Received: 22 November 2017 | Revised: 13 March 2018 | Accepted: 27 April 2018

DOI: 10.1111/ddi.12782

BIODIVERSITY RESEARCH

WILEY Diversity and Distributions

Finding the right fit: Comparative cetacean distribution models using multiple data sources and statistical approaches

Solene Derville[1,2,3,4] | Leigh G. Torres[3] | Corina Iovan[1] | Claire Garrigue[1,4]

[1]UMR ENTROPIE (IRD, Université de La Réunion, CNRS), Nouméa Cedex, New Caledonia

[2]Collège Doctoral, Sorbonne Université, Paris, France

[3]Department of Fisheries and Wildlife, Marine Mammal Institute, Oregon State University, HMSC, Newport, OR, USA

[4]Operation Cétacés, Nouméa, New Caledonia

Correspondence
Solene Derville, UMR ENTROPIE (IRD, Université de La Réunion, CNRS), 101 promenade Roger Laroque, BPA5, 98848 Nouméa Cedex, Nouvelle-Calédonie.
Email: solene.derville@ird.fr

Funding information
New Caledonian Government; Ministère Français de la transition écologique et solidaire; Vale S.A.; New Caledonian Provinces; World Wildlife Fund; International Fund for Animal Welfare; Fondation d'Entreprises Total

Editor: Jane Elith

Abstract

Aim: Accurate predictions of cetacean distributions are essential to their conservation but are limited by statistical challenges and a paucity of data. This study aimed at comparing the capacity of various statistical algorithms to deal with biases commonly found in nonsystematic cetacean surveys and to evaluate the potential for citizen science data to improve habitat modelling and predictions. An endangered population of humpback whales (*Megaptera novaeangliae*) in their breeding ground was used as a case study.

Location: New Caledonia, Oceania.

Methods: Five statistical algorithms were used to model the habitat preferences of humpback whales from 1,360 sightings collected over 14 years of nonsystematic research surveys. Three different background sampling approaches were tested when developing models from 625 crowdsourced sightings to assess methods accounting for citizen science spatial sampling bias. Model evaluation was conducted through cross-validation and prediction to an independent satellite tracking dataset.

Results: Algorithms differed in complexity of the environmental relationships modelled, ecological interpretability and transferability. While parameter tuning had a great effect on model performances, GLMs generally had low predictive performance, SVMs were particularly hard to interpret, and BRTs had high descriptive power but showed signs of overfitting. MAXENT and especially GAMs provided a valuable complexity trade-off, accurate predictions and were ecologically intelligible. Models showed that humpback whales favoured cool (22–23°C) and shallow waters (0–100 m deep) in coastal as well as offshore areas. Citizen science models converged with research survey models, specifically when accounting for spatial sampling bias.

Main conclusions: Marine megafauna distribution models present specific challenges that may be addressed through integrative evaluation, independent testing and appropriately tuned statistical algorithms. Specifically, controlling overfitting is a priority when predicting cetacean distributions for large-scale conservation perspectives. Citizen science data appear to be a powerful tool to describe cetacean habitat.

KEYWORDS
citizen science, generalized regression, humpback whales, machine learning, species distribution models, support vector machines

1 | INTRODUCTION

Species distribution models (SDMs) have become an indispensable tool for ecologists and conservationists to describe the complex ecological relationships between species and their environment, and to predict distributions over multiple spatial (e.g., Mannocci, Monestiez, Spitz, & Ridoux, 2015) and temporal scales (e.g., Legrand et al., 2016; Morán-Ordóñez, Lahoz-Monfort, Elith, & Wintle, 2017). Correlative SDMs rely on statistical algorithms to fit empirical observations of species occurrence to environmental conditions (Austin, 2007; Elith & Leathwick, 2009; Guisan & Zimmermann, 2000; Guisan et al., 2013). Considering the great potential for SDMs to inform conservation, a growing field of research has emerged to develop applicable models and improve their predictive performance. A multitude of statistical algorithms are now available to build SDMs: profile models (e.g., ecological niche factor analysis (ENFA), Hirzel, Hausser, Chessel, & Perrin, 2002), regression models (e.g., generalized linear models (GLMs), generalized additive models (GAMs), Hastie & Tibshirani, 1990; multivariate adaptive regression splines (MARS), Friedman, 1991), machine learning (e.g., maximum entropy (MAXENT), Phillips, Anderson, & Schapire, 2006; boosted regression trees (BRTs), Friedman, 2001; random forests (RFs), Breiman, 2001; support vector machines (SVMs), Boser, Guyon, & Vapnik, 1992) and Bayesian approaches (occupancy models, MacKenzie, 2006), among others. These methods have been compared empirically (Aguirre-Gutiérrez et al., 2013; Elith et al., 2006; Oppel et al., 2012; Phillips et al., 2009) and with simulated data (Elith & Graham, 2009; García-Callejas & Araújo, 2016; Qiao, Soberón, & Peterson, 2015) in various contexts. Most studies have stressed the existing trade-off between the descriptive and predictive performance of all models, hence emphasizing the fact that model evaluation and transferability are data- and study-specific (Qiao et al., 2015).

The descriptive and predictive power of SDMs has proved particularly useful to understanding the spatial patterns of rare species or species living in ecosystems that are technically challenging to survey (Dunn, Buchanan, Cuthbert, Whittingham, & McGowan, 2015; Engler, Guisan, & Rechsteiner, 2004; Stirling, Boulcott, Scott, & Wright, 2016). Given their wide-ranging behaviour, their rarity and the remote habitats they live in, cetaceans fall in this category (Redfern et al., 2006), with added observational challenges due to the high proportion of time they spend below the surface. Also, as many cetacean species are in need of protection from emerging anthropogenic threats (Avila, Kaschner, & Dormann, 2018), SDMs are greatly valued for their ability to predict probabilities of presence in unsurveyed locations where spatial management is needed (Breen, Brown, Reid, & Rogan, 2017; Gomez et al., 2017; Mannocci, Roberts, Miller, & Halpin, 2017; Redfern et al., 2017). However, cetacean distribution models have unique statistical challenges that warrant specific methodological exploration. Robust predictions have been derived from density surface models (Miller, Burt, Rexstad, & Thomas, 2013), but a large proportion of cetacean research efforts worldwide is not designed to collect data compatible with this approach (e.g., distance measurements, systematic effort). Indeed, nonsystematic cetacean surveys conducted at-sea are often characterized by a heterogeneous spatiotemporal distribution of effort, which can be biased towards easily accessible habitats, areas and times with better weather, or known areas of use (Corkeron et al., 2011). As a result, cetacean habitat datasets tend to display patterns of spatial autocorrelation (Dormann et al., 2007), hierarchical structures (Roberts et al., 2017) and unmeasured confounding effects (e.g., detection distance depending on vessel type, weather, etc.) that can affect SDMs.

MAXENT is among the most popular approach to SDMs (Radosavljevic & Anderson, 2014) and has been applied on numerous occasions for cetacean habitat modelling (e.g., Lindsay et al., 2016; Smith et al., 2012). GLMs and, more recently, GAMs have been applied successfully to many cetacean species (e.g., right whales, Rayment, Dawson, & Webster, 2015; harbour porpoise, Gilles et al., 2016; blue whales, Redfern et al., 2017). BRTs have been less frequently applied to cetacean studies (Derville, Constantine, Baker, Oremus, & Torres, 2016; Torres et al., 2013). At last, SVMs have received less attention in the SDM community (Drake, Randin, & Guisan, 2006) and have never been applied to cetacean habitat modelling. A few comparative analyses of SDMs algorithms have been conducted using cetacean survey data (Macleod, Mandleberg, Schweder, Bannon, & Pierce, 2008; Praca, Gannier, Das, & Laran, 2009; Zanardo, Parra, Passadore, & Möller, 2017), but no recent attempt has explored the ability of increasingly popular modelling methods, such as machine learning, to deal with the biases inherent in the data used for modelling cetacean habitats at large scales.

Many of the technical challenges of data collection in marine ecosystems can be overcome by combining data from multiple sources (Pacifici et al., 2016). To this extent, citizen science may be a promising opportunity to increase the quantity and spatial extent of cetacean observations for habitat modelling efforts (Tiago, Pereira, & Capinha, 2017). Citizen science, as a form of crowdsourcing, can be broadly defined as "the engagement of nonprofessionals in scientific research" (Miller-Rushing, Primack, & Bonney, 2012), and the method may vary from fully trained and equipped volunteers operating in well-defined study areas, to anecdotal reports of observations by members of the general public. In cetacean research, sighting data may be gained from the general public, fishing operators, ferries, oil and gas platforms, cargo ships or whale-watching operators. Citizen science geographical data have been used successfully to study cetacean behaviour and ecology on several occasions (Bruce, Albright, Sheehan, & Blewitt, 2014; Thorne et al., 2012; Tobeña, Prieto, Machete, & Silva, 2016; Torres et al., 2013), but their application to SDMs is fraught with an array of statistical challenges (Bird et al., 2014). Indeed, the probability of recording a species at a given site is always based on both the probability of species occurrence and of an observer recording the data. In citizen science, the sampling effort is rarely recorded, and as a result, it is often hard to determine whether a higher encounter rate at a site is due to high habitat suitability or simply to a higher observer effort (Bird et al., 2014). The correct implementation of methods to account for uneven survey effort, particularly when it was not explicitly quantified, is crucial for cetacean SDMs because highly mobile species are

thought to be especially sensitive to background sampling (Brotons, Thuiller, Araújo, & Hirzel, 2004).

This study investigates the distribution of an emblematic species, the humpback whale *Megaptera novaeangliae*, in New Caledonia, south-western Pacific Ocean. Humpback whales that spend the austral winter in New Caledonian waters are part of the Oceanian breeding population and are classified as endangered by the IUCN (Childerhouse et al., 2009). Furthermore, the recently created Natural Park of the Coral Sea (Decree GNC:2014-1063) requires in-depth knowledge of the spatial distribution and habitats of migratory megafauna to support large-scale management in the region. Fourteen years of whale observations recorded through boat-based nonsystematic research surveys and crowdsourcing are used to model the habitat preferences of humpback whales in the New Caledonian Economic Exclusive Zone (EEZ) through a presence-background SDM approach. The aim of this study was to (a) compare the performance of different SDMs statistical algorithms using a typical cetacean survey dataset and (b) evaluate the potential for crowdsourced cetacean observations to describe and predict habitat preferences using various background sampling techniques that account for sampling bias. An independent humpback whale satellite tracking dataset is tested for robust validation of the modelling approaches.

2 | METHODS

2.1 | Study area

Located in south-western Pacific Ocean (Figure 1) the New Caledonian EEZ spans more than 1.3 million km^2 and is characterized by a complex seabed topography. The area includes a main island, "Grande Terre," as well as remote reef complexes such as the Chesterfield-Bellona plateaus (60 m deep on average), seamounts such as Antigonia seamount (60 m deep), and shallow banks such as the Fairway-Landsdowne banks (200–0 m deep). The mainland is surrounded by a barrier reef that delineates large lagoons. Shallow waters are therefore found both nearshore and offshore (defined here as waters at least 10 km away from any reef or land). New Caledonia is visited every austral winter by a humpback whale breeding substock that is part of the endangered Oceanian population (Childerhouse et al., 2009).

2.2 | Data collection

Data processing and statistical analysis were performed with R (version 3.3.2, R Core Team, 2016), QGIS (version 2.18.3, QGIS Development Team, 2016) and ARCMAP (version 10.3, ESRI, 2016).

2.2.1 | Research surveys dataset

At-sea humpback whale surveys were conducted from June to October, over 14 years between 2003 and 2016. The survey effort was nonsystematic as it did not follow transect lines (see "haphazard" surveys in Corkeron et al., 2011) and was conducted in closing mode (cetaceans were approached after detection). The location of survey effort was determined to maximize chances of whale encounter while accounting for common cetacean survey limitations: weather conditions, harbour proximity and vessel capacity (e.g., Derville et al., 2016). As a result, effort and observations were spatially biased towards coastal and reef areas, a data

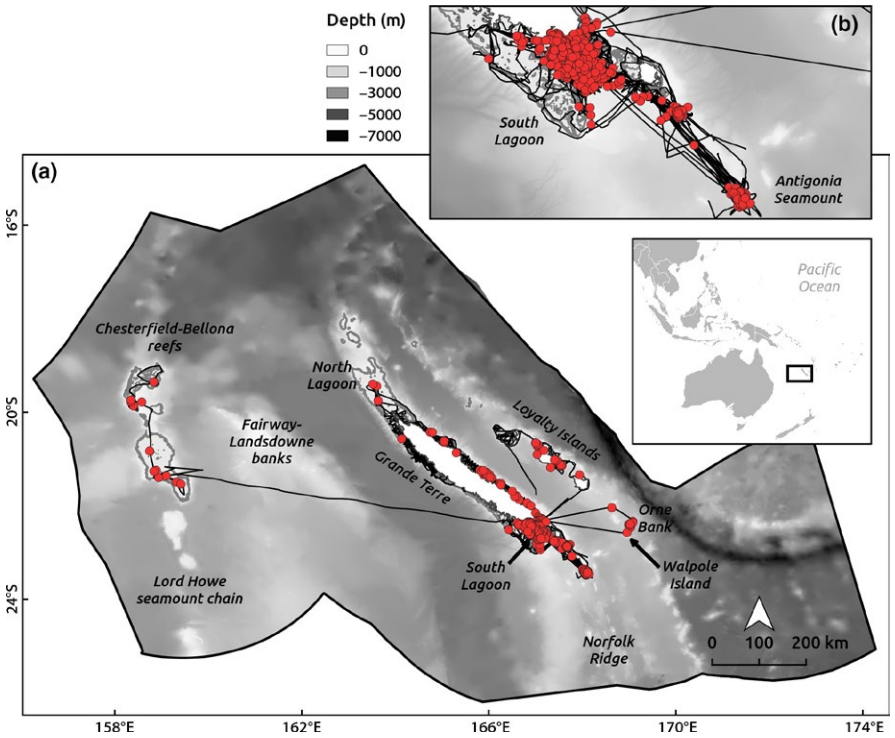

FIGURE 1 Research surveys and humpback whale observations (2003–2016) in New Caledonian waters (a). Shallow reefs are shown in grey over a depth raster. GPS tracklines of the boat are shown in black, with most of the survey effort concentrated in the South lagoon (b). Humpback whale observations are represented with red points (n = 1,360) [Colour figure can be viewed at wileyonlinelibrary.com]

clustering pattern commonly found in cetacean sea survey datasets (Kaschner, Quick, Jewell, Williams, & Harris, 2012). Most of the surveys (65%) were conducted in the South Lagoon (Figure 1). Small semirigid hulled inflatable boats were typically used (76% survey days), with three to five trained observers aboard (see Garrigue, Greaves, & Chambellant, 2001). To a lesser extent (24% survey days), larger vessels such as catamarans and oceanographic vessels were used to survey other areas of the New Caledonian EEZ (Figure 1). Cetaceans were searched for by naked eye in Beaufort sea states ≤3. All GPS boat tracklines were standardized to display one position per minute (initial resolution ranging from 1 position/30 s to 1/min). Presence locations were recorded as the position of the vessel for each whale group encounter. Encounters are considered independent events, as repeated observations of the same individual whale within a survey day rarely occurred (Derville, Torres, & Garrigue, 2018).

2.2.2 | Citizen science dataset

Crowdsourced sightings of marine mammals included in this analysis were recorded from June to October 2003–2016 through a marine mammal observation network coordinated by NGO Opération Cétacés since 1991 www.operationcetaces.nc. Sightings were conserved when: (a) the volunteer provided a picture allowing an accurate identification of the species, (b) the volunteer had advanced cetacean species identification skills or (c) enough description was provided to perform species identification with little doubt (e.g., shape of the fluke/dorsal, specific surface activities). Precise GPS positions were recorded in 50% of cases. Other sightings were positioned within 2 km confidence in 82% cases (up to 5 km max) using the description of the locations (usually referencing small reefs/bays) projected in a GIS website (https://explorateur-carto.georep.nc/).

2.2.3 | ARGOS tracking dataset

Adult humpback whales were tagged in coastal and offshore waters around New Caledonia from 2007 to 2016, in August and September (n = 43, for more details see Garrigue, Clapham, Geyer, Kennedy, & Zerbini, 2015) with implantable transmitters (SPOT5, SPLASH-10 ©Wildlife Computers). Whales of both sexes were equally sampled (21 females, 21 males and one unknown), including females with a calf (n = 14). ARGOS locations of lowest quality (classes "B" and "Z"; Nicholls, Robertson, & Murray, 2007), overlapping with land or implying unrealistic speeds (>12 km/h), were removed.

2.2.4 | Environmental data

Dynamic environmental conditions averaged at a monthly temporal scale were included in this analysis based on hypothesized humpback whale preferences. A monthly scale was considered a good temporal trade-off to capture coarse scale intra and interannual oceanographic processes (e.g., El Niño Southern Oscillation phenomenon) that could affect whales in their tropical breeding latitudes (Fernandez, Yesson, Gannier, Miller, & Azevedo, 2017; Mannocci, Boustany, et al), while allowing for almost gap-free remotely sensed maps. Sea surface temperature (SST) and diffuse attenuation at 490 nm (K490) were extracted from remotely sensed data sources at weekly resolutions and averaged per month from June to October of each year (Table 1). SST has frequently been correlated with many top predator distributions (Scales et al., 2014) and specifically breeding humpback whales (Bortolotto, Danilewicz, Hammond, Thomas,

TABLE 1 Predictor variables implemented in the habitat preference models for humpback whales in New Caledonian waters

Predictor	Description	Unit	Resolution	Source
SST	Sea surface temperature	°C	0.04° monthly	NOAA[a] SWFSC ERD (MODIS)[b] https://oceancolor.gsfc.nasa.gov/
K490	Diffuse attenuation at 490 nm	–	0.04° monthly	NASA[c]/GSFC (MODIS)[d] https://oceancolor.gsfc.nasa.gov/
DEPTH	Depth	m	500 m	DTSI[e] + NOAA ETOPO Composit www.ngdc.noaa.gov/
DISSURF	Distance to closest land/reef	km	500 m	Millennium Coral Reef Mapping www.imars.marine.usf.edu/MC/
S.AVG	Mean slope	rad	5 km mw	
S.COV	Coefficient of variation of the slope	–	5 km mw	
A.AVG	Mean aspect (slope orientation)	rad	5 km mw	
CPRO	Profile curvature[f]	–	5 km mw	

Notes. mw, moving window.
[a]National Oceanographic and Atmospheric Agency.
[b]Moderate Resolution Imaging Spectroradiometer, dataset reference: erdMH1sstd8day.
[c]National Aeronautics and Space Administration.
[d]Dataset reference: erdMH1kd4908day.
[e]Direction des Technologies et des Services de l'Information.
[f]For more details on curvature, see: https://desktop.arcgis.com/en/arcmap/10.3/manage-data/raster-and-images/curvature-function.htm

& Zerbini, 2017; Rasmussen et al., 2007; Smith et al., 2012). K490, which is a measure of turbidity, has also been linked with cetacean distribution (Mendez, Rosenbaum, Subramaniam, Yackulic, & Bordino, 2010). K490 tends to be systematically higher inside the tropical lagoon environment (see Supporting Information Appendix S1) and was therefore included as a proxy of suitable humpback whale habitat in shallow lagoons (Lindsay et al., 2016).

Depth (DEPTH) was primarily extracted from a 500-m resolution bathymetric chart, and small gaps were filled with the ETOPO 1 maps (Table 1). Several topographic variables were derived from bathymetry to best capture the seabed topographic complexity (Bouchet, Meeuwig, Salgado Kent, Letessier, & Jenner, 2015) of the unique New Caledonian region. Mean slope (S.AVG), coefficient of variation of the slope (S.COV) and mean aspect (A.AVG, orientation of the slope) were calculated using a 5 × 5 km moving window. Euclidean distance to the closest land or shallow reef (DISSURF) was calculated from coastline and reef shapefiles (Andréfouët, Chagnaud, Chauvin, & Kranenburg, 2008). At last, profile curvature (C.PRO) was calculated using the ARCMAP "3D Analyst Tool" and averaged over a 5 × 5 km moving window to estimate the convexity of the slope and reveal terracing of seabed structures such as seamounts (Table 1).

To ensure consistency across statistical algorithms, all environmental variables were scaled and centred, by subtracting the mean and dividing by the standard deviation calculated over the full presence-background dataset. At last, Pearson's coefficients were calculated between environmental variables in the presence-background dataset to prevent collinearity (control that $r < 0.5$ for all variables).

2.3 | Modelling habitat preferences

2.3.1 | Using research survey data

Humpback whale occurrence data collected during research surveys were modelled relative to environmental conditions with five algorithms: GLM, GAM, BRT, MAXENT and SVM. While nonsystematic cetacean surveys are generally not designed to record data as presence–absence, they often include some sort of sampling effort estimation, through the recording of times on effort and boat GPS tracklines. Here, the areas surrounding boat tracklines were used to characterize available environmental conditions between sighting locations (presence) and area surveyed (background, e.g., Derville et al., 2015; Torres, Read, & Halpen, 2008). Tracklines were segmented into on- and off-effort sections. A set of points, denoted background points (a.k.a "pseudo-absences"), was sampled within the on-effort survey track strip-width, spanning 4 km to each side of the tracklines to reflect the average detection distance of the semi-inflatable boat used in most surveys (pers. comm. Garrigue), although detection distance might have been larger with the bigger research vessels. Daily samples of background points were generated with a minimum distance of 1 km from each other, but independent of presence locations. The number of background points was proportional to the time on effort per survey day (on average 35 points per 5-hr survey). Combined background and presence points constituted a binomial dataset of 18,046 data points.

Cross-validation is a common model evaluation procedure and a powerful tool to account for hierarchical structures within the dataset, such as spatial autocorrelation (Roberts et al., 2017). Here, Monte Carlo cross-validation accounted for dependencies in the observation data, namely the daily autocorrelation resulting from daily clusters in the extent and intensity of the survey effort. The dataset was divided into 638 blocks containing presence and background points for each day of survey. Fifty training datasets containing 90% of randomly selected days of survey were sampled without replacement. As a result, each training dataset of the cross-validation contained many blocks (each block is a survey day) and was paired with an evaluation dataset containing the remaining blocks. Presence and background points were weighted to control for prevalence, so that the sum of weights on presences was equal to the sum of weights on background points in each training dataset (Elith, Kearney, & Phillips, 2010).

Boosted regression trees, SVMs and MAXENT models were subject to a preliminary tuning stage ensuring optimal performance within the scope of our training datasets. In the GLMs, each predictor was included as a cubic orthogonal polynomial (see Supporting Information Appendix S2). In the GAMs, restricted maximum likelihood was used to optimize parameter estimates for the thin-plate regression splines (see Supporting Information Appendix S2). All models were first tested using a set of nine predictors, including eight environmental variables plus Julian day, then run using a smaller set of predictors after removing the ones that contributed the least (See Supporting Information Appendix S3). Julian day was added to the set of predictors to account for the seasonal phenology of humpback whales in breeding areas that results in a peak of prevalence in August. The contribution of each predictor was directly provided in the R summaries for BRTs and MAXENT models, but assessed using the "caret" R package (version 6.0) for GLMs/GAMs. For SVMs, the recursive feature elimination algorithm (Guyon, Weston, Barnhill, & Vapnik, 2002) was applied for linear kernels only (as this method is not available for radial kernel SVMs), and the resulting ranking criteria were rescaled to sum to 100. For GLMs, the contribution of the three orthogonal polynomial terms was summed per predictor. All contributions were averaged over the 50 cross-validation runs.

Partial dependence plots were produced for each predictor variable and averaged over the 50 cross-validation runs of each statistical algorithm. These plots allow the graphical visualization of the marginal effect of a given variable on the response while all other predictors are held constant at their mean sampled value (Friedman, 2001). They provide a useful ecological interpretation of SDMs, though should be regarded with caution when strong interactions exist between the predictors (Goldstein, Kapelner, Bleich, & Pitkin, 2015).

2.3.2 | Using citizen data

Three different sampling approaches were tested to generate background points, hereon referred to as "UNIFORM", "TARGET" and

"POP." The number of background points was set separately for each approach to generate the same background density as in the research survey models (estimated to a minimum of 0.02 point/km^2). In the UNIFORM sampling approach, 36,300 background points (equivalent to 605 per month) were randomly sampled over the entire New Caledonian EEZ (covering 1.6 M km^2, Figure 2b). The TARGET sampling is based on a popular method developed by Phillips et al. (2009) in which the spatial bias in the sightings data is transferred to the background data by approximating areas where the probability of detection is nonzero. In practice, the areas of background sampling may be limited to those where sightings of species within the same taxonomic group have been reported by the public. Here, 2,340 background points (equivalent to 39 per month) were sampled in 25 km buffers surrounding all marine mammal observations in the citizen science New Caledonian dataset (n = 818 sightings across 15 marine mammal species, including humpback whales, background area covering 0.1 M km^2, Figure 2c). At last, the POP sampling approach was designed to correct the spatial bias in crowdsourced sightings by including a proxy of human densities in the background data (Figure 2d). This approach relies on the assumption that sampling is biased towards waters that are more accessible/closer to human settlements or that are more attractive to people. In New Caledonia, most of the population concentrates in the mainland "Grande Terre," specifically in the capital Noumea (Figure 2a). Also, lagoons and waters surrounding the reef's outer edge are popular sites for recreational activities. The POP background sampling was designed to sample 36,300 background points over the EEZ proportionally to local human density (see Supporting Information Appendix S4).

The relationship between the observations of humpback whale groups by citizens and environmental conditions was modelled using GAMs with the same settings as the research survey GAMs. Monte Carlo cross-validation was applied over 50 randomly sampled training and evaluation datasets representing respectively 90% and 10% of the total datasets stratified by months. Weights for presence and background points were applied similarly to the research survey models. GAMs were applied to seven predictors: DEPTH, DISSURF, S.AVG, S.COV, K490, SST and month to account for humpback whale migratory phenology.

2.4 | Validation and prediction

The descriptive power of each model was assessed by calculating the area under the ROC curve (AUC) of the training datasets ("int. AUC"). AUC measures the capacity of the models to classify between presence and background points and ranges between 0 and 1 (Swets, 1988). This metric allows a "threshold-independent" evaluation of model performance, a useful characteristic for model comparison. Predictive performance was assessed by calculating AUC over the evaluation datasets ("ext.AUC"), which is the withheld data portion in each cross-validation iteration. The absolute value of the difference between ext.AUC and int.AUC was also calculated to assess the degree of overfitting in the model ("diff.AUC", Warren & Seifert, 2010). A threshold to convert continuous predicted probabilities into a binomial output was estimated for each model run, using the threshold value that maximized specificity (true negative rate) and sensitivity (true positive rate) over the evaluation dataset predictions (Liu, Newell, & White, 2016). Using this threshold, two metrics of predictive performance were derived: the sensitivity of models when predicting ARGOS tracking locations ("sensitivity. ARGOS", in % correctly classified as presences), and the true statistic skill when predicting the evaluation datasets ("TSS"; Allouche, Tsoar, & Kadmon, 2006). Following the tuning of BRTs, SVMs and MAXENT models, two different settings were selected for each approach: the model with highest ext.AUC was considered the best predictive model (annotated ".pred"), while the model with the lowest diff.AUC was considered the most stable model (annotated ".stable"). At last, the predictive performance of the citizen science models was tested relative to the 50 research survey evaluation datasets, hence allowing the estimation of AUC values ("comp.AUC") comparable to the research survey ext.AUC.

Humpback whale habitat suitability was predicted on a grid with 500 × 500 m cells covering the EEZ. For this purpose, SST and K490 were averaged beforehand over June to October, from 2003 to 2016. Julian day and month were fixed to the date of the peak of humpback whale presence for the research survey and citizen science models, respectively, on 28th August. Predicted layers for each model were averaged over the 50 cross-validation runs

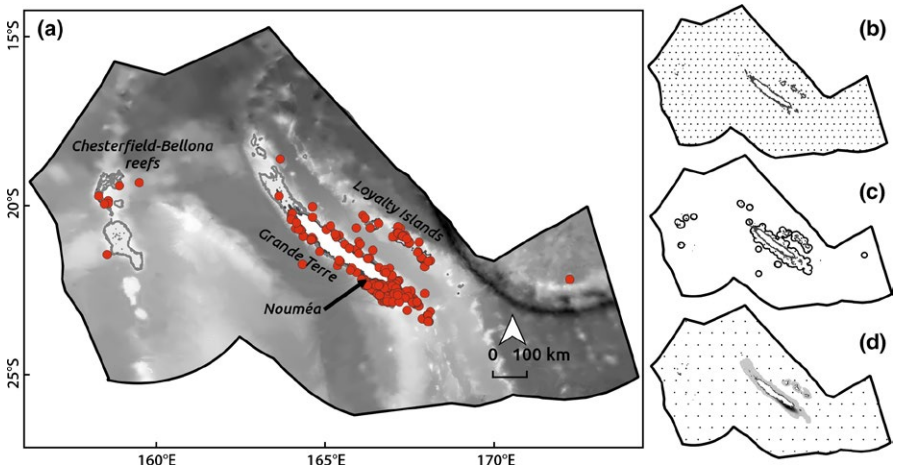

FIGURE 2 Citizen science observations of humpback whale groups (2003–2016) in New Caledonian waters (a). Observations are represented with red points (n = 625). Schematics of the three background sampling methods are provided: the UNIFORM sampling (b), the TARGET sampling restricted to areas surrounding sightings (c) and the POP sampling weighted in proportion to human densities (d). In the last approach, darker shades of grey represent a higher probability of sampling [Colour figure can be viewed at wileyonlinelibrary.com]

(Roberts et al., 2017), and the standard deviations of predictions were mapped to report uncertainty (see Supporting Information Appendix S3). The similarity between average predicted maps was assessed using Pearson's coefficients. Environmental extrapolation was not limited in the predictions per se, but the areas where environmental conditions strayed outside their training ranges were highlighted in the final maps of habitat suitability to be interpreted with caution (e.g., Mannocci, Roberts et al., 2017).

3 | RESULTS

Dedicated research surveys covered 49,843 km across 14 years and 638 days of effort (see Supporting Information Appendix S3, Table 1). Survey effort covered 21% of the EEZ waters and encountered a total of 1,360 humpback whale groups (annual mean = 97 ± *SD* 40 groups). A total of 625 humpback whale group encounters were recorded opportunistically by citizen scientists (annual mean = 45 ± *SD* 28). Sightings were recorded predominantly by park rangers (29%) and whale-watching operators (22%). After filtering the 43 raw ARGOS tracks, 1,539 locations of 4,180 were conserved.

3.1 | Modelling habitat preferences from research survey data

All models were first applied to the set of nine predictors; then, the predictors that contributed the least, CPRO and A.AVG, were removed for further analysis (Supporting Information Appendix S3).

The comparison of parameter tunings for BRTs, SVMs and MAXENT models showed a trade-off between diff.AUC and ext.AUC/TSS (Table 2). For a given algorithm, the impact of tuning on all evaluation metrics was large; for instance, MAXENT models showed a 9% increase in ext.AUC when applied with hinge features in comparison with linear features only. Models, selected for their "predictive performance" (high ext.AUC and TSS), were less "stable" from training to evaluation (larger diff.AUC).

The same trade-off was present in the broad comparison of the five statistical algorithms. Diff.AUC was highest for BRTs, and the SVM.pred model, reflecting increased overfitting of the relationships. Statistical algorithms can be ranked in increasing diff.AUC: SVM.stable—MAXENT.stable, GAM, GLM—MAXENT.pred, SVM.pred, BRT.stable, BRT.pred; and in decreasing ext.AUC: BRT.pred, SVM.pred, BRT.stable, MAXENT.pred, GAM, GLM, MAXENT.stable, SVM.stable. TSS was correlated to ext.AUC (n = 8, Pearson's r = 0.98) and was surprisingly high for GAMs considering its medium ext.AUC. Sensitivity calculated over the ARGOS data tended to be lower in more complex models that had high ext.AUC (BRTs, SVM.pred, GAMs).

The five statistical algorithms mostly agreed on the relative contribution of the main variables. DEPTH, DISSURF, and SST were the major contributors, together accounting for 54%–96% of the contributions (Table 3). Yet, both algorithm type and tuning impacted the predictor's contributions. Contrary to GLMs and GAMs where DISSURF was preponderant, BRTs found that DEPTH was the most important predictor, with very little effect of DISSURF. In an interesting manner, K490 had a relatively high contribution in BRTs and

TABLE 2 Parameters and validation metrics of habitat preference models for humpback whales in New Caledonian waters. The mean and (±) standard deviation of each metric is calculated over 50 runs of the cross-validation. For SVMs, BRTs and MAXENT models, metrics for the parameterization that led to the best diff.AUC ("stable model") and ext.AUC ("predictive model") are reported

	Tuning	int.AUC	ext.AUC	diff.AUC	TSS	sensitivity.argos %
Research survey model						
GLM		0.724 ± 0.003	0.714 ± 0.032	0.011 ± 0.035	0.349 ± 0.053	61.8 ± 6.3
GAM		0.736 ± 0.003	0.727 ± 0.031	0.009 ± 0.034	0.373 ± 0.05	42.7 ± 4.9
$MAXENT_{.stable}$	linear, beta 1[a]	0.675 ± 0.005	0.675 ± 0.041	0 ± 0.046	0.274 ± 0.063	53.3 ± 9.8
$MAXENT_{.pred}$	hinge, beta 1[a]	0.747 ± 0.004	0.736 ± 0.031	0.011 ± 0.034	0.364 ± 0.055	46.1 ± 6.2
$SVM_{.stable}$	linear, cost 0.01[b]	0.669 ± 0.005	0.669 ± 0.041	0 ± 0.046	0.27 ± 0.062	70.9 ± 14.5
$SVM_{.pred}$	radial, cost 10[b]	0.772 ± 0.003	0.744 ± 0.029	0.028 ± 0.032	0.39 ± 0.047	42.8 ± 7.0
$BRT_{.stable}$	lr 0.005, tc 1[c]	0.767 ± 0.004	0.738 ± 0.033	0.029 ± 0.036	0.364 ± 0.056	43.9 ± 7.8
$BRT_{.pred}$	lr 0.005, tc 3[c]	0.843 ± 0.004	0.775 ± 0.027	0.069 ± 0.029	0.425 ± 0.045	40.8 ± 5.9
Citizen science models						
UNIFORM		0.990 ± 0.001	0.990 ± 0.005	0.001 ± 0.006	0.936 ± 0.021	47.0 ± 6.8
POP		0.947 ± 0.003	0.937 ± 0.017	0.010 ± 0.02	0.754 ± 0.041	46.0 ± 10.1
TARGET		0.927 ± 0.004	0.919 ± 0.027	0.009 ± 0.031	0.733 ± 0.075	43.9 ± 12.3

Notes. [a]MAXENT models were applied with a linear or hinge feature and beta parameter equal to 1.
[b]SVMs were applied with linear or radial kernel type and cost of constraint violation equal to 0.01 or 10.
[c]BRTs were applied with a learning rate of 0.005 and a tree complexity of 1 or 3.

TABLE 3 Mean contribution of environmental variables to habitat preference models for humpback whales in New Caledonian waters. Values are ranked and scaled to 100 separately for each algorithm (greatest influence in bold). Coefficients of variation (%) of the mean contribution calculated over 50 cross-validation runs are indicated by ±. For SVMs, BRTs and MAXENT models, contributions for the parameterization that led to the best diff.AUC ("stable model") and ext.AUC ("predictive model") are reported

[a]	S.AVG	S.COV	JULIAN/MONTH	K490	SST	DISSURF	DEPTH
Research survey model							
GLM	5.5 ± 17.7%	11.0 ± 8.8%	9.9 ± 11.1%	**19.4 ± 9.7%**	**19.0 ± 8.8%**	**21.6 ± 13.5%**	13.6 ± 13.0%
GAM	2.2 ± 27.3%	2.3 ± 21.7%	9.8 ± 11.2%	10.7 ± 23.4%	**22.9 ± 8.7%**	**28.4 ± 10.2%**	**23.7 ± 9.7%**
$MAXENT_{.stable}$	7.7 ± 34.7%	0.4 ± 54.6%	0.2 ± 127.2%	0.9 ± 28.6%	**40.8 ± 7.4%**	**28.9 ± 10.0%**	**21.2 ± 13.6%**
$MAXENT_{.pred}$	1.2 ± 30.0%	1.4 ± 60.9%	4.1 ± 17.4%	2.4 ± 20.5%	**23.8 ± 9.6%**	**20.4 ± 6.6%**	**46.6 ± 4.4%**
$SVM_{.stable}$	2.5 ± 32.4%	0.4 ± 38.4	0.3 ± 89.5%	0.6 ± 36.8%	**75.1 ± 1.7%**	**12.9 ± 6.4%**	**8.2 ± 22.3%**
$BRT_{.stable}$	6.1 ± 8.6%	5.6 ± 13.6%	2.4 ± 13.4%	**20.9 ± 5.5%**	**27.0 ± 6.8%**	2.6 ± 12.8%	**35.5 ± 4.3%**
$BRT_{.pred}$	6.9 ± 6.5%	**17.4 ± 5.1%**	4.6 ± 7.0%	16.5 ± 5.5%	**23.9 ± 6.5%**	5.2 ± 6.3%	**25.6 ± 4.5%**
Citizen science models							
UNIFORM	0.6 ± 66.7%	1.9 ± 21.1%	1.3 ± 46.2%	**13.8 ± 12.3%**	9.4 ± 17.0%	**37.2 ± 7.8%**	**35.7 ± 12.6%**
POP	1.7 ± 35.3%	1 ± 30.0%	**11.1 ± 17.1%**	**55 ± 5.5%**	6.2 ± 14.5%	7.7 ± 15.6%	**17.4 ± 16.1%**
TARGET	1.6 ± 37.5%	2.1 ± 33.3%	4.4 ± 38.6%	**39.9 ± 7.8%**	**20.7 ± 16.9%**	1.7 ± 88.2%	**29.5 ± 20.7%**

Notes. [a]Average slope (S.AVG), Julian date (JULIAN) for research survey models or month of year (MONTH) for citizen science models, coefficient of variation of the slope (S.COV), diffuse attenuation as turbidity index (K490), sea surface temperature (SST), distance to closest reef or land (DISSURF) and depth (DEPTH).

GLMs. Tuning affected contributions: MAXENT.stable favoured SST, while MAXENT.pred favoured DEPTH.

Ecological relationships between humpback whale occurrence and environmental conditions (Figure 3) showed different trends across the five statistical algorithms and varying complexity. In relation to overfitting trends revealed by high diff.AUC and ext.AUC in Table 2, BRTs showed noisy response curves. On the contrary, GLMs, SVMs and MAXENT models captured the general trends in the relationships but missed some specific features. For instance, habitat suitability globally increased with increasing DISSURF in BRTs, GLMs and MAXENT models, whereas SVMs predicted high suitability only for small DISSURF values (around 20 km). GAMs predicted a bimodal relationship to DISSURF, with a high suitability around 35 km, then between 130 and 200 km, and a decrease for larger distances.

Overall, humpback whales favoured shallow waters about 0–100 m deep, and relatively cold-water temperatures, between 22°C and 23°C. Models demonstrated that whales had a preference for relatively flat seabeds (low S.AVG), of medium to relatively high topographic complexity (S.COV 1 – 2% and above), which could represent the top of banks, seamounts or reef lagoons. At last, the probability of occurrence increased with lower values of K490, but most models demonstrated a peak between 0.1 and 0.2, denoting a preference for medium turbidity.

The algorithms differed in their predictions over certain zones (Figure 4 and Supporting Information Appendix S3), such as the Loyalty Islands, which were suitable in GAMs and SVMs but not in the other approaches. GAMs, BRTs and MAXENT models predicted smoother gradients over the study area, while GLMs predicted low suitability in most lagoons and SVMs had strong cut-offs in the predicted values. The algorithms also differed in their predictions into unsampled environmental space (dashed areas, Figure 4): BRTs and MAXENT models predicted a high suitability for the whole southern part of the study area, while GLMs predicted high suitability everywhere in the extrapolation zone. The extrapolations from GAMs appeared to be mostly driven by the bathymetric pattern. In general, spatial overlap between ARGOS tracking locations and areas of high habitat suitability was high for all models (e.g., Figure 4e), especially South of the mainland. Excluding areas of extrapolation, the five models agreed on humpback whale preference for shallow waters, which resulted in high habitat suitability predictions for reef complexes (Chesterfield-Bellona, North Lagoon, South Lagoon), banks (Fairway-Landsdowne, Orne bank), coastal waters (Loyalty Islands) and shallow seamounts of the Lord Howe seamount chain and Norfolk Ridge.

3.2 | Modelling habitat preferences from citizen science data

The three citizen science models had high AUC (> 0.90, Table 2). The UNIFORM model had the best predictive performance (highest ext. AUC and sensitivity.argos), followed by the POP and TARGET models. Most important, the TARGET model and to a lesser extent the POP model better predicted research survey occurrences (comp. AUC = 0.573 ± 0.006 and 0.541 ± 0.003, respectively) than the UNIFORM models (comp.AUC = 0.538 ± 0.004).

The three citizen science models differed in the relative contribution of predictors (Table 3). The TARGET model was the most similar to the research survey models, with SST and DEPTH having a great influence. DISSURF was a major contributor to the UNIFORM model only. At last, in all three models, K490 was among the most influential predictors.

Predicted maps of habitat suitability (Figure 5) were very similar between the UNIFORM and POP models (Figure 5a,c,

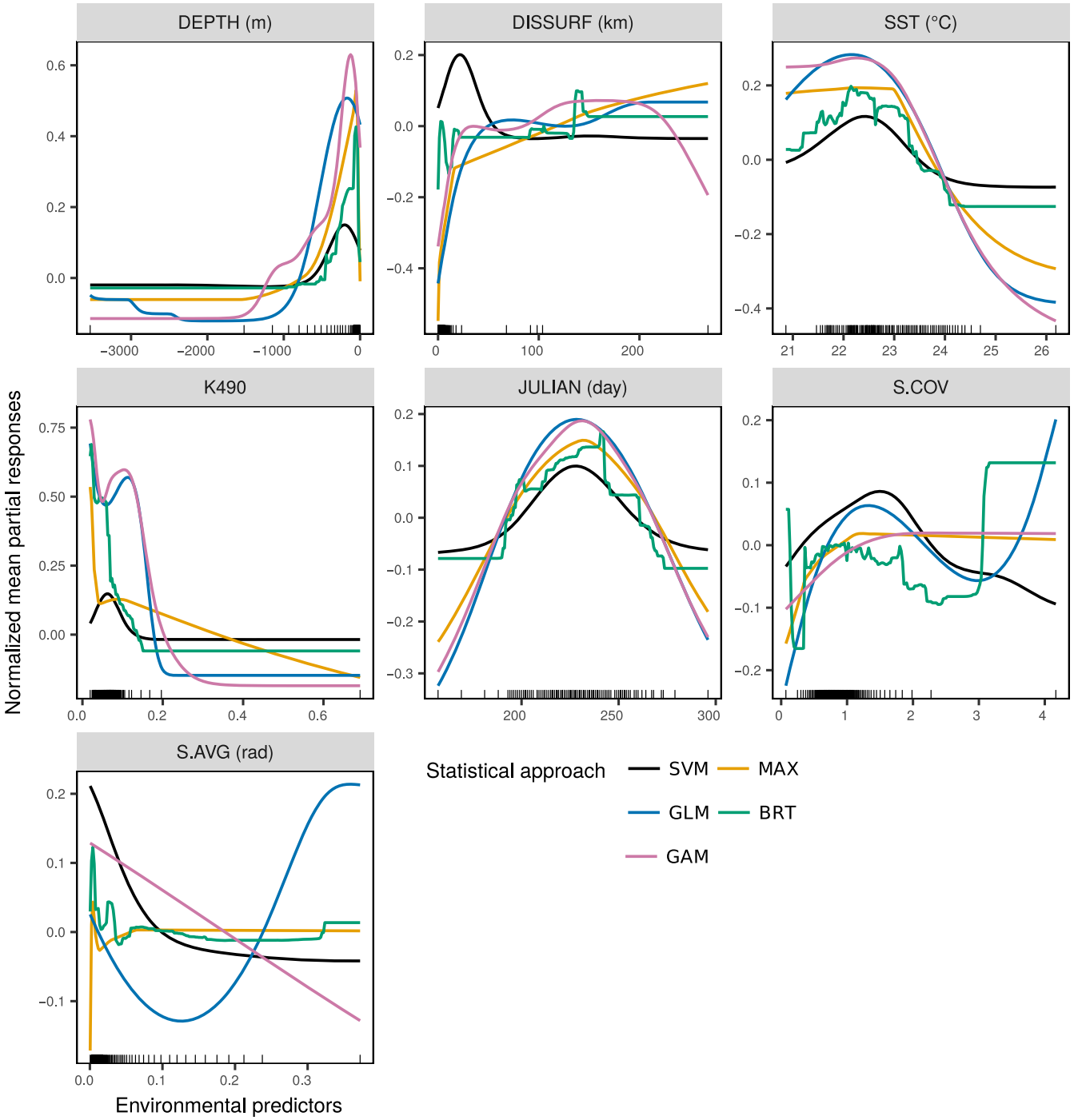

FIGURE 3 Mean partial dependence plots obtained by five statistical algorithms to model humpback whale occurrence from research survey data with respect to environmental variables: DEPTH: depth; DISSURF: distance to closest reef or land; S.AVG: mean slope; S.COV: coefficient of variation of the slope; SST: sea surface temperature and K490 = diffuse attenuation at 490 nm (turbidity). Solid lines represent the mean marginal effect of each variable relative to the probability of presence, over 50 cross-validation runs. Probabilities on the y-axis originally ranging from 0 to 1 were normalized per model to be centred on zero. Rug plots show the distribution of values in the full presence-background research survey dataset, in percentiles, and provide a measure of confidence on the fitted responses. For SVMs, BRTs and MAXENT models, only the plots obtained with the "predictive" tuning (highest ext.AUC) are reported [Colour figure can be viewed at wileyonlinelibrary.com]

Pearson's r = 0.98). Despite being affected by environmental extrapolation over part of the study area (Figure 5b), the TARGET models prediction maps fitted more closely with the research survey maps (Figures 4e, 5b, Pearson's coefficient: r = 0.74), with offshore shallow waters such as the Fairway-Landsdowne bank showing particularly high suitability. The three citizen science models predicted all waters located in reef or coastal habitats to be suitable.

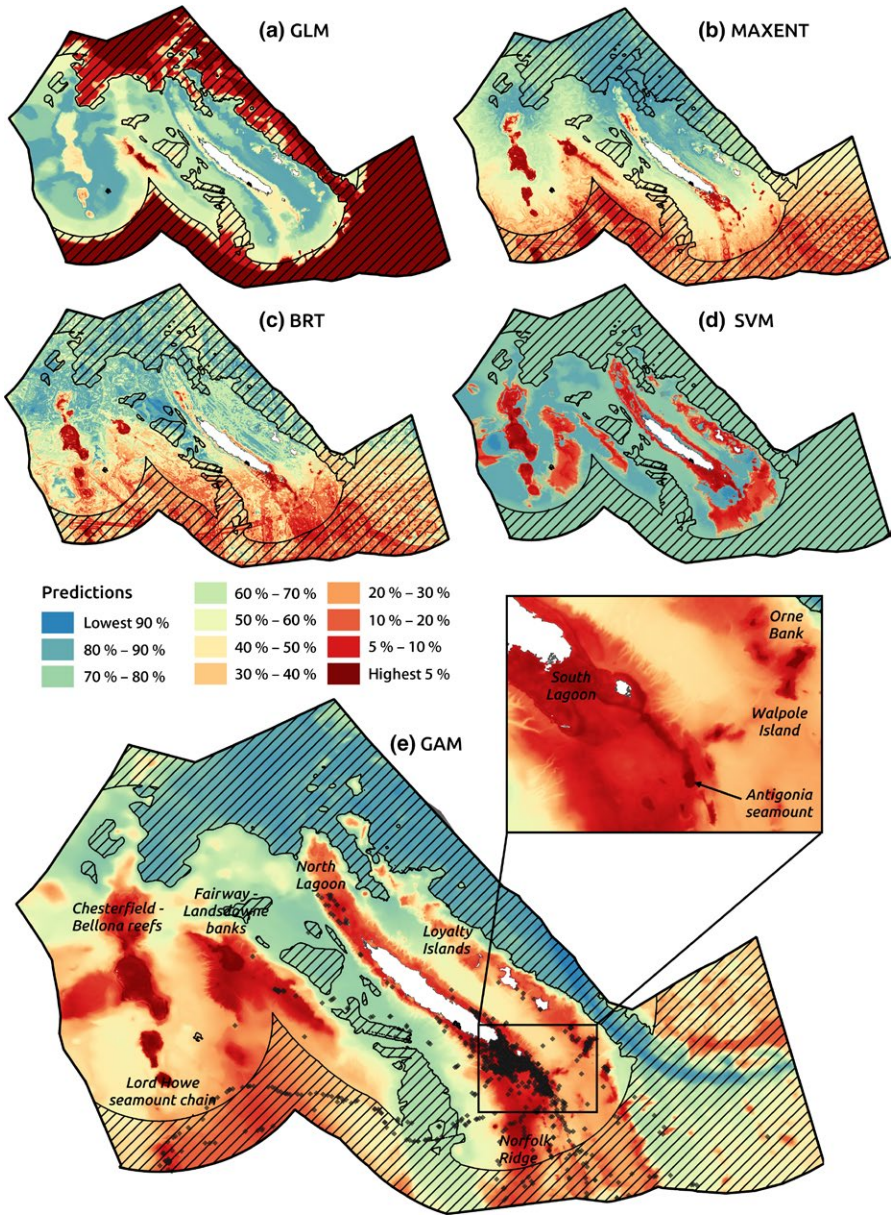

FIGURE 4 Maps of mean predicted humpback whale habitat suitability from research survey models. Habitat suitability was averaged over 50 cross-validation runs for each statistical algorithm, and a coloured log-scale was applied to values ranging from 0 to 1. Colours represent fixed percentages of probability distributions of the suitability predicted values (e.g., the highest 10% corresponds to the decile with highest values over each map). Areas of extrapolation where at least one environmental variable expanded outside the range observed in the training dataset are dashed. Filtered positions from satellite tags deployed in the region are shown with black squares in panel (e). For SVMs, BRTs and MAXENT models, only the plots obtained with the "predictive" tuning (highest ext.AUC) are reported [Colour figure can be viewed at wileyonlinelibrary.com]

4 | DISCUSSION

The multisource New Caledonian humpback whale dataset allowed an in-depth methodological investigation of practices (background sampling, statistical algorithms, model tuning, evaluation and predictions) to generate informative SDMs using nonsystematic and citizen science data for cetacean species. Derived results are broadly applicable to other marine megafauna modelling efforts as observations collected during nonsystematic surveys and through citizen science are representative of worldwide research efforts to study marine mammals. Statistical algorithm comparisons performed on the research survey dataset revealed differences in the complexity of the environmental relationships modelled, the ecological interpretability of outputs and model transferability across large geographical scales. Although citizen science models did not perform as well as the research survey models, they predicted similar humpback whale suitable habitats and benefited from specifically tuned background sampling approaches that account for spatial bias of effort.

In nonsystematic closing mode surveys, covariates affecting detection may not be precisely recorded (e.g., sea state, vessel type/height, number of observers) and may vary within and between surveys days. While presence-background approaches should not be considered a solution to imperfect detection (Monk, 2014), they can be applied safely as long as detection probability is not directly correlated to the habitat variables of interest. Such correlation may exist if a cetacean species spends more time at the surface when resting/feeding in specific habitats for instance. A general balance between model complexity and generality was observed, in concordance with the conceptual framework detailed by Guisan and Zimmermann (2000). Models that more closely fit the relationships in the training data were less efficient at model extrapolation

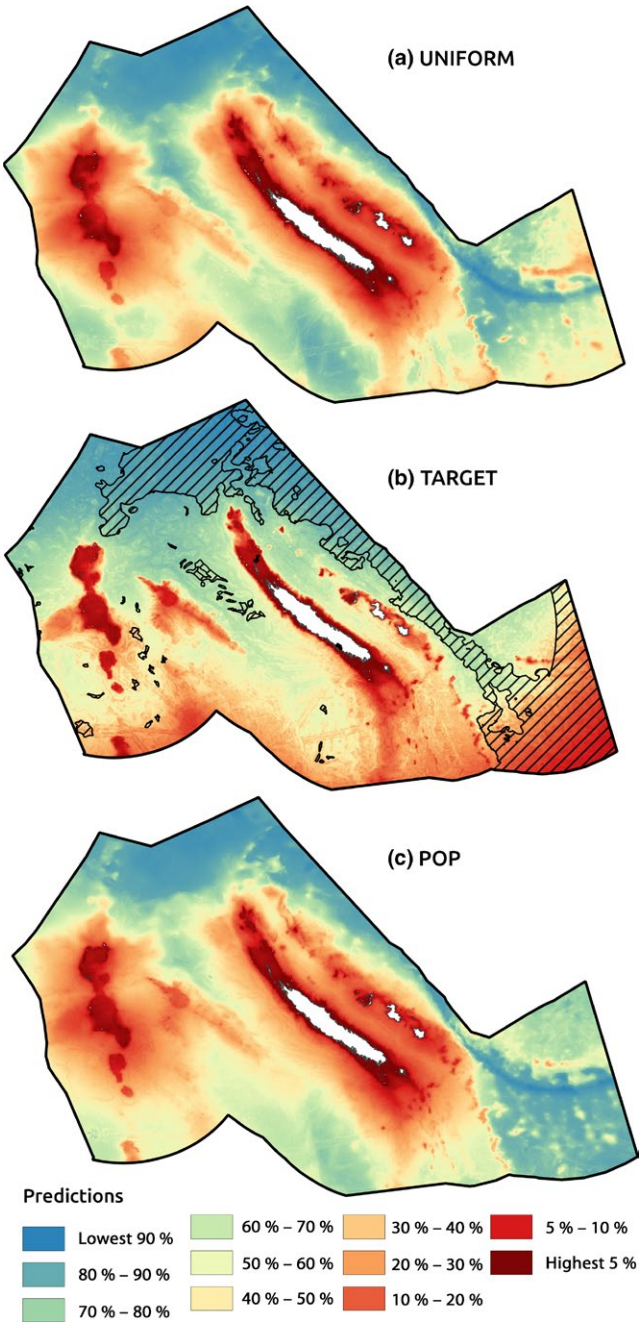

FIGURE 5 Maps of mean predicted humpback whale habitat suitability from citizen science models. Habitat suitability was averaged over 50 cross-validation runs for each statistical algorithm, and a coloured log-scale was applied to values ranging from 0 to 1. Colours represent fixed percentages of probability distributions of the suitability values (e.g., the highest 5% corresponds to the half-of-decile with highest values over each map). Areas of extrapolation where at least one environmental variable expanded outside the range observed in the training dataset are dashed [Colour figure can be viewed at wileyonlinelibrary.com]

to new data, a relationship found both when comparing different statistical algorithms and different tunings of a given statistical algorithm. Whatever the parameterization, BRTs systematically suffered from overfitting and as a result displayed noisy partial dependence plots and predicted maps. The complexity of SVMs and MAXENT models strongly depended on tuning; for instance, radial kernel SVMs were overfitted, whereas the linear kernel version ranked the lowest in explanatory power, along with GLMs, and MAXENT models applied with linear features. The performances of MAXENT models applied with hinge features and of GAMs were intermediate in terms of predictive performance and stability, as measured by ext.AUC and diff.AUC. While GLMs and GAMs were not tested with different parameterizations in this study, it must be noted that tuning may also affect regression-based methods (e.g., through polynomial degree or smoothing basis size).

Considering that many marine SDMs are applied in a spatial conservation planning context (Cleguer, Grech, Garrigue, & Marsh, 2015; Gomez et al., 2017; La Manna, Ronchetti, & Sarà, 2016; Pérez-Jorge et al., 2015; Robinson et al., 2011), it appears that statistical algorithms that intrinsically limit overfitting should be prioritized. Indeed, managers are confronted with extrapolation needs, and SDMs are often implemented to predict the presence of a species in a place/time in which data are not available (Mannocci et al., 2015; Redfern et al., 2017). For instance, with proper tuning, all algorithms predicted the Fairway-Landsdowne banks to be a favourable area for humpback whales. The discovery of this new potential area of humpback whale use is supported by the satellite tracking of two humpback whales (Garrigue et al., 2015), and will help target future research efforts and inform conservation policy. Furthermore, given their wide ranges and mobility, migratory cetacean species are likely to have broad fundamental ecological niches (Guisan & Zimmermann, 2000). Yet, broad niches are generally more difficult to model than narrow ones (Morán-Ordóñez et al., 2017), specifically with MAXENT (Qiao et al., 2015). In this context, overfitting the species–environment relationships in a given study area is likely to strongly affect the transferability of the models (Torres et al., 2015) and underestimate the breadth of the species' niches. On the contrary, approaches such as GAMs and MAXENT with hinge features were capable of modelling humpback whale habitats with a relatively high level of complexity, while conserving a good transferability to novel geographical areas. While using the restricted maximum-likelihood method successfully penalized overfitting in this case study, the complexity of the GAM-fitted responses may be further controlled by tuning the basis size for smoothing (e.g., Mannocci, Roberts et al., 2017), hence also providing the opportunity to include explicit knowledge regarding the species' response to environmental gradients (Austin, 2007).

At last, our statistical comparison underlines that there is no such thing as a universally "best" SDM approach (Qiao et al., 2015). The study goal should be clearly identified upfront, whether it is to produce accurate and/or precise spatial predictions or description of local species–environment relationships. Then, model selection depends on two main issues: the use of evaluation metrics and critical ecological thinking. This study confirms that model evaluation should rely on metrics that promote the best predictive performance while minimizing overfitting. AUC is advantageous because of its threshold-independent nature, but its interpretation in a presence-background

context is not straightforward (Jiménez-Valverde, 2012; Phillips et al., 2006). Diff.AUC cannot be interpreted as easily as in Warren and Seifert (2010) when prevalence and presence-background overlap vary between the training and the evaluation dataset. However, diff.AUC may be used to relatively compare transferability between models as long as it is averaged over consistent cross-validation runs. At last, the combination of diff.AUC with TSS and ext.AUC appeared like a good trade-off to reveal both stability and predictive performance of the models. Moreover, using a truly independent validation dataset can be challenging (Roberts et al., 2017) but ensures the robust estimation of predictive error. Tracking data may constitute such independent data to evaluate or supplement habitat models (e.g., Louzao et al., 2009; Pinto et al., 2016) although it is inherently limited to measuring model sensitivity (i.e., capacity of the model to predict tracking locations as presences), unless other metrics are derived from tracking locations (Pinto et al., 2016). The tracking data have to be contemporaneous with the model calibration dataset and unbiased by sex, social class or tagging location. In this study, most tags were deployed in the South Lagoon (n = 34, 76%); hence, 30% of the track positions were located in this area. As a result, model predictive performance was relatively high for any model that predicted high suitability in the South Lagoon. At last, ARGOS location error tends to be relatively high when tracking large whales (most locations are of quality "B" with precision >50 km; Nicholls et al., 2007). Hence, prior to using these locations for validation of a habitat model, variables could be averaged in the vicinity of the location, or imprecise positions could be filtered out (as was the case in this study). At last, the visual inspection of predicted maps overlapped with the tracks actually proved more useful than the quantification of predictions to this dataset.

Also model evaluation must include the close examination of the variables' relative contributions, partial dependence plots and spatially projected predictions. Indeed, models with similar performances have been found to predict distributions differently because of different functional relationships (Elith & Graham, 2009) and/or because the relative contribution of variables differed (Zanardo et al., 2017). Here, SVMs seem to have deserved their "black-box" reputation (Goldstein et al., 2015) as their ecological interpretation was arduous. For instance, contributions of the predictor variables could only be assessed when using linear kernels, whereas the radial kernels that provided the best predictive performance could not be interpreted as easily. On the contrary, although showing signs of overfitting, BRTs are more interpretable machine-learning approaches that were the only models to identify DEPTH as the dominant variable over DISSURF. In line with this trend, although they relied more on DISSURF than DEPTH, GAMs captured a multimodal relationship relative to DISSURF, revealing preferences for coastal as well as remote waters more than 100 km from shore. While this relationship should be regarded with caution considering the spatially skewed survey effort (favouring specific study areas, such as Antigonia or the South Lagoon), it also shows that complex environmental relationships might be revealed with increased effort in offshore waters. The preference for coastal waters has been extensively documented in humpback whale breeding grounds (Bortolotto et al., 2017; Cartwright et al., 2012; Guidino, Llapapasca, Silva, Alcorta, & Pacheco, 2014; Smith et al., 2012; Trudelle et al., 2016) but only recently has satellite telemetry revealed the use of waters far from any coast or reef (Dulau et al., 2017; Garrigue et al., 2015; Trudelle et al., 2016). Through robust and independent niche modelling, this study confirms that humpback whales are not constrained by proximity to sheltered shorelines, but rather by depth, as whales appear to be preferentially found in shallow waters, both in coastal and offshore areas—a pattern clearly captured by BRTs and GAMs.

Citizen science models aligned with the main ecological relationships highlighted in the research survey models. K490 was particularly influential compared to the research survey models, which could be explained by the high proportion of whales observed by the general public in the lagoons surrounding the mainland that are characterized by relatively high turbidity compared to the open ocean. When sampling bias was corrected in the TARGET method, ecological relationships converged with the research survey model and SST was also found to be particularly influent. The preferred SST range in research survey models (22°C–23°C) was similar to ranges found in neighbouring breeding grounds (GBR, Smith et al., 2012) but relatively low compared to worldwide breeding temperatures reported by Rasmussen et al. (2007). However, as recurrently highlighted in cetacean SDMs (Becker et al., 2017; Redfern et al., 2006) it is hard to differentiate the direct effect of a variable such as SST, from indirect effects due to a correlation with other unmeasured variables, including competition, prey distribution and social interaction.

At last, citizen science models of humpback whale habitat preferences showed promising predictive capacities compared to the research survey models, yet were contingent upon background sampling. Given the wider distribution of background points compared to the research survey dataset, int.AUC and ext.AUC metrics appeared to be inflated (Barve et al., 2011), and the use of comp.AUC was crucial to a robust model evaluation. The TARGET model, which accounted for spatial bias, performed better than the simple UNIFORM model to predict new independent data (comp.AUC) and showed the best ecological match to research survey predictions. However, it is also detrimentally restricted by environmental extrapolation and the background sampling buffer size is likely to have an impact on predictive performance (Barve et al., 2011; Fourcade, Engler, Rödder, & Secondi, 2014). With smaller sample sizes, the predictive capacity of the TARGET model to large areas is likely to decrease. The POP model appears like an interesting alternative in such cases, as it does not restrict the environmental space in which background is sampled, but still accounts for sampling bias. In a conceptual manner, the POP model reflects the assumption that human activity concentrates in coastal areas in the vicinity of cities (Halpern et al., 2015). This assumption is similar in essence to using distance to roads (Phillips et al., 2009) or distance to the coastline (Fithian, Elith, Hastie, & Keith, 2015) as a proxy for land-based observation density. Indeed, the issue of accessibility of

study sites to volunteers has been addressed in land-based datasets (e.g., Tulloch, Mustin, Possingham, Szabo, & Wilson, 2013) but less so in marine studies (Robinson et al., 2011). A variety of other methods have been developed to account for spatial bias in presence-only SDMs. For instance, spatial filtering has been shown to improve predictive performance in several land-based study cases (resampling presence points Boria, Olson, Goodman, & Anderson, 2014; Fourcade et al., 2014; Kramer-Schadt et al., 2013) but was not tested here because it was not considered adapted to the generally small sample sizes recorded in cetacean citizen science programmes. We found that using the TARGET (based on Phillips et al., 2009) and POP sampling methods provided simple and adaptable solutions to account for sampling bias in a cetacean citizen science context.

5 | CONCLUSION

This study provides an in-depth investigation of statistical approaches to highlight the technical challenges associated with cetacean habitat modelling. All algorithms suggested that the endangered New Caledonian population of humpback whales displays a preference for relatively cool and shallow waters regardless of distance to reefs or coasts. Algorithms displayed a range of predictive and descriptive capacity that depended on parameter tuning. BRTs generally characterized ecologically meaningful species–environment relationships, but predictions were fraught with overfitting. SVMs fitted the data closely when using radial kernels, but lacked interpretability and transferability. GAMs stood out as an interesting trade-off with ecologically interpretable results that maintained complexity at a reasonable level to allow good predictive performance over unsampled areas, which is a crucial characteristic in a conservation planning perspective. Considering the wide breadth of migratory cetacean fundamental niches, we conclude that cetacean SDMs produced for conservation purposes should specifically prevent overfitting in order to conserve some transferability to novel geographical areas. Overfitting may be prevented using stratified cross-validation, evaluation with an independent dataset, and an appropriate statistical algorithm and parameter tuning. At last, this study also emphasized the role of citizen science to study wide-ranging species such as cetaceans over large spatial scales. Habitat preference models based on citizen science observations converged with models based on research survey when spatial sampling bias was accounted for in the models. The development of citizen science programmes in marine environments and their application to species distribution models therefore appear like a low-cost and socially valuable research tool and contributor to marine policy.

ACKNOWLEDGMENTS

We gratefully acknowledge the numerous volunteers who participated in fieldwork, especially D. Boillon, C. Bonneville, M. Chambellant, R. Dodémont, M. Oremus, V. Pérard and A. Schaffar. We thank M. Mangeas, R. Pouteau and F. Sullivan for advice and reviewing of this manuscript. We thank everyone who contributed to the citizen science marine mammal dataset in New Caledonia, especially the Protection du Lagon and Caledonie Charter teams. Financial support was provided by Fondation d'Entreprises Total, International Fund for Animal Welfare, the Ministère de la Transition Ecologique et Solidaire, the New Caledonian Government, the New Caledonian Provinces, Vale S.A. and the World Wildlife Fund for Nature. Fieldwork was undertaken under permits issued by the Environment Departments of the New Caledonian provinces and the New Caledonian government.

DATA ACCESSIBILITY

The data used in this manuscript are available via the online Zenodo repository (https://zenodo.org/communities/umr-entropie/ https://doi.org/10.528/zenodo.1065016).

ORCID


Solene Derville http://orcid.org/0000-0002-0380-7921

REFERENCES

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. https://doi.org/10.1371/journal.pone.0063708

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. https://doi.org/10.1111/j.1365-2664.2006.01214.x

Andréfouët, S., Chagnaud, N., Chauvin, C., & Kranenburg, C. J. (2008). *Atlas of French Overseas Coral Reefs*. Nouméa, New Caledonia. Retrieved from http://umr-entropie.ird.nc/index.php/home/ressources/mcrmp/atlas-outre-mer-francais

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

Avila, I. C., Kaschner, K., & Dormann, C. F. (2018). Current global risks to marine mammals: Taking stock of the threats. *Biological Conservation*, *221*(February), 44–58. https://doi.org/10.1016/j.biocon.2018.02.021

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. https://doi.org/10.1016/j.ecolmodel.2011.02.011

Becker, E. A., Forney, K. A., Thayre, B. J., Debich, A., Campbell, G. S., Whitaker, K., … Hildebrand, J. A. (2017). Habitat-based density models for three cetacean species off Southern California illustrate pronounced seasonal differences. *Frontiers in Marine Science*, *4*, 121. https://doi.org/10.3389/fmars.2017.00121

Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., … Frusher, S. (2014). Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, *173*, 144–154. https://doi.org/10.1016/j.biocon.2013.07.037

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. https://doi.org/10.1016/j.ecolmodel.2013.12.012

Bortolotto, G. A., Danilewicz, D., Hammond, P. S., Thomas, L., & Zerbini, A. N. (2017). Whale distribution in a breeding area: spatial models of habitat use and abundance of western South Atlantic humpback whales. *Marine Ecology Progress Series*, *585*, 213–227. https://doi.org/10.3354/meps12393

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory – COLT '92*, 144–152. https://doi.org/10.1145/130385.130401

Bouchet, P. J., Meeuwig, J. J., Salgado Kent, C. P., Letessier, T. B., & Jenner, C. K. (2015). Topographic determinants of mobile vertebrate predator hotspots: current knowledge and future directions. *Biological Reviews*, *90*(3), 699–728. https://doi.org/10.1111/brv.12130

Breen, P., Brown, S., Reid, D., & Rogan, E. (2017). Where is the risk? Integrating a spatial distribution model and a risk assessment to identify areas of cetacean interaction with fisheries in the northeast Atlantic. *Ocean and Coastal Management*, *136*, 148–155. https://doi.org/10.1016/j.ocecoaman.2016.12.001

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1017/CBO9781107415324.004

Brotons, 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. https://doi.org/10.1111/j.0906-7590.2004.03764.x

Bruce, E., Albright, L., Sheehan, S., & Blewitt, M. (2014). Distribution patterns of migrating humpback whales (Megaptera novaeangliae) in Jervis Bay, Australia: A spatial analysis using geographical citizen science data. *Applied Geography*, *54*(February), 83–95. https://doi.org/10.1016/j.apgeog.2014.06.014

Cartwright, R., Gillespie, B., Labonte, K., Mangold, T., Venema, A., Eden, K., & Sullivan, M. (2012). Between a Rock and a hard place : Habitat selection in female-calf humpback whale (*Megaptera novaeangliae*) Pairs on the Hawaiian Breeding Grounds. *PLoS ONE*, *7*(5), e38004. https://doi.org/10.1371/journal.pone.0038004

Childerhouse, S., Jackson, J., Baker, C. S., Gales, N., Clapham, P. J., & Brownell, R. J. (2009). Megaptera novaeangliae (Oceania subpopulation). In: IUCN 2009 IUCN Red List of Threatened Species Version 2009 2. Retrieved from www.iucnredlist.org

Cleguer, C., Grech, A., Garrigue, C., & Marsh, H. (2015). Spatial mismatch between marine protected areas and dugongs in New Caledonia. *Biological Conservation*, *184*, 154–162. https://doi.org/10.1016/j.biocon.2015.01.007

Corkeron, P. J., Minton, G., Collins, T., Findlay, K., Willson, A., & Baldwin, R. (2011). Spatial models of sparse data to inform cetacean conservation planning: An example from Oman. *Endangered Species Research*, *15*, 39–52. https://doi.org/10.3354/esr00367

Derville, S., Constantine, R., Baker, C. S., Dietrich-Steel, D., Oremus, M., & Torres, L. G. (2015). Fine-scale habitat use of the critically endangered Māui dolphins: describing and predicting spatial distribution in a coastal environment. *21rst Biennial Conference on the Biology of Marine Mammals*, San Francisco, USA.

Derville, S., Constantine, R., Baker, C. S., Oremus, M., & Torres, L. G. (2016). Environmental correlates of nearshore habitat distribution by the Critically Endangered Maui dolphin. *Marine Ecology Progress Series*, *551*, 261–275. https://doi.org/10.3354/meps11736

Derville, S., Torres, L., & Garrigue, C. (2018). Social segregation of humpback whales in contrasted coastal and oceanic breeding habitats. *Journal of Mammalogy*, *99*(1), 41–54. https://doi.org/10.1093/jmammal/gyx185

Dormann, C. F., McPherson, J. M., Araújo, M. B., Bivand, R., Bolliger, J., Carl, G., … Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography*, *30*(5), 609–628. https://doi.org/10.1111/j.2007.0906-7590.05171.x

Drake, J. M., Randin, C., & Guisan, A. (2006). Modelling ecological niches with support vector machines. *Journal of Applied Ecology*, *43*(3), 424–432. https://doi.org/10.1111/j.1365-2664.2006.01141.x

Dulau, V., Pinet, P., Geyer, Y., Fayan, J., Mongin, P., Cottarel, G., … Cerchio, S. (2017). Continuous movement behavior of humpback whales during the breeding season in the southwest Indian Ocean: On the road again!. *Movement Ecology*, *5*(1), 11. https://doi.org/10.1186/s40462-017-0101-5

Dunn, J. C., Buchanan, G. M., Cuthbert, R. J., Whittingham, M. J., & McGowan, P. J. K. (2015). Mapping the potential distribution of the Critically Endangered Himalayan Quail Ophrysia superciliosa using proxy species and species distribution modelling. *Bird Conservation International*, *25*(4), 466–478. https://doi.org/10.1017/S095927091400046X

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. https://doi.org/10.1111/j.1600-0587.2008.05505.x

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

Elith, J., Kearney, M., & Phillips, S. (2010). The art of modelling range-shifting species. *Methods in Ecology and Evolution*, *1*(4), 330–342. https://doi.org/10.1111/j.2041-210X.2010.00036.x

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

Engler, R., Guisan, A., & Rechsteiner, L. (2004). An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, *41*(2), 263–274. https://doi.org/10.1111/j.0021-8901.2004.00881.x

Environmental Systems Research Institute (ESRI). (2016). ArcGIS Release 10.3. Redlands, CA.

Fernandez, M., Yesson, C., Gannier, A., Miller, P. I., & Azevedo, J. M. N. (2017). The importance of temporal resolution for niche modelling in dynamic marine environments. *Journal of Biogeography*, *44*(12), 2816–2827. https://doi.org/10.1111/jbi.13080

Fithian, W., Elith, J., Hastie, T., & Keith, D. A. (2015). Bias correction in species distribution models: Pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, *6*, 424–438. Retrieved from http://arxiv.org/abs/1403.7274

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), e97122. https://doi.org/10.1371/journal.pone.0097122

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, *19*(1), 1–67.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. Retrieved from http://www.jstor.org/stable/2699986

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. https://doi.org/10.1016/j.ecolmodel.2015.06.002

Garrigue, C., Clapham, P. J., Geyer, Y., Kennedy, A. S., & Zerbini, A. N. (2015). Satellite tracking reveals novel migratory patterns and the importance of seamounts for endangered South Pacific Humpback Whales. *Royal Society Open Science*, *2*, 150489.

Garrigue, C., Greaves, J., & Chambellant, M. (2001). Characteristics of the New Caledonian Humpback whale population. *Memoirs of the Queensland Museum*, *47*(2), ISSN 0079-8835.

Gilles, A., Viquerat, S., Becker, E. A., Forney, K. A., Geelhoed, S. C. V., Haelters, J., … Aarts, G. (2016). Seasonal habitat-based density models for a marine top predator, the harbour porpoise, in a dynamic environment. *Ecosphere*, *7*(6), e01367. https://doi.org/10.13748/j.cnki.issn1007-7693.2014.04.012

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, *24*(1), 44–65. https://doi.org/10.1080/10618600.2014.907095

Gomez, C., Lawson, J., Kouwenberg, A. L., Moors-Murphy, H., Buren, A., Fuentes-Yaco, C., … Wimmer, T. (2017). Predicted distribution of whales at risk: Identifying priority areas to enhance cetacean monitoring in the Northwest Atlantic Ocean. *Endangered Species Research*, *32*(1), 437–458. https://doi.org/10.3354/esr00823

Guidino, C., Llapapasca, M. A., Silva, S., Alcorta, B., & Pacheco, A. S. (2014). Patterns of spatial and temporal distribution of humpback whales at the southern limit of the Southeast Pacific breeding area. *PLoS One*, *9*(11), e112627. https://doi.org/10.1371/journal.pone.0112627

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

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, *135*, 147–186. https://doi.org/10.1016/S0304-3800(00)00354-9

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1–3), 389–422. https://doi.org/10.1023/A:1012487302797

Halpern, B. S., Frazier, M., Potapenko, J., Casey, K. S., Koenig, K., Longo, C., … Walbridge, S. (2015). Spatial and temporal changes in cumulative human impacts on the world's ocean. *Nature Communications*, *6*(May), 7615. https://doi.org/10.1038/ncomms8615

Hastie, T. J., & Tibshirani, R. J. (1990). Generalized additive models. In *Monographs on statistics and applied probability* (p. 352). London: Chapman and Hall/CRC. Retrieved from http://books.google.com/books?hl=fr&lr=&id=qa29r1Ze1coC&pgis=1

Hirzel, A. H., Hausser, J., Chessel, D., & Perrin, N. (2002). Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? *Ecology*, *83*(7), 2027–2036. https://doi.org/10.1890/0012-9658(2002) 083[2027:ENFAHT]2.0.CO;2

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. https://doi.org/10.1111/j.1466-8238.2011.00683.x

Kaschner, K., Quick, N. J., Jewell, R., Williams, R., & Harris, C. M. (2012). Global coverage of Cetacean line-transect surveys: Status Quo, data gaps and future challenges. *PLoS One*, *7*(9), e44075. https://doi.org/10.1371/journal.pone.0044075

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

La Manna, G., Ronchetti, F., & Sarà, G. (2016). Predicting common bottlenose dolphin habitat preference to dynamically adapt management measures from a Marine Spatial Planning perspective. *Ocean and Coastal Management*, *130*, 317–327. https://doi.org/10.1016/j.ocecoaman.2016.07.004

Legrand, B., Benneveau, A., Jaeger, A., Pinet, P., Potin, G., Jaquemet, S., & Le Corre, M. (2016). Current wintering habitat of an endemic seabird of Réunion Island, Barau's petrel Pterodroma baraui, and predicted changes induced by global warming. *Marine Ecology Progress Series*, *550*(April), 235–248. https://doi.org/10.3354/meps11710

Lindsay, R. E., Constantine, R., Robbins, J., Mattila, D. K., Tagarino, A., & Dennis, T. (2016). Characterising essential breeding habitat for whales informs the development of large-scale Marine Protected Areas in the South Pacific. *Marine Ecology Progress Series*, *548*, 263–275. https://doi.org/10.3354/meps11663

Liu, C., Newell, G., & White, M. (2016). On the selection of thresholds for predicting species occurrence with presence-only data. *Ecology and Evolution*, *6*(1), 337–348. https://doi.org/10.1002/ece3.1878

Louzao, M., Bécares, J., Rodríguez, B., Hyrenbach, K. D., Ruiz, A., & Arcos, J. M. (2009). Combining vessel-based surveys and tracking data to identify key marine areas for seabirds. *Marine Ecology Progress Series*, *391*, 183–197. https://doi.org/10.3354/meps08124

MacKenzie, D. I. (2006). *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence*. Burlington, MA: Academic Press.

Macleod, C. D., Mandleberg, L., Schweder, C., Bannon, S. M., & Pierce, G. J. (2008). A comparison of approaches for modelling the occurrence of marine animals. *Hydrobiologia*, *612*, 21–32. https://doi.org/10.1007/s10750-008-9491-0

Mannocci, L., Boustany, A. M., Roberts, J. J., Palacios, D. M., Dunn, D. C., Halpin, P. N., … Winship, A. J. (2017). Temporal resolutions in species distribution models of highly mobile marine animals: Recommendations for ecologists and managers. *Diversity and Distributions*, *23*(10), 1098–1109. https://doi.org/10.1111/ddi.12609

Mannocci, L., Monestiez, P., Spitz, J., & Ridoux, V. (2015). Extrapolating cetacean densities beyond surveyed regions: Habitat-based predictions in the circumtropical belt. *Journal of Biogeography*, *42*(7), 1267–1280. https://doi.org/10.1111/jbi.12530

Mannocci, L., Roberts, J. J., Miller, D. L., & Halpin, P. N. (2017). Extrapolating cetacean densities to quantitatively assess human impacts on populations in the high seas. *Conservation Biology*, *31*(3), 601–614. https://doi.org/10.1111/cobi.12856.This

Mendez, M., Rosembaum, H. C., Subramaniam, A., Yackulic, C., & Bordino, P. (2010). Isolation by environmental distance in mobile marine species: molecular ecology of franciscana doplhins at their southern range. *Molecular ecology*, *19*, 2212–2228. https://doi.org/10.1111/j.1365-294X.2010.04647.x

Miller, D. L., Burt, M. L., Rexstad, E. A., & Thomas, L. (2013). Spatial models for distance sampling data: Recent developments and future directions. *Methods in Ecology and Evolution*, *4*(11), 1001–1010. https://doi.org/10.1111/2041-210X.12105

Miller-Rushing, A., Primack, R., & Bonney, R. (2012). The history of public participation in ecological research. *Frontiers in Ecology and the Environment*, *10*(6), 285–290. https://doi.org/10.1890/110278

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. https://doi.org/10.1111/faf.12039

Morán-Ordóñez, A., Lahoz-Monfort, J. J., Elith, J., & Wintle, B. A. (2017). Evaluating 318 continental-scale species distribution models over a 60-year prediction horizon: what factors influence the reliability of predictions? *Global Ecology and Biogeography*, *26*(3), 371–384. https://doi.org/10.1111/geb.12545

Nicholls, D. G., Robertson, C. J., & Murray, M. D. (2007). Measuring accuracy and precision for CLS: Argos satellite telemetry locations. *Notornis*, *54*(3), 137–157.

Oppel, S., Meirinho, A., Ramírez, I., Gardner, B., O'Connell, A. F., Miller, P. I., & Louzao, M. (2012). Comparison of five modelling techniques to predict the spatial distribution and abundance of seabirds. *Biological Conservation*, *156*, 94–104. https://doi.org/10.1016/j.biocon.2011.11.013

Pacifici, K., Reich, B. J., Miller, D. A. W., Gardner, B., Stauffer, G., Singh, S., … Collazo, J. A. (2016). Integrating multiple data sources in species distribution modeling : A framework for data fusion. *Ecology*, *98*(3), 840–850.

Pérez-Jorge, S., Pereira, T., Corne, C., Wijtten, Z., Omar, M., Katello, J., … Louzao, M. (2015). Can static habitat protection encompass critical areas for highly mobile marine top predators? Insights from Coastal East Africa. *PLoS One*, *10*(7), e0133265. https://doi.org/10.1371/journal.pone.0133265

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*, 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J. R., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological Applications*, *19*(1), 181–197. https://doi.org/10.1890/07-2153.1

Pinto, C., Thorburn, J. A., Neat, F., Wright, P. J., Wright, S., Scott, B. E., … Travis, J. M. J. (2016). Using individual tracking data to validate the predictions of species distribution models. *Diversity and Distributions*, *2*, 682–693. https://doi.org/10.1111/ddi.12437

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

QGIS Development Team. (2016). QGIS Geographic Information System. Open Source Geospatial Foundation Project. Version Es. Retrieved from http://qgis.osgeo.org

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. https://doi.org/10.1111/2041-210X.12397

R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.r-project.org/.

Radosavljevic, A., & Anderson, R. P. (2014). Making better Maxent models of species distributions: Complexity, overfitting and evaluation. *Journal of Biogeography*, *41*(4), 629–643. https://doi.org/10.1111/jbi.12227

Rasmussen, K., Palacios, D. M., Calambokidis, J., Saborío, M. T., Dalla Rosa, L., Secchi, E. R., … Stone, G. S. (2007). Southern Hemisphere humpback whales wintering off Central America: Insights from water temperature into the longest mammalian migration. *Biology Letters*, *3*(3), 302–305. https://doi.org/10.1098/rsbl.2007.0067

Rayment, W., Dawson, S., & Webster, T. (2015). Breeding status affects fine-scale habitat selection of southern right whales on their wintering grounds. *Journal of Biogeography 42*, 463–474. https://doi.org/10.1111/jbi.12443

Redfern, J. V., Ferguson, M. C., Becker, E. A., Hyrenbach, K. D., Good, C., Barlow, J., … Werner, F. (2006). Techniques for cetacean-habitat modeling. *Marine Ecology Progress Series*, *310*, 271–295. https://doi.org/10.3354/meps310271

Redfern, J. V., Moore, T. J., Fiedler, P. C., de Vos, A., Brownell, R. L., Forney, K. A., … Ballance, L. T. (2017). Predicting cetacean distributions in data-poor marine ecosystems. *Diversity and Distributions*, *23*(4), 394–408. https://doi.org/10.1111/ddi.12537

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., … Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical or phylogenetic structure. *Ecography*, 1–17. https://doi.org/10.1111/ecog.02881

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. https://doi.org/10.1111/j.1466-8238.2010.00636.x

Scales, K. L., Miller, P. I., Hawkes, L. A., Ingram, S. N., Sims, D. W., & Votier, S. C. (2014). On the front line: Frontal zones as priority at-sea conservation areas for mobile marine vertebrates. *Journal of Applied Ecology*, *51*(6), 1575–1583. https://doi.org/10.1111/1365-2664.12330

Smith, J., Grantham, H., Gales, N., Double, M., Noad, M., & Paton, D. (2012). Identification of humpback whale breeding and calving habitat in the Great Barrier Reef. *Marine Ecology Progress Series*, *447*(Harwood 2001), 259–272. https://doi.org/10.3354/meps09462

Stirling, D. A., Boulcott, P., Scott, B. E., & Wright, P. J. (2016). Using verified species distribution models to inform the conservation of a rare marine species. *Diversity and Distributions*, *22*(7), 808–822. https://doi.org/10.1111/ddi.12447

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285–1293.

Thorne, L. H., Johnston, D. W., Urban, D. L., Tyne, J., Bejder, L., Baird, R. W., … Chapla Hill, M. (2012). Predictive modeling of spinner dolphin (*Stenella longirostris*) resting habitat in the main Hawaiian Islands. *PLoS One*, *7*(8), https://doi.org/10.1371/journal.pone.0043167

Tiago, P., Pereira, H. M., & Capinha, C. (2017). Using citizen science data to estimate climatic niches and species distributions. *Basic and Applied Ecology*, *20*, 75–85. https://doi.org/10.1016/j.baae.2017.04.001

Tobeña, M., Prieto, R., Machete, M., & Silva, M. A. (2016). Modeling the potential distribution and richness of Cetaceans in the Azores from fisheries observer program data. *Frontiers in Marine Science*, *3*(202), 1–19. https://doi.org/10.3389/fmars.2016.00202

Torres, L. G., Read, A. J., & Halpen, P. (2008). Fine-scale habitat modelling of top marine predator: Do prey data improve predictive capacity? *Ecological Applications*, *18*(7), 1702–1717. https://doi.org/10.1890/07-1455.1

Torres, L. G., Smith, T. D., Sutton, P., MacDiarmid, A., Bannister, J., & Miyashita, T. (2013). From exploitation to conservation: Habitat models using whaling data predict distribution patterns and threat exposure of an endangered whale. *Diversity and Distribution*, *19*, 1138–1152. https://doi.org/10.1111/ddi.12069

Torres, L. G., Sutton, P. J. H., Thompson, D. R., Delord, K., Weimerskirch, H., Sagar, P. M., … Phillips, R. A. (2015). Poor transferability of species distribution models for a pelagic predator, the grey petrel, indicates contrasting habitat preferences across Ocean Basins. *PLoS One*, *10*(3), e0120014. https://doi.org/10.1371/journal.pone.0120014

Trudelle, L., Cerchio, S., Zerbini, A. N., Geyer, Y., Mayer, F., Jung, J., … Adam, O. (2016). Influence of environmental parameters on movements and habitat utilization of humpback whales in the Madagascar breeding ground Subject Category: Subject Areas: Royal Society Open. *Science*, *3*, 160616.

Tulloch, A. I. T., Mustin, K., Possingham, H. P., Szabo, J. K., & Wilson, K. A. (2013). To boldly go where no volunteer has gone before: Predicting volunteer activity to prioritize surveys at the landscape scale. *Diversity and Distributions*, *19*(4), 465–480. https://doi.org/10.1111/j.1472-4642.2012.00947.x

Warren, D. L., & Seifert, S. N. (2010). Ecological niche modeling in Maxent: The importance of model complexity and the performance of model selection criteria. *Ecological Applications*, *21*(2), 335–342. https://doi.org/10.1890/10-1171.1

Zanardo, N., Parra, G. J., Passadore, C., & Möller, L. M. (2017). Ensemble modelling of southern Australian bottlenose dolphin Tursiops sp. distribution reveals important habitats and their potential ecological function. *Marine Ecology Progress Series*, *569*, 253–266. https://doi.org/10.3354/meps12091

BIOSKETCH

Solène Derville is a PhD student at the French National Research Institute for Sustainable Development where she conducts research on the spatial ecology of humpback whales in Oceania. She is interested in innovative and multidisciplinary methods to study the multiscale space use of marine megafauna. The ENTROPIE team (IRD) is leading research on biodiversity, functional ecology, conservation and management of tropical marine ecosystems. The Geospatial Ecology of Marine Megafauna Laboratory (GEMM lab, OSU) focuses on the ecology, behaviour, health, and conservation of marine megafauna including cetaceans, pinnipeds, seabirds and sharks.

Author contributions: C.G. collected the data; S.D., C.G. and L.G.T. conceived the ideas and designed the methodology; S.D., L.G.T and C.I. analysed the data; S.D. led the writing.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Derville S, Torres LG, Iovan C, Garrigue C. Finding the right fit: Comparative cetacean distribution models using multiple data sources and statistical approaches. *Divers Distrib.* 2018;24:1657–1673. https://doi.org/10.1111/ddi.12782