries Science Center.

Barraclough, R.K., 2000. Distance sampling: a discussion document produced for the department of conservation. In: Science & Research Internal Report 175, Department of Conservation, P.O. Box 10-420, Wellington, New Zealand, pp. 1–26.

Berger, A., 1996. A brief MaxEnt tutorial, Available at: http://www-2.cs.cmu.edu/afs/cs/user/aberger/www/html/tutorial/tutorial.html.

Bio, A., Alkemade, R., Barendregt, A., 1998. Determining alternative models for vegetation response analysis: a non-parametric approach. J. Veg. Sci. 9 (1), 5–16.

Brotons, L., Thuiller, W., Araujo, M.B., Hirzel, A.H., 2004. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. Ecography 27, 437–448.

Brown, D.G., 1994. Predicting vegetation types at treeline using topography and biophysical disturbance variables. J. Veg. Sci. 5 (5), 641–656.

Buckland, S.T., 2004. Advanced Distance Sampling. Oxford University Press.

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D., Thomas, L., 2001. Introduction to Distance Sampling Estimating Abundance of Biological Populations. Oxford University Press.

Candela, L., Castelli, D., Coro, G., Lelii, L., Mangiacrapa, F., Marioli, V., Pagano, P., 2015a. An infrastructure-oriented approach for supporting biodiversity research. Ecol. Inf. 26, 162–172.

Candela, L., Castelli, D., Coro, G., Lelii, L., Mangiacrapa, F., Marioli, V., Pagano, P., 2015b. An infrastructure-oriented approach for supporting biodiversity research. Ecol. Inf. 26, 162–172.

Candela, L., Castelli, D., Coro, G., Pagano, P., Sinibaldi, F., 2013. Species distribution modeling in the cloud. Concurr. Comput.

Cappo, M., De'ath, G., Boyle, S., Aumend, J., Olbrich, R., Hoedt, F., Perna, C., Brunskill, G., 2005. Development of a robust classifier of freshwater residence in barramundi (*Lates calcarifer*) life histories using elemental ratios in scales and boosted regression trees. Mar. Freshw. Res. 56 (5), 713–723.

Casal, C.M.V., Kesner-Reyes, K., Palomares, M.L.D., Bailly, N., Froese, R., 2013. Predicting species distribution using fishbase, sealifebase and aquamaps. FishBase http://www.fishbase.us/Download/CBD_FB_SLB_AqMaps.pdf.

Cheung, W.W., Lam, V.W., Sarmiento, J.L., Kearney, K., Watson, R., Pauly, D., 2009. Projecting global marine biodiversity impacts under climate change scenarios. Fish Fish. 10 (3), 235–251.

Cheung, W.W., Lam, V.W., Sarmiento, J.L., Kearney, K., Watson, R., Zeller, D., Pauly, D., 2010. Large-scale redistribution of maximum fisheries catch potential in the global ocean under climate change. Glob. Change Biol. 16 (1), 24–35.

Cohen, D., Inada, T., Iwamoto, T., Scialabba, N., 1990. An annotated and illustrated catalogue of cods, hakes, grenadiers and other gadiform fishes known to date. FAO Species Cat. 10, 442.

Corkeron, P.J., Collins, G.M.T., Findlay, K., Willson, A., Baldwin, R., 2011. Spatial models of sparse data to inform cetacean conservation planning: an example from Oman. Endanger. Spec. Res. 15 (1), 39–52.

Coro, G., Candela, L., Pagano, P., Italiano, A., Liccardo, L., 2014a. Parallelizing the execution of native data mining algorithms for computational biology. Concurr. Comput.

Coro, G., Gioia, A., Pagano, P., Candela, L., 2013a. A service for statistical analysis of marine data in a distributed e-infrastructure. Boll. Geofis. Teor. Appl. 54 (1), 68–70.

Coro, G., Magliozzi, C., 2015a. A script to estimate realistic absence and presence records from OBIS, Downloadable via the high-availability CNR data provisioning system at http://data.d4science.org/uri-resolver/id?filename=absencesspecieslistprod.r&smp-id=551020dbe4b0ffc120bc4837&contenttype=text%2fx-rsrc.

Coro, G., Magliozzi, C., 2015b. List of 280 benchmark species to test the absence locations estimation algorithm, Downloadable via the high-availability CNR data provisioning system at http://goo.gl/98ZR4H.

Coro, G., Magliozzi, C., Ellenbroek, A., Kaschner, K., Pagano, P., 2015a. Automatic classification of climate change effects on marine species distributions in 2050 using the aquamaps model. Environ. Ecol. Stat., 1–26, http://dx.doi.org/10.1007/s10651-015-0333-8.

Coro, G., Magliozzi, C., Ellenbroek, A., Pagano, P., 2015b. Improving data quality to build a robust distribution model for *Architeuthis dux*. Ecol. Model. 305, 29–39.

Coro, G., Pagano, P., Ellenbroek, A., 2013b. Automatic procedures to assist in manual review of marine species distribution maps. In: Adaptive and Natural Computing Algorithms. Springer, pp. 346–355.

Coro, G., Pagano, P., Ellenbroek, A., 2013c. Combining simulated expert knowledge with neural networks to produce ecological niche models for *Latimeria chalumnae*. Ecol. Model. 268, 55–63.

Coro, G., Pagano, P., Ellenbroek, A., 2014b. Comparing heterogeneous distribution maps for marine species. GIScience Remote Sens. 51 (5), 593–611.

Coro, G., Webb, T.J., Appeltans, W., Bailly, N., Cattrijsse, A., Pagano, P., 2015c. Classifying degrees of species commonness: north sea fish as a case study. Ecol. Model. 312, 272–280.

D4Science, 2015. The D4Science services portal. services.d4science.org.

Daan, N., Bromley, P., Hislop, J., Nielsen, N., 1990. Ecology of north sea fish. Neth. J. Sea Res. 26 (2), 343–386.

Dedecker, A.P., Goethals, P.L., Gabriels, W., De Pauw, N., 2004. Optimization of artificial neural network (ANN) model design for prediction of macroinvertebrates in the zwalm river basin (Flanders, Belgium). Ecol. Model. 174 (1), 161–173.

DeYoung, B., Rose, G., 1993. On recruitment and distribution of Atlantic cod (*Gadus morhua*) off Newfoundland. Can. J. Fish. Aquat. Sci. 50 (12), 2729–2741.

Drinkwater, K.F., 2005. The response of Atlantic cod (*Gadus morhua*) to future climate change. ICES J. Mar. Sci. 62 (7), 1327–1337.

Edwards, J.L., Lane, M.A., Nielsen, E.S., 2000. Interoperability of biodiversity databases: biodiversity information on every desktop. Science 289 (5488), 2312–2314.

Elith, J., Graham, C.H., Anderson, R.P., et al., 2006. Novel methods improve prediction of species distributions from occurrence data. Ecography 29, 129–151.

Elith, J., Leathwick, J., 2007. Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. Divers. Distrib. 13 (3), 265–275.

Elith, J., Leathwick, J.R., 2009. Species distribution models: ecological explanation and prediction across space and time. Ann. Rev. Ecol. Evol. Syst. 40 (1), 677.

Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E., Yates, C.J., 2011. A statistical explanation of MaxEnt for ecologists. Divers. Distrib. 17 (January (1)), 43–57.

Engler, R., Guisan, A., Rechsteiner, L., 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. J. Appl. Ecol. 41 (2), 263–274.

Fahay, M., Berrien, P., Johnson, D., Morse, W., 1999. Essential fish habitat source document: Atlantic cod, *Gadus morhua*, life history and habitat characteristics. NOAA Tech. Mem. NMFS-NE 124, 41.

FAO, 2015. Fact Sheets, Available at: http://www.fao.org/newsroom/en/facts/index.html.

FAO, 2015. The Fact Sheet species of the Food and Agriculture Organization of the United Nations. http://www.fao.org/fishery/species/search/en.

Ferrier, S., 2002. Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? Syst. Biol. 51 (2), 331–363.

Forney, K.A., Becker, E.A., Foley, D.G., Barlow, J., Oleson, E.M., 2015. Habitat-based models of cetacean density and distribution in the central north pacific. Endanger. Species Res. 27 (1), 1–20.

Fox, C.J., Taylor, M., Dickey-Collas, M., Fossum, P., Kraus, G., Rohlf, N., Munk, P., van Damme, C.J., Bolle, L.J., Maxwell, D.L., et al., 2008. Mapping the spawning grounds of north sea cod (*Gadus morhua*) by direct and indirect means. Proc. R. Soc. B: Biol. Sci. 275 (1642), 1543–1548.

Franklin, J., 2010. Mapping Species Distributions: Spatial Inference and Prediction. Cambridge University Press.

Frescino, T.S., Edwards, T.C., Moisen, G.G., 2001. Modeling spatially explicit forest structural attributes using generalized additive models. J. Veg. Sci. 12 (1), 15–26.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat., 1189–1232.
Friedman, J.H., Meulman, J.J., 2003. Multiple additive regression trees with application in epidemiology. Stat. Med. 22 (9), 1365–1381.
Froese, R., Coro, G., Kleisner, K., Demirel, N., 2014. Revisiting safe biological limits in fisheries. Fish Fish.
Fromentin, J.-M., Ibanez, F., Legendre, P., 1993. A phytosociological method for interpreting plankton data. Mar. Ecol. Prog. Ser. 93, 285–306.
Fujioka, E., Berghe, E.V., Donnelly, B., Castillo, J., Cleary, J., Holmes, C., McKnight, S., Halpin, P., 2012. Advancing global marine biogeography research with open-source GIS software and cloud computing. Trans. GIS 16 (2), 143–160.
Garzon, M.B., Blazek, R., Neteler, M., De Dios, R.S., Ollero, H.S., Furlanello, C., 2006. Predicting habitat suitability with machine learning models: the potential area of *Pinus sylvestris* L. in the Iberian peninsula. Ecol. Model. 197 (3), 383–393.
GBIF, 2014. Global Biodiversity Information Facility (GBIF). gbif.org.
GBIF, 2015. GBIF occurrences distribution. http://www.gbif.org/occurrence.
Gibson, L., Barrett, B., Burbidge, A., 2007. Dealing with uncertain absences in habitat modelling: a case study of a rare ground-dwelling parrot. Divers. Distrib. 13, 704–713.
Grassle, J., 2000. The Ocean Biogeographic Information System (OBIS): an on-line, worldwide atlas for accessing, modeling and mapping marine biological data in a multidimensional geographic context. Oceanography 13 (3), 5–7.
Gregory, R.S., Anderson, J.T., Dalley, E.L., 1997. Distribution of juvenile Atlantic cod (*Gadus morhua*) relative to available habitat in Placentia bay, Newfoundland. Northwest Atl. Fish. Organ. 29, 3–12.
Guisan, A., Edwards, T.C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol. Model. 157 (2), 89–100.
Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecol. Lett. 8 (9), 993–1009.
Guisan, A., Zimmermann, N., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135, 147–186.
Halpin, P., Read, A., Best, B., Hyrenbach, K., Fujioka, E., Coyne, M., Crowder, L., Freeman, S., Spoerri, C., et al., 2006. OBIS-SEAMAP: developing a biogeographic research data commons for the ecological studies of marine mammals, seabirds, and sea turtles. Mar. Ecol. Prog. Ser. 316 (23), 246.
Halpin, P.N., 2009. obis-seamap. Oceanography 22 (2), 104.
Hastie, T., Tibshirani, R., Friedman, J., Franklin, J., 2005. The elements of statistical learning: data mining, inference and prediction. Math. Intell. 27 (2), 83–85.
Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models, vol. 43. CRC Press.
Hedger, R., McKenzie, E., Heath, M., Wright, P., Scott, B., Gallego, A., Andrews, J., 2004. Analysis of the spatial distributions of mature cod (*Gadus morhua*) and haddock (*Melanogrammus aeglefinus*) abundance in the north sea (1980–1999) using generalised additive models. Fish. Res. 70 (1), 17–25.
Hilbert, D.W., Ostendorf, B., 2001. The utility of artificial neural networks for modelling the distribution of vegetation in past, present and future climates. Ecol. Model. 146 (1), 311–327.
Holl, S., Plum, H., 2009. Postgis. GeoInformatics 3 (2009), 34–36.
Ibanez, F., 1982. A new application of information theory for the description of a plankton chronological series). J. Plankton Res. 4 (3), 619–632.
Jones, M.C., Dye, S.R., Pinnegar, J.K., Warren, R., Cheung, W.W., 2012. Modelling commercial fish distributions: prediction and assessment using different approaches. Ecol. Model. 225, 133–145.
Jónsson, J., 1996. Tagging of cod (*Gadus morhua*) in Icelandic waters 1948–1986: tagging of haddock (*Gadus aeglefinus*) in Icelandic waters 1953–1965. Hafrannsóknastofnunin.
Kaschner, K., Ready, J., Agbayani, E., Rius, J., Kesner-Reyes, K., Eastwood, P., South, A., Kullander, S., Rees, T., Close, C., et al., 2008. Aquamaps: Predicted range maps for aquatic species, Available at: http://www.aquamaps.org, Version 8, 2010.
Katsanevakis, S., 2007. Density surface modelling with line transect sampling as a tool for abundance estimation of marine benthic species: the pinna nobilis example in a marine lake. Mar. Biol. 152 (1), 77–85.
Knijn, R.J., Boon, T.W., Heessen, H.J., Hislop, J.R., 1993. Atlas of north sea fishes. ICES Cooper. Res. Report 194, 268.
Knutsen, H., André, C., Jorde, P.E., Skogen, M.D., Thuróczy, E., Stenseth, N.C., 2004. Transport of north sea cod larvae into the skagerrak coastal populations. Proc. R. Soc. Lond. B: Biol. Sci. 271 (1546), 1337–1344.
Kot, C.Y., Fujioka, E., Hazen, L.J., Best, B.D., Read, A.J., Halpin, P.N., 09 2010. Spatio-temporal gap analysis of obis-seamap project data: assessment and way forward. PLoS ONE 5 (9), e12990, http://dx.doi.org/10.1371/journal.pone.0012990.
Lassalle, G., Béguer, M., Beaulaton, L., Rochard, E., 2008. Diadromous fish conservation plans need to consider global warming issues: an approach using biogeographical models. Biol. Conserv. 141 (4), 1105–1118.
Leathwick, J., Elith, J., Hastie, T., 2006. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. Ecol. Model. 199 (2), 188–196.
Leathwick, J., Rowe, D., Richardson, J., Elith, J., Hastie, T., 2005. Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. Freshw. Biol. 50 (12), 2034–2052.
Leathwick, J.R., 1998. Are new zealand's nothofagus species in equilibrium with their environment? J. Veg. Sci. 9 (5), 719–732.
Lehmann, A., Overton, J.M., Leathwick, J.R., 2002. Grasp: generalized regression analysis and spatial prediction. Ecol. Model. 157 (2), 189–207.
Lippitt, C.D., Rogan, J., Toledano, J., Sangermano, F., Eastman, J.R., Mastro, V., Sawyer, A., 2008. Incorporating anthropogenic variables into a species distribution model to map gypsy moth risk. Ecol. Model. 210 (3), 339–350.
McCullagh, P., 1984. Generalized linear models. Eur. J. Oper. Res. 16 (3), 285–292.
Moisen, G.G., Frescino, T.S., 2002. Comparing five modelling techniques for predicting forest characteristics. Ecol. Model. 157 (2), 209–225.
Muñoz, J., Felicísimo, Á.M., 2004. Comparison of statistical methods commonly used in predictive modelling. J. Veg. Sci. 15 (2), 285–292.
Neat, F.C., Wright, P.J., Zuur, A.F., Gibb, I.M., Gibb, F.M., Tulett, D., Righton, D.A., Turner, R.J., 2006. Residency and depth movements of a coastal group of atlantic cod (*Gadus morhua* l.). Mar. Biol. 148 (3), 643–654.
Neuenfeldt, S., Righton, D., Neat, F., Wright, P., Svedäng, H., Michalsen, K., Subbey, S., Steingrund, P., Thorsteinsson, V., Pampoulie, C., et al., 2013. Analysing migrations of atlantic cod *Gadus morhua* in the north-east atlantic ocean: then, now and the future. J. Fish Biol. 82 (3), 741–763.
OBIS, 2015a. Quality control of OBIS data. http://www.iobis.org/node/47.
OBIS, 2015b. Statistics of the OBIS dataset. http://www.iobis.org/about/statistics.
OBIS, 2015c. The OBIS Web Portal search interface. http://iobis.org/mapper/.
O'Brien, C.M., Fox, C.J., Planque, B., Casey, J., 2000. Fisheries: climate variability and north sea cod. Nature 404 (6774), 142.
Olden, J.D., Lawler, J.J., Poff, N.L., 2008. Machine learning methods without tears: a primer for ecologists. Q. Rev. Biol. 83 (2), 171–193.
Olsen, E., Knutsen, H., Gjøsæter, J., Jorde, P., Knutsen, J., Stenseth, N., 2004. Life-history variation among local populations of atlantic cod from the norwegian skagerrak coast. J. Fish Biol. 64 (6), 1725–1730.
Özesmi, S.L., Özesmi, U., 1999. An artificial neural network approach to spatial habitat modelling with interspecific interaction. Ecol. Model. 116 (1), 15–31.
Pálsson, Ó.K., Thorsteinsson, V., 2003. Migration patterns, ambient temperature, and growth of icelandic cod (*Gadus morhua*): evidence from storage tag data. Can. J. Fish. Aquat. Sci. 60 (11), 1409–1423.
Pearson, R., Dawson, T., Berry, P., Harrison, P., 2002. Species: a spatial evaluation of climate impact on the envelope of species. Ecol. Model. 154 (3), 289–300.
Pearson, R.G., 2012. Species distribution modeling for conservation educators and practitioners, Available at: http://ncep.amnh.org.
Pearson, R.G., Dawson, T.P., 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Glob. Ecol. Biogeogr. 12 (5), 361–371.
Pearson, R.G., Dawson, T.P., Liu, C., 2004. Modelling species distributions in britain: a hierarchical integration of climate and land-cover data. Ecography 27 (3), 285–298.
Phillips, S.J., Dudik, M., 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31, 161–175.
Pihl, L., Ulmestrand, M., 1993. Migration pattern of juvenile cod (*Gadus morhua*) on the swedish west coast. ICES J. Mar. Sci. 50 (1), 63–70.
Platts, P.J., McClean, C.J., Lovett, J.C., Marchant, R., 2008. Predicting tree distributions in an east african biodiversity hotspot: model selection, data bias and envelope uncertainty. Ecol. Model. 218 (1), 121–134.
Quantum GIS, 2011. Quantum GIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org.
R Core Team, 2015. R: A language and environment for statistical computing. http://www.R-project.org.
Ready, J., Kaschner, K., South, A.B., Eastwood, P.D., Rees, T., Rius, J., Agbayani, E., Kullander, S., Froese, R., 2010. Predicting the distributions of marine organisms at the global scale. Ecol. Model. 221 (3), 467–478.
Ricard, D., Branton, R.M., Clark, D.W., Hurley, P., 2010. Extracting groundfish survey indices from the Ocean Biogeographic Information System (OBIS): an example from Fisheries and Oceans Canada. ICES J. Mar. Sci. 67 (4), 638–645.
Righton, D., Andersen, K.H., Neat, F., Thorsteinsson, V., Steingrund, P., Svedäng, H., Michalsen, K., Hinrichsen, H.-H., Bendall, V., Neuenfeldt, S., et al., 2010. Thermal niche of atlantic cod *Gadus morhua*: limits, tolerance and optima. Mar. Ecol. Prog. Ser.
Ruzzante, D.E., Taggart, C.T., Doyle, R.W., Cook, D., 2001. Stability in the historical pattern of genetic structure of newfoundland cod (*Gadus morhua*) despite the catastrophic decline in population size from 1964 to 1994. Conserv. Genet. 2 (3), 257–269.
Schweder, T., 2007. Advanced distance sampling: estimating abundance of biological populations. J. Am. Stat. Assoc. 102 (478), 763–764.
Smith, S.J., Page, F.H., 1996. Associations between atlantic cod (*Gadus morhua*) and hydrographic variables: implications for the management of the 4vsw cod stock. ICES J. Mar. Sci. 53 (3), 597–614.
Stockwell, D., 1999. The garp modelling system: problems and solutions to automated spatial prediction. Int. J. Geogr. Inf. Sci. 13 (2), 143–158.
Svedäng, H., Righton, D., Jonsson, P., 2007. Migratory behaviour of atlantic cod *Gadus morhua*: natal homing is the prime stock-separating mechanism. Mar. Ecol. Prog. Ser. 345, 1–12.
The AquaMaps Consortium, 2014. The AquaMaps, Available at: http://www.aquamaps.org.
Thomas, L., Buckland, S.T., Burnham, K.P., Anderson, D.R., Laake, J.L., Borchers, D.L., Strindberg, S., 2002. Distance sampling. Encyclopedia of environmetrics.
Tsontos, V.M., Kiefer, D.A., 2002. The gulf of maine biogeographical information system project: developing a spatial data management framework in support of OBIS. Oceanol. Acta 25 (5), 199–206.
Vanden Berghe, E., 2015. A script to estimate sampling absence and presence records from OBIS, Downloadable via the high-availability CNR data provisioning system at http://data.d4science.org/uri-resolver/id?filename=absencesspecieslistprod.r&smp-id=551020dbe4b0ffc120bc4837&contenttype=text%2fx-rsrc.

Vanden Berghe, E., Grassle, J., Stocks, K., Halpin, P., Lang da Silveira, F., 2010a. Integrating biological data into ocean observing systems: the future role of OBIS. Proc. OceanObs, 9.
Vanden Berghe, E., Stocks, K.I., Grassle, J.F., 2010b. Data integration: the ocean biogeographic information system. In: Life in the World's Oceans: Diversity, Distribution, and Abundance., pp. 333.
Wheeler, A., Du Heaume, V., 1969. The Fishes of the British Isles and North-West Europe. Macmillan, London.
Yee, T.W., Mitchell, N.D., 1991. Generalized additive models in plant ecology. J. Veg. Sci. 2 (5), 587–602.
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. Ecol. Model. 157 (2), 261–280.
Zeller, D., Froese, R., Pauly, D., 2005. On losing and recovering fisheries and marine science data. Mar. Policy 29 (1), 69–73.