IOP Conference Series: Earth and
Environmental Science

PAPER • OPEN ACCESS

Predicting cetacean habitats beyond surveyed regions in Indonesian waters: species distribution model transferability may not always be preferable

To cite this article: Achmad Sahri *et al* 2023 *IOP Conf. Ser.: Earth Environ. Sci.* **1276** 012054

View the article online for updates and enhancements.

Predicting cetacean habitats beyond surveyed regions in Indonesian waters: species distribution model transferability may not always be preferable

Achmad Sahri[1*], Floor Maters[2,3] Mochamad I H Putra[4], Putu L K Mustika[5,6], Danielle Kreb[7] and Ron J van Lammeren[2]

[1]Research Center for Oceanography, National Research and Innovation Agency, Jl. Pasir Putih 1, Ancol Timur, Jakarta 14430, Indonesia
[2]Laboratory of Geo-Information Science and Remote Sensing, Wageningen University and Research, P.O. Box 47, 6700 AA Wageningen, the Netherlands
[3]Coastal and river engineering, Royal Haskoning DHV, the Netherlands
[4]Elasmobranch and Charismatic Species Program, Konservasi Indonesia. Jl. Pejaten Barat No. 16 A, Pasar Minggu, Jakarta Selatan 12550, Indonesia
[5]Cetacean Sirenian Indonesia (CETASI), Jakarta Utara, Indonesia
[6]James Cook University, College of Business, Law and Governance, Townsville, Queensland, Australia
[7]Yayasan Konservasi RASI (Rare Aquatic Species of Indonesia), Samarinda, Kalimantan Timur, Indonesia
*Corresponding author's email address

Abstract. Understanding the distribution of cetaceans in Indonesian waters is imperative for their conservation management, however such information is lacking for the country. Our study predicted the species distributions of two cetacean species (common bottlenose dolphin and sperm whale) beyond the surveyed regions in Indonesian Exclusive Economic Zone (EEZ). This was done by using a combination of presence-only data, randomly generated pseudo-absences and environmental predictors variables within the Biomod2 framework in R. Ten potential predictor variables were identified, of which five were selected after correlation tests. Local Random Forest models were built to the extent of four small study regions, and later projected to the whole Indonesian EEZ. The common bottlenose dolphin local models showed preference for areas close to the coast and shallower waters. Sperm whale local predictions were located further into the open waters and at deeper waters. The extrapolated predictions into the Indonesian EEZ, however, showed some unexpected results. The high occurrences for common bottlenose dolphins were not only located close to the islands, but also more into open waters. In contrast, sperm whale distributions have high occurrences near coastal areas and in the vicinity of islands than in the open oceans. This information suggested that the transferability of species distribution models may not always be preferable, because provide low accuracy. Sighting data, choices of variables and model settings influenced the outcome of the extrapolated models. Despite the unpreferable of the extrapolations, the results are still beneficial for cetacean conservation purposes, since the study was able to identify potential habitats in unsurveyed regions.

1. Introduction

It is globally recognized that conservation is needed to preserve marine biodiversity, since the health of worldwide marine ecosystems is deteriorating and causes the biodiversity loss [1,2]. One group of marine animals of great conservation concern are the cetaceans: whales, dolphins and porpoises [3]. The International Union for Conservation of Nature has listed 27 (30%) out of the currently 89 recognized cetacean species as being threatened by extinction on their Red List of threatened animals [4-6].

Indonesian waters are important for cetacean conservation as they harbor 36 cetacean species [7], entail 40% of all existing cetaceans [5]. All marine mammal species living in the Indonesian waters are protected species according to the country's regulations [7]. On the international level, the United Nations Convention on the Law of the Sea (UNCLOS) provides a legal framework to protect marine ecosystems all over the world. This convention demands all coastal countries to protect the marine environment in their own Exclusive Economic Zones (EEZs) [8], such as through the establishment of Marine Protected Areas (MPAs). The Convention on Biological Diversity (CBD) also calls on parties to establish a system of protected areas to conserve biological diversity [7].

To improve the effectiveness of cetacean conservation, detailed information on their distribution and abundance is required [9,10]. Cetaceans are top predators and because of this position in the food web, they are of ecological importance for the conservation of marine biodiversity [11-13]. However, data on the occurrence of cetaceans is sparse [6,11]. This lack of information is due to many factors such as the low density and elusive trait of the taxa [14], its mobile capability that are hard to detect [6,14,15], and time and financial constraints to perform dedicated cetacean surveys [16].

To overcome this problem, species distribution models (SDMs) are used by many ecologists and biogeographers [17-19]. SDMs predict species distributions by relating sighting data to environmental predictors [10,11,20]. The sighting data provides better understanding of species environmental favors [21]. These SDMs give the opportunity to extrapolate the observed distributions into unsurveyed locations and interpret the data-gaps [11]. Most explorative studies however have focused on species distribution modelling simply filling gaps within the surveyed areas, as these generate reliable results more easily [22,23]. Using SDMs to extrapolate distributions to unsurveyed areas is a subject that has been studied less, but is emerging because of the call for models that go beyond the studied regions to solve large-scale conservation issues, such as cetacean distributions in unsurveyed Indonesian waters. Heikkinen et al. [23] studied ten different modelling methods to assess the transferability (extrapolative accuracy) of the SDM methods. Some models did predict better than others for unsurveyed area, however, Heikkinen et al. [23] could not provide a general rule to specify SDMs that provides consistent and improved transferability. Mannocci et al. [11] successfully extrapolated the distributions of three cetacean guilds into unsurveyed areas in the circumtropical belt. Despite these promising results, Redfern et al. [22] found that their model predictions of blue whale habitats, were not transferable into unsurveyed areas. However, they were able to identify potential habitats in regions without sighting data using data from multiple ecosystems [22]. The extrapolation of cetacean distributions is beneficial and could be applied in other areas and for other cetacean species. We predicted cetacean habitats beyond surveyed regions in Indonesian waters to assess the model transferability for two cetacean species.

2. Materials and Methods

2.1. Study Area

Four regions in Indonesia adjusted from the seascapes and ecoregions [24] were selected due to data availability for building local models i.e., Bird's Head, Lesser Sunda, North East Borneo and South East Sulawesi (Figure 1). The Exclusive Economic Zones (EEZs) of Indonesia (Figure 1) were used as the extent of the local model extrapolations to understand the distributions of cetaceans in unsurveyed regions.

2.2. Cetacean occurrence data and pseudo-absence data

The two cetacean species selected for this study were the common bottlenose dolphin (*Tursiops truncatus*) and the sperm whale (*Physeter macrocephalus*). These two cetacean species were not only chosen because of the adequate amount of sighting data, but also because they have quite distinct habitat preferences. Common bottlenose dolphin lives in shallow and coastal waters during the entire day, while sperm whales almost always inhabit the deeper open waters rather than shallow coastal areas [5]. These habitat preference differences may thus give variations during the SDM process. The occurrence data used in this study was gathered from Sahri et al. [25] of which was collected between 2000 and 2018. All data only consisted of presence data.

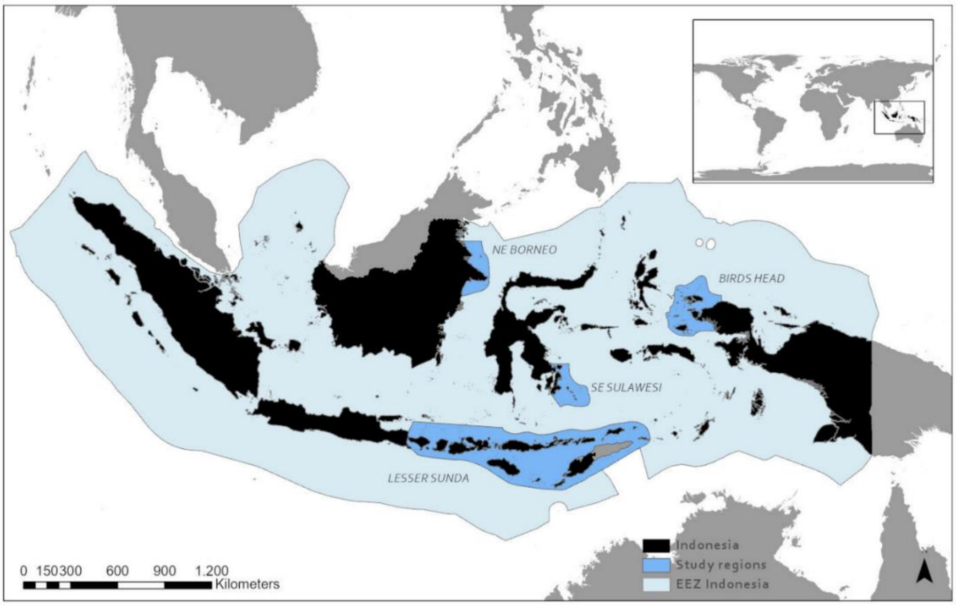

Figure 1. The four study regions within Indonesia as expressed in bright blue and the Exclusive Economic Zones (EEZs) of Indonesia as depicted in light blue.

For the common bottlenose dolphin, 289 sighting points were available and for the sperm whale, 94 points (Table 1). No sighting points were available in Bird's Head for the sperm whale, while only 10 sighting points of sperm whales were available in North-East Borneo. Jimenez-Valverde [26] suggested the minimum amount of sighting points for SDMs is 20 sighting points. It was thus decided that the amount of 10 points for sperm whales in North-East Borneo was not enough to compute a meaningful model. Although the two cetacean species may migrate differently over the particular seasons, this temporal dimension was not taken into account in our study due to the small number of samples. The data showed too low sightings per season and did not distinguish individuals or pods. For these reasons, meaningful seasonal distribution differences could not be generated.

Table 1. Sighting data of two cetacean species used in SDMs obtained from Sahri et al. [25].

	Bird's Head	Lesser Sunda	NE Borneo	SE Sulawesi	Total
Common bottlenose dolphin					
Initial sighting points	127	69	66	27	289
Sighting points left after rarefying	33	26	26	19	
Pseudo-absences	3300	2600	2600	1900	
SAC points removed	94	43	40	8	
Sperm whale					
Initial sighting points		64	(10)	20	94
Sighting points left after rarefying		34		15	
Pseudo-absences		3400		1500	
SAC points removed		30		5	

Before using the initial sighting data in the SDMs, the spatial auto correlation (SAC) of the sightings was examined. Decreasing SAC can prevent the creation of incorrect models resulting in wrong assumptions about the effects of environmental conditions on the species distribution [27]. This examination was done by rarefying the sighting data [28] using "Spatially rarefy occurrence data tool" in SDMtoolbox version 2.4 in ArcGIS [29].

Since absence records were not available but were needed for SDMs, 'pseudo-absence' records were created by randomly sampling points in the study regions, but excluding the locations where presence records were known [30-32]. Pseudo-absence records were created for each species in all four study regions. Both Barbet-Massin et al. [33] and Lobo & Tognelli [34] suggested to use a large amount of pseudo-absences in SDMs. However, as only a limited amount of presence points was available for this study, it was decided to use an amount of pseudo-absence points 100 times of the amount of presences as selected for SDMs (Table 1), following Lobo & Tognelli [34], to avoid a modeling outbalance [33,34]. The presences and pseudo-absences weighted in a balanced manner has resulted very good models as reported in previous cetacean studies [35-37]. These pseudo-absences were created using the "create random point" tool in ArcGIS.

2.3. Environmental predictor variables

Environmental datasets were also required to run SDMs. The dataset included topographic and climatic variables of importance in predicting the distribution of a species due to locations a species prefers [38]. The best environmental variable would be the spatial distribution of the prey of the cetacean species of interest. However, such information was lacking and thus environmental predictor variables were used that serve as an indicator of prey presence as done in previous studies [11,38]. Finally, based on a literature study, ten potential variables were selected: bathymetry, slope, distance to isobaths (200, 1000 and 2500 m), distance to the coast, distance to shelf, sea surface temperature, chlorophyll-a and sea surface salinity (Table 2). Table 2 provides the reasons and sources of the variables.

Table 2. Environmental predictor variables as used in this study.

Environmental variable	Source	Reason of including variable
Bathymetry (m)	a	Shallow waters are preferred by common bottlenose dolphins, while sperm whales prefer deeper waters [5]
Slope (%)	b	Steep slopes are associated with higher primary production and availability of prey [11,45]
Distance to isobaths (km): 200 m 1000 m 2500 m	b	Common bottlenose dolphins prefer waters that are not very deep, while sperm whales prefer deeper waters [5,46,47]
Distance to (km): Coast	c	Low distances to the coast are preferred by bottlenose dolphins, while sperm whales prefer greater distances [5]
Shelf	d	Upwelling around the shelf breaks can be associated with higher availability of prey [48]
Sea surface temperature annual (SST) (◦C)	e	Surface temperature affects the availability of cetacean's prey [11]
Chlorophyll-a annual (Chl-a) (mg.m^{-3})	e	High concentrations of chlorophyll-a indicate a high concentration of phytoplankton [11]
Sea surface salinity annual (SSS) (PSU)	f	Surface salinity affects the prey aggregation of cetaceans [22]

[a] GEBCO (https://www.gebco.net/).
[b] GEBCO (https://www.gebco.net/) and ArcGIS derived.
[c] Indonesian Geospatial Information Agency and ArcGIS derived.
[d] Seafloor Geomorphic Features Map [39] and ArcGIS derived.
[e] Aqua MODIS (https://oceancolor.gsfc.nasa.gov).
[f] SODA 3.3.1 (http://apdrc.soest.hawaii.edu).

The bathymetric dataset was derived from the General Bathymetric Chart of the Ocean (GEBCO). This same dataset was used to compute the second and third environmental layers: the slope and the distances to the isobaths. The distance to the coast was derived from the Indonesian Geospatial Information Agency. The shelf data was derived from the Global Seafloor Geomorphic Features Map [39]. Sea surface temperature (SST) and chlorophyll-a (Chl-a) data was gathered from MODIS-aqua

(2000-2018). Sea surface salinity (SSS) data was derived from the SODA 3.3.1 model (2000-2018). All datasets were quadratically tessellated by a 1 x 1 km resolution.

It is important to check whether the environmental predictor variables did not contain any high inter-correlations (multicollinearity), as this will affect the modelling outcome [18,38,40,41]. The multicollinearity was checked by calculating the Pearson's correlations among the environmental predictor variables in all four study regions and were shown in heatmaps (Figure A1 in Appendices). Correlations with a Pearson coefficient of 0.75 or higher are known as high to very high correlations [42]. All layer combinations that had a correlation coefficient higher than 0.75 were thus reduced to only one environmental dataset [25]. It resulted in the use of only five out of ten potential predictor variables in SDMs, i.e., bathymetry, slope, distance to coast, chlorophyll-a and sea surface salinity. It was in line with the many suggestions to use a limited amount of predictor variables when extrapolating into unsurveyed areas, to avoid too complex models [43] and to achieve a higher model transferability [43,44].

2.4. Species Distribution Model: Random Forest (RF)

Random Forest (RF) was chosen as a suitable modelling technique as the measurement scale fits the data used in this study (our preliminary study, not shown) and decreased over-fitting problems and was therefore known as an accurate modelling technique. RF is known to perform well with handling non-linear data, which is needed when modelling species distributions [49,50]. RF is a type of correlative and predictive modeling class. A correlative model estimates the environment that a species favors by relating the species occurrence to environmental predictors that can be expected to influence the species appearance [21]. A predictive model finds a relationship between the species presence and variables, then use this relation to predict the distribution of a species [10,17,19].

Our RF models were created using the Biomod2 package in R. In the Biomod2 framework, some choices of settings were made before running the models. For the RF algorithm, the first choice to be made was between regression and classification. With RF regression, predictions were made to a continuous output variable, while with RF classification the output was a discrete variable [51]. For our study, a discrete output variable was generated, which results in the choice of selecting RF classification. Another decision to be made was the number of trees to grow. It is important to not set this number too low, as this would disable each input row to get predicted several times [51]. The default number of trees within Biomod2 is 500 trees, and was good for our study as the number was high enough to let each input row got predicted multiple times and to serve the randomness to prevent overfitting [52]. A number of 500 trees thus was grown in each model. The best number of variables (*mtry*) that were randomly sampled as option for each split differ for RF regression and RF classification. Breiman [51] recommended a value of (sqrt(p) for (*mtry*) for classification models, with p being the number of variables. Lastly, a node size of 1 was chosen, which was the recommended value for classification models. The node size is the minimum size of terminal nodes, in the case of increasing this number the trees that are grown get smaller [51,52].

For each species and study regions, 10 evaluation runs were performed. The data was split into two subsets, one to calculate the model (70% of the data) and one to evaluate these models (30% of the data). The 70/30 division is a default as used in many SDMs studies. This data split in combination with the evaluation runs and the 500 trees that were built each run granted for randomness and thus a robust test of the models [52]. Thuiller et al. [52] especially recommended this strategy when no independent dataset for evaluation is available. Ideally, an independent dataset with presence and absence points to evaluate the outcome of the models will give a better performance, but for our study no independent dataset was available. We also decided not to use the occurrence points that were deleted after the rarefying process as dataset for evaluation. In some study regions, a small amount of occurrence points was removed (e.g., 5 for sperm whales in SE Sulawesi, Table 2), and thus was not suitable for model evaluation.

2.5. *Extrapolation to unsurveyed regions*

The prediction of species distributions in the whole Indonesian EEZ was done by extrapolating the local models for each species into the extent of the Indonesian EEZ. In ecology, extrapolations are usually done over space, as in this study, and into a different time frame e.g., the future or past [10,53]. The latter was not the scope of this study. To project the model output over space, we reused the same environmental variables with the extent of the Indonesian EEZ. Within this process, sighting data was used to predict distributions into an environment that was not surveyed. The extrapolation of cetacean distributions to Indonesian EEZ was done using the "Biomod2 projection" tool in R. The projection tool unfortunately only accepts the Random Forest output of one study region at the same time. Thus, the extrapolations were done per individual study region and then were averaged to create one map per species. For example, the extrapolated projection of common bottlenose dolphin was based on the Random Forest results of the four study regions and only on two regions for the sperm whale.

2.6. *Evaluation of SDMs*

Before drawing conclusions about the output that SDMs generate, it is important to assess the robustness of the model. Pearce & Ferrier [54] argued that there are two aspects that need to be measured when evaluating the performance of SDMs: discrimination ability and reliability. To examine the feasibility of the Random Forest models of this study, the models were evaluated by using the area under the receiver operating characteristic curve (AUC) and the true skill statistic (TSS). The AUC is one of the most widely used accuracy indices, and considered the standard index for SDMs [55]. The AUC of an SDM is the likelihood that the SDM will rate a randomly selected presence location higher than a randomly selected absence location [54,56]. The TSS is an index that compares the number of accurate predictions minus those due to random guessing to an assumed set of excellent predictions [57]. Both measures were chosen because they widely use in evaluating species distribution models [58,59]. Models with AUC values of >0.7 and TSS >0.4 can be considered as meaningful models [60].

The Random Forest models were not only evaluated using these two-evaluation metrics as they cannot fully be relied on due to the scarcity of data and the characteristic of pseudo-absences. Additionally, the models' sensitivity was examined by evaluating the importance of each of the environmental variables and by creating response curves using the "response.plot2" function in Biomod2 [52,61]. These curves gave insight into the extent in which each variable contributed to the outcome of the models, presenting the sensitive level of the models regarding each individual variable.

3. Results

3.1. *Performance of Random Forest Model*

The Random Forest models of the common bottlenose dolphin in Lesser Sunda scored best with an AUC value of 0.858 and a TSS value of 0.676. The common bottlenose dolphin model in Bird's Head also met the thresholds of AUC and TSS although with lower values (0.778 and 0.543, respectively) than those of in Lesser Sunda. Both common bottlenose dolphin models in NE Borneo and SE Sulawesi were not considered as meaningful models according to their AUC and TSS values. The sperm whale models in Lesser Sunda had an AUC of 0.834 and TSS of 0.580 indicating meaningful models. The sperm whale models in SE Sulawesi, as in common bottlenose dolphin, all scored below the AUC and TSS thresholds.

Table 3. Averaged AUC and TSS values (AUC >0.7 and TSS >0.4 in **bold** letters).

	Bird's Head	Lesser Sunda	NE Borneo	SE Sulawesi
Common bottlenose dolphin				
AUC	**0.778**	**0.858**	0.665	0.643
TSS	**0.543**	**0.676**	0.359	0.325
Sperm whale				
AUC		**0.834**		0.598
TSS		**0.580**		0.316

3.2. Variable Importance

'Distance to coast' was the most important in predicting the common bottlenose dolphin distributions in three out of four study regions ranging from 30.3% (Bird's Head) to 39% (SE Sulawesi). Only in NE Borneo, chlorophyll-a was the most important variable for the common bottlenose dolphin, with a value of 27.6%. The second most important variable differed over the study regions, being sea surface salinity (SSS) in Bird's Head (19.3%), bathymetry in both Lesser Sunda (24.5%) and SE Sulawesi (24.6%), and distance to coast in NE Borneo (23.2%). The third most important variable for Bird's Head and NE Borneo was bathymetry (16.8% and 23.1%). Chlorophyll-a was the third variable for both Lesser Sunda and SE Sulawesi with 21.1% and 17.1%, respectively. Slope was the variable of least importance for all regions, ranging from 9.4% (SE Sulawesi) to 16.9% (Bird's Head).

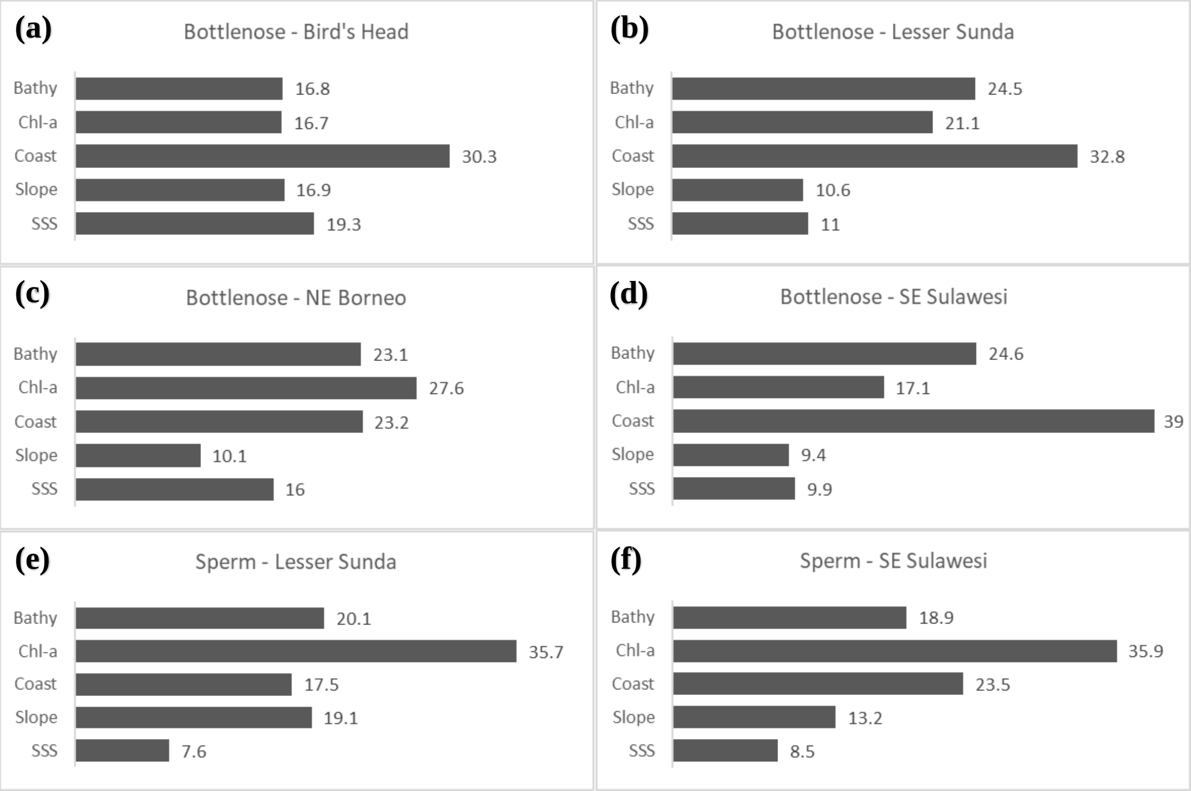

Figure 2. Relative importance of the variables (normalized to sum 100%) per region for common bottlenose dolphin (a-d) and sperm whale (e-f).

The most important variable in predicting sperm whale distributions in both Lesser Sunda and SE Sulawesi was chlorophyll-a with slightly similar values of 35.7% and 35.9%. The second most important variable differed per region, being bathymetry in Lesser Sunda (20.1%) and distance to coast in SE Sulawesi (23.5%). Slope was the third most important variable for sperm whale models in Lesser Sunda. Contrasting with the common bottlenose dolphin models, slope in sperm whale model was found to be one of most important variables. Bathymetry scored as the third most important variable in SE Sulawesi with value of 18.9%. Sea surface salinity (SSS) was the variable of least importance for two regions with values no more than 8.5 %.

3.3. Response Curves

The array of response curves (Figure 3) shows the results of the ten model runs for each variable and each study region. The curves show how the probability of occurrence changes by varying one variable and how sensitive the model was to that variable. Interactions between variables were not included here [52]. Overall, many variabilities exist between species and study regions.

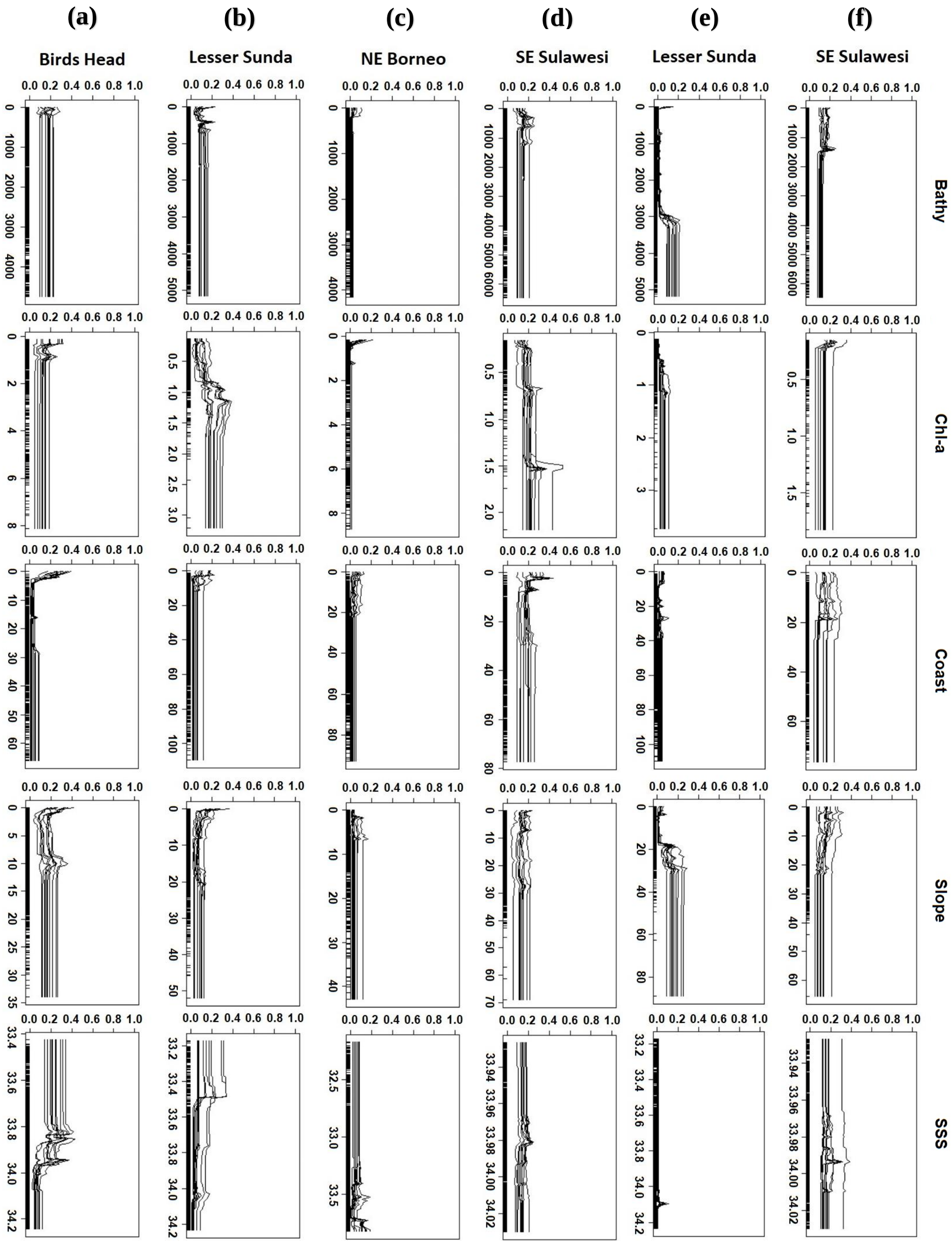

Figure 3. Response Curves of Random Forest models per region for common bottlenose dolphin (a-d) and sperm whale (e-f). The y-axis shows the probability, the x-axis the value range of the variable, subsequently units in m, mg.m^{-3}, km, %, and PSU. The value range may vary per region.

The bathymetry response curves of the common bottlenose dolphin did not show much variation and seem to be quite constant across regions with high probability of occurrence increased in waters < 800 m depth. The bathymetry curves for the sperm whale models differ a lot over the two study regions. In Lesser Sunda, the probability of occurrence increases around a depth of 3000 m, while the response curves of SE Sulawesi show a slight increase between 0 and 1500 m.

The chlorophyll-a response curves are very different between regions and species. In Bird's Head models for common bottlenose dolphins show no detectible changes. In Lesser Sunda, an increase of probability was seen between 1 and 1.5 mg.m^{-3} for common bottlenose dolphin, while a negligible increase was visible around the value of 1 mg.m^{-3} for sperm whale. The NE Borneo curves of the common bottlenose dolphin did not show any preference of chlorophyll-a level. The curves of SE Sulawesi for the common bottlenose dolphin showed small increases of probability around 0.7 and 1.5 mg.m^{-3}, while the sperm whale probability did not change much.

When looking at the distance to coast response curves, common bottlenose dolphins tend to prefer near-shore areas (within 8 km). As expected, the response curves for sperm whale showed a preference of areas that are on greater distance from the coast (25 km) than the dolphin. However, in Lesser Sunda the preference for distance to coast was not clearly visible, while in SE Sulawesi the preference was visible.

The response curves of the slope for common bottlenose dolphin show that this species preferred areas with gentle slope in all regions. The response curves for NE Borneo did not really show any increases of occurrence probability. The response curves of sperm whales showed opposite patterns for two regions. While in Lesser Sunda an increase in probability of occurrence started around a slope of 20%, the increase in probability of occurrence in SE Sulawesi started around areas without any slope and decreased around a slope of 20%.

The sea surface salinity (SSS) response curves for common bottlenose dolphin were a bit more variable across the regions with small increases at salinity level around 33.8 PSU in Bird's Head, 33.4 PSU in Lesser Sunda, 33.5 in NE Borneo, and 33.96 PSU in SE Sulawesi. The SSS curves of sperm whales in Lesser Sunda show a small increase at a salinity level of 34.1 PSU, while being completely constant over the rest of the range. The curves in SE Sulawesi show a bit more fluctuation but also show higher probability of occurrence around salinity levels of 33.96-34.1 PSU.

3.4. Cetacean Distributions in Local Models

Random Forest local models resulted different predicted distributions for two cetacean species in different regions. The differences in distribution were analyzed by visualizing averaged occurrence probability maps and its variability and comparing these results over the two cetacean species. The variability of the modelling results can be seen from the standard deviations of the averaged models for each species (Figure A2 in Appendices). It is clear that the standard deviations are similar to their respective averaged probabilities (Figure 4 and Figure A2 in Appendices) i.e., areas with higher probabilities have a higher variability as well and vice versa. The lower the variability the more certain the predicted models.

In the Bird's Head, the common bottlenose dolphins were distributed around the islands and shores, and were unlikely to be sighted away from the coast. In Lesser Sunda, the two cetacean species show different distributions based on their probability maps. High probabilities of common bottlenose dolphins were located close to the coast, up to 8 km away from the coast. This species was also found between islands. The high probabilities of sperm whale were not predicted to occur between islands. Instead, the distribution was predicted farther away from coastal areas and more into the open waters, up to 40 km away from the coast.

In NE Borneo, probabilities of common bottlenose dolphin were especially high around the small islands, although some high probabilities were also visible further away from the land into the water. The high probabilities were not predicted further than 30 km from the coast. High probabilities were also visible in between islands, suggesting that the dolphin was predicted to also forage between the islands. In SE Sulawesi, common bottlenose dolphin preferred coastal areas rather than the open waters with high probabilities were more clustered closely to the islands. In contrast, high probabilities of sperm whale occurrence were predicted to be further away from the coast and islands (up to 55 km) into more open waters. This information indicates that the Random Forest models have succeeded in predicting differences in the species distributions of the two cetacean species of interest.

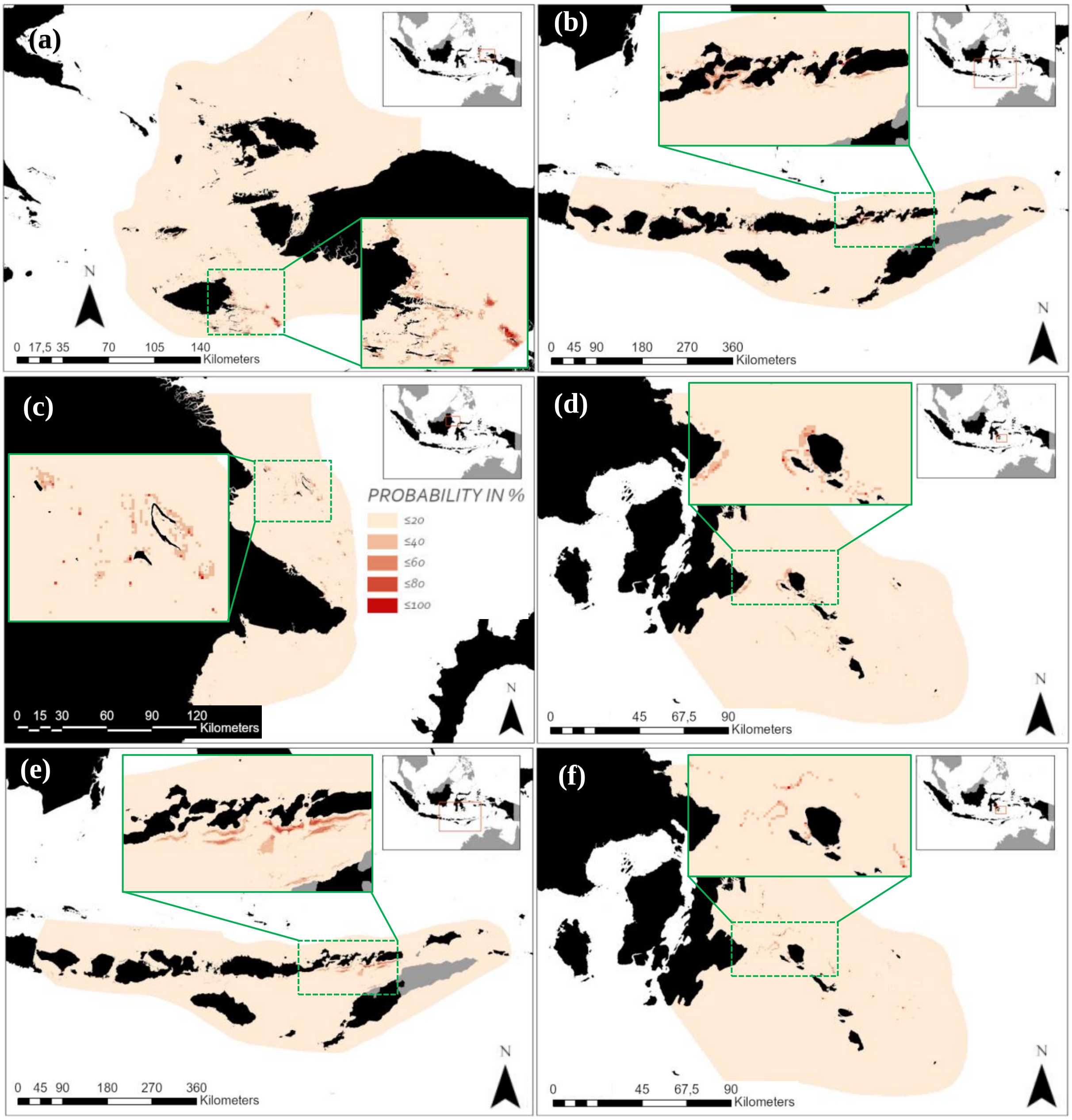

Figure 4. Local Random Forest models per region for common bottlenose dolphin (a-d) and sperm whale (e-f). Cells with a light red colour represent low occurrence probabilities, while cells with a darker red colour show high probability of occurrence. The close-up view in each window is presented.

3.5. Extrapolated Distribution Probabilities to Unsurveyed Regions

When comparing the results, it was evident that the extrapolation predictions of both cetacean species were not comparable to the local model results. The extrapolated distribution of the common bottlenose dolphin to the extent of the Indonesian EEZ showed that a concentration of occurrence probabilities of ≥ 40 % was visible in the North-West and South-East of Indonesia (Figure 5a). The cells with a probability of ≥ 60 % are not only located close to the islands, but also a bit further from the coastal areas in the open water. This is surprising as the local model results showed a predicted preference for locations that close to the coast. The variability maps of the extrapolated models for the common bottlenose dolphin were similar to the local models with high occurrence probabilities show high variabilities as well (Figure A3 in Appendices). Higher variability is thus especially visible in the North-West and South-East of Indonesia, but barely exceeding standard deviations of 20 (Figure A3 in Appendices).

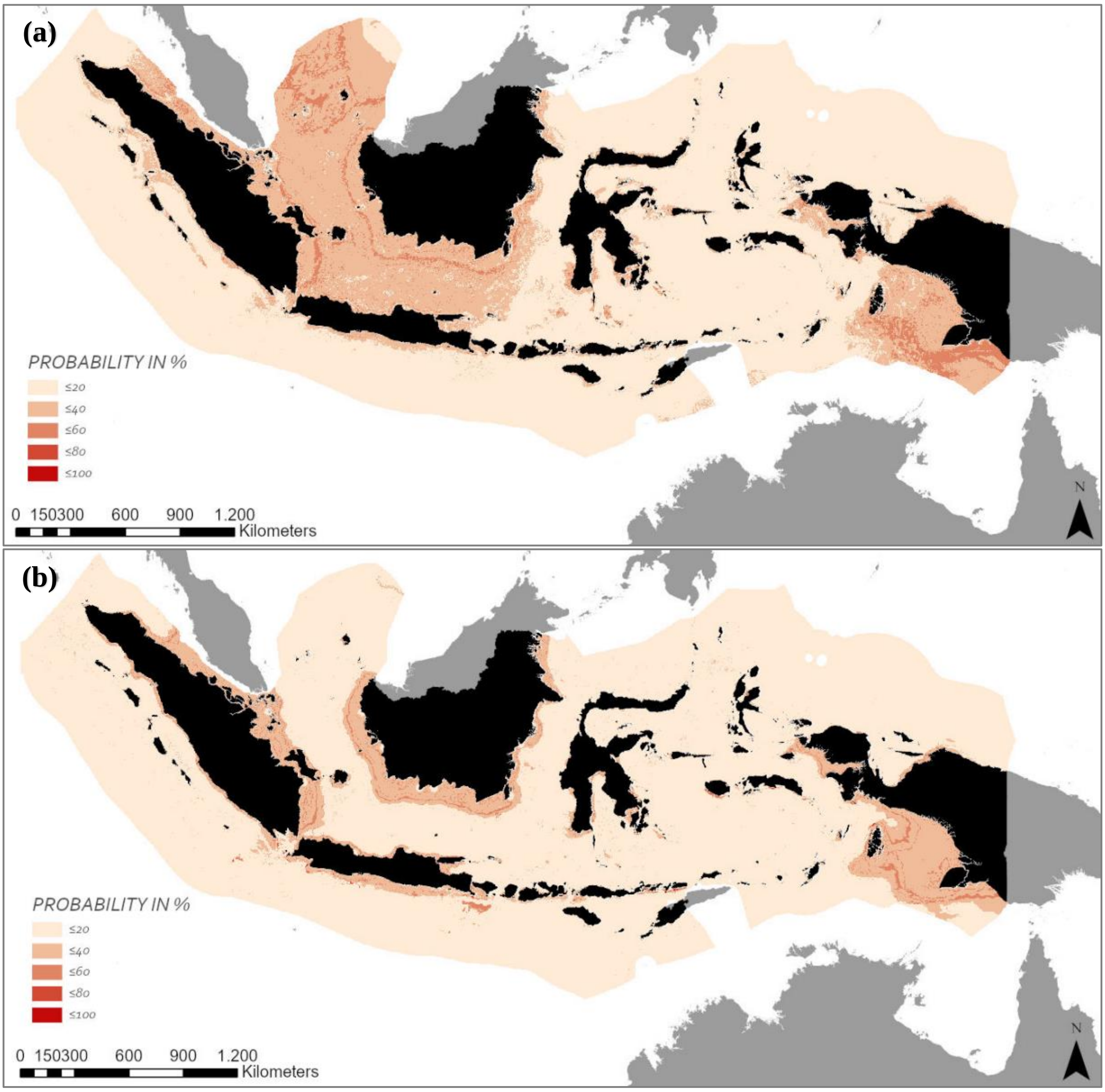

Figure 5. Projected models by extrapolating local models to Indonesian EEZ: common bottlenose dolphin (a) and sperm whale (b).

For sperm whales, the extrapolated distribution clearly indicates that higher probabilities are concentrated around coastal areas (Figure 5b). Occurrence probabilities of ≥ 40 % are visible in areas that are near to the islands, but not limited to very short distances to the islands. Occurrence probabilities that are higher than ≥ 40 % are located further away from the coast. However, the occurrences are still within the vicinity of islands as opposed to the open oceans. This finding is contradictory to the species local models which showed occurrence farther away from the coast, into the open waters. High probabilities of ≥ 60 % barely exist in the maps for sperm whale, with several of these probabilities spotted in East Sumatera, South and West Kalimantan, South Bali and South Papua (Figure 5b). The variability of the probabilities increases as the probability increases (Figure A3 in Appendices).

4. Discussion

This study successfully predicted the species distributions of two cetacean species, common bottlenose dolphin and sperm whale, in different smaller regions by local models based on predictive Random Forest modelling (local RF). These local RF models have been extrapolated into unsurveyed Indonesian EEZ. Overall, many variabilities exist in the predicted distributions between species and among study regions. Since the two species have different habitat preferences, an obvious difference in predicted distributions was expected. The local RF models were indeed able to predict differences between the distributions of the two species. For instance, the common bottlenose dolphin was predicted mainly in near coastal areas and in the waters between small islands, while the sperm whale was predicted to be found in open waters far away from the coast and barely found between small islands (Figure 4). These findings correspond to the habitat preferences of the species that can be found in the literature [5,46,47].

The extrapolated distributions of the two cetacean species from local models into unsurveyed Indonesian EEZ, however, did show some unexpected results (Figure 5). High probabilities of common bottlenose dolphin were located at some distance from coastal areas, and differed from local RF models that showed high probabilities around coastal areas. A surprising result also occurred for the sperm whale, as the extrapolated predictions were mainly located around the islands close to the coast and not in the open waters. It should be noted that these contradicting results may be caused by the sperm whale sighting data that mainly contains presences that are located close to the coast. Guisan & Thuiller [18] suggested that species occurrence data needs to be accurate when modelling species distribution as any location errors or poor representativeness of the survey data (as in our study) may impact model performance. Our results indicated that model transferability from local models to unsurveyed regions was not always preferable (i.e., result in low accuracy) as also found in previous studies. For instance, Redfern et al. [22] reported that their models predicting blue whale habitats were not transferable into unsurveyed areas.

The choices made in our study, including setting and complexity of the model, have influenced the output of the models. We first examined the relationship between sighting data and environmental predictors by creating local RF models to the extent of smaller study regions. These found relationships were then used to project the distributions over space to the extent of the Indonesian EEZ waters. This was done as no sighting data was available beyond the local study regions. Mannocci et al. [11] applied the same method and created GAM models for three local study regions before extrapolating distributions beyond these regions and creating worldwide predictions. The found relationships, were influenced by the size of local study regions (Figure 1) and extrapolation results differed when altering the size of these local study regions. This means that the results of our study are very sensitive to the four selected local study regions, and therefore caution should be taken when interpreting extrapolated species distributions.

It would be interesting to examine how sensitive the extrapolation results are to the chosen local study regions and whether skipping the extra step of first examining the relationships at local study regions and directly extrapolating into the entire area of interest generates better results. This direct extrapolation approach has been applied in other studies (e.g., Redfern et al. [22]), but had its own drawbacks. Model results of this approach were based on sighting data that was clustered to specific areas and large parts had not any presence data at all. It was possible that the species of subject reacted differently to its environment in different locations [18,53], and the applied extrapolation approach did not eliminate this drawback. However, it gave the opportunity to first examine the relationships at a local level and assessed whether they are similar or not, before extrapolating these relationships beyond surveyed regions. In our study, the RF algorithm was chosen for predicting species distributions. Although the RF algorithm was a suitable technique, many SDM studies used multiple algorithms to enable reviewing which algorithm leads to the best results [10]. This was not the scope of this study, but it may be of interest to compare different modelling techniques and examine which one performs best at predicting distributions beyond surveyed areas [23]. By doing this, it is also possible to examine how sensitive the data is to the modelling technique chosen and whether they lead to very distinct outcomes.

The RF model was, due to the Biomod2 framework, unable to use simultaneously the different region based SDM outcomes to project the species distribution on the full Indonesian EEZ extent. Biomod2 only enables the use of the SDM outcome of one single study region at the same time to project the species distribution. Consequently, the found relationships at the different study regions were not used at the same time to project into space, and averaging did not lead to the best results. Extrapolating the projections simultaneously ensures that new predictions are made based on the relationships found nearby and these relationships have often a higher reliability [61]. It would be much better if the outcomes of all study regions could have been used simultaneously to project the distributions to the full extent of the Indonesian EEZ. Mannocci et al. [11] have used their local results simultaneously during the extrapolation part of their study. For future work, such Biomod2 option is crucial.

The sighting data used in this study was sampled in a non-systematic way, which may have resulted in sighting data that did not represent the actual cetacean distributions accurately. Sighting data collected in a systematic way would have been used for modelling, for example, Mannocci et al. [11] used sighting data that was systematically collected. The sighting data also collected over 18 years that, intrinsically, such time series may resemble increasing human impact and climatic modifications. However, a long time-span is usual when studying species that are rare and hard to observe. For example, Redfern et al. [22] and Sahri et al. [25] used sighting data for modelling cetacean distributions that was collected more than 15 years with relatively satisfied results.

The models used in this study could also be enhanced if absence data would be available. With absence data available, areas can be assigned as 100% true absences and give more insights in areas that are not suitable for the species of interest [21,53,61]. Since no absence data was available, pseudo-absences were generated randomly with an amount of 100 times of the presences. The effect of using a different amount of pseudo-absences in extrapolated models to unsurveyed regions has not been studied, but might be interesting to examine in future studies. Our study has not taken into account the temporal dimension of the presences and absences. To further increase the knowledge on the distributions of the two cetacean species, a temporal verification should be done taking into account the temporal behavior of the two species [5]. Most sighting data has information on the date of the sightings and this can be used to explore temporal or seasonal differences in species distributions in future research.

Cetaceans are species that are hard to study because they are highly mobile animals and mostly being underwater [62,63]. Most information on cetacean occurrences elsewhere, including in Indonesia come from the animals being at the water surface or from deceased bodies that have washed ashore. This study contributes to gaining a better understanding of cetacean distributions and habitat preferences from species distribution modelling. The models revealed their occurrence probabilities and the importance of environmental variables. The environmental predictors in this study were selected to represent the habitat preferences of the cetacean species. The choice of the specific variables has a large impact on the results of this study. It may be possible that environmental predictors that are not incorporated in this study are better at representing the habitat preferences of the cetacean species. Mannocci et al. [11] for example chose 14 different environmental variables, with some of the variables they used have not been used in this study, such as wind speed, silicate-nitrate ratio and net primary production. Dransfield et al. [40] even chose 20 environmental variables as candidates for modelling. To choose the best environmental predictors a complete understanding of the underlying mechanisms of the distribution of the cetaceans is needed. The literature review of this study on the habitat preferences of cetaceans represented what is known about the preferences of the cetacean species at this moment of time, but this information still knows uncertainties [5]. One type of environmental predictor that would improve the model is a dataset representing the distribution of the cetaceans' prey [38].

Predicting cetacean habitats beyond the sampled areas was unpreferable in this study. This result was influenced by the lack of empirical support from the data that was used and lead to modelling results that do not fit biological reality [53]. Deceptive extrapolation results can be caused by inaccurate sampling methods or the usage of unsuitable modelling methods or variables. It is important to understand the underlying mechanisms of a species' distribution in order to include useful environmental variables in the model. Sometimes ecological understanding of the species may be available, but the required environmental data is lacking [53,64]. Inaccurate sampling schemes can result in the failure to correctly capture the relevant environmental variables and can result into unpreferable extrapolation outputs. Besides, the assumption often exists that species are at equilibrium with their environment, which means that the species inhabit all suitable locations available and that the sighting data perfectly represents the species' actual habitat [18,53]. The most suitable habitats may however stay unoccupied, due to disturbances, migratory behavior, or other factors. Bouchet et al. [53] mentioned an example that West Australian bottlenose dolphins stay in areas that contain less prey, but are safer during times of high shark abundances. It may thus be possible, when extrapolating distributions into unsurveyed areas that favorable areas are identified but none of the species actually inhabit these areas. It is also possible, when extrapolating on broader scale, that locations are included where the species react differently to the habitat.

In contrast to using more variables, Mannocci et al. [11] suggested that extrapolation models should not be too complex and only use a limited number of covariates. These simple models are especially preferred when aiming to increase the ecological realism and to interpret the model more easily. More complex extrapolation models that include a great number of covariates often create more extensive results, and the risk for overfitting is also large for these models. Bouchet et al. [53] suggested that complex overfitting models suit for pinpointing areas for reintroducing rare species, while simple models are more suitable for identifying habitat locations of rare species. No species distribution model will thus perform perfectly, but choices should be made based on the data availability and quality, as well as the purpose of the modelling.

The best way to evaluate the predictions of a model was to validate the outcomes externally by using a separate independent dataset of sightings [11]. A dataset for this external validation, however, was not available for this study. An online habitat model, Aquamaps (www.aquamaps.org), provides predictions of the distributions of many marine species, including cetaceans. These predictions are based on a species distribution model developed by Kaschner et al. [65] with different input of occurrence data and environmental predictor variables. These online maps thus can be used for comparisons of our modelling results. The AquaMaps results of the common bottlenose dolphin are a bit more similar to the results of our study. The areas with a high predicted occurrence probability in our extrapolation maps are also predicted to be high probability areas in the AquaMaps predictions. In contrary for the sperm whale, the predictions of our study and of AquaMaps do totally differ. The areas that were predicted to have high sperm whale occurrence probabilities in our study have a low occurrence probability in AquaMaps. AquaMaps also predicts high occurrence probabilities only in areas close to the open waters and not in between islands. These AquaMaps results comply more with the known habitat preferences of sperm whales as these animals are known to prefer areas that are away from the coastal areas [5,46].

Despite the limitation of transferability, it can be inevitable when dealing with (endangered) species living in remote and poorly accessible areas. For instance, this study was still able to identify potential habitats in regions without sighting data (Figure 5), and the extrapolation of cetacean distributions are essential for conservation purposes [22,53]. Most important is that the assumptions of extrapolation are understood to correctly interpret results and embed them in conservation work. SDM applications to extrapolate species distributions into unsurveyed areas has received less attention, but become apparent due to the call for models that predict beyond the studied regions to address large-scale conservation challenges, such as cetacean distributions in unsurveyed Indonesian EEZ waters.

Acknowledgements
The sighting data used in our study was obtained from Sahri et al. [25] that previously acquired through data sharing agreement and is therefore not publicly accessible. We would like to thank once again the organizations mentioned in that reference: The Nature Conservancy Indonesia, WWF Indonesia, Wakatobi National Park Authority, Yayasan Konservasi RASI (Rare Aquatic Species of Indonesia), Cetacean Sirenian Indonesia (CETASI), and Marine Megafauna Research Group - Misool Foundation, Savu Sea Program Indonesia. We also thank the GIMA joint MSc program of Technical University Delft, Twente University, Utrecht University and Wageningen University of the Netherlands to support this research topic.

Appendices

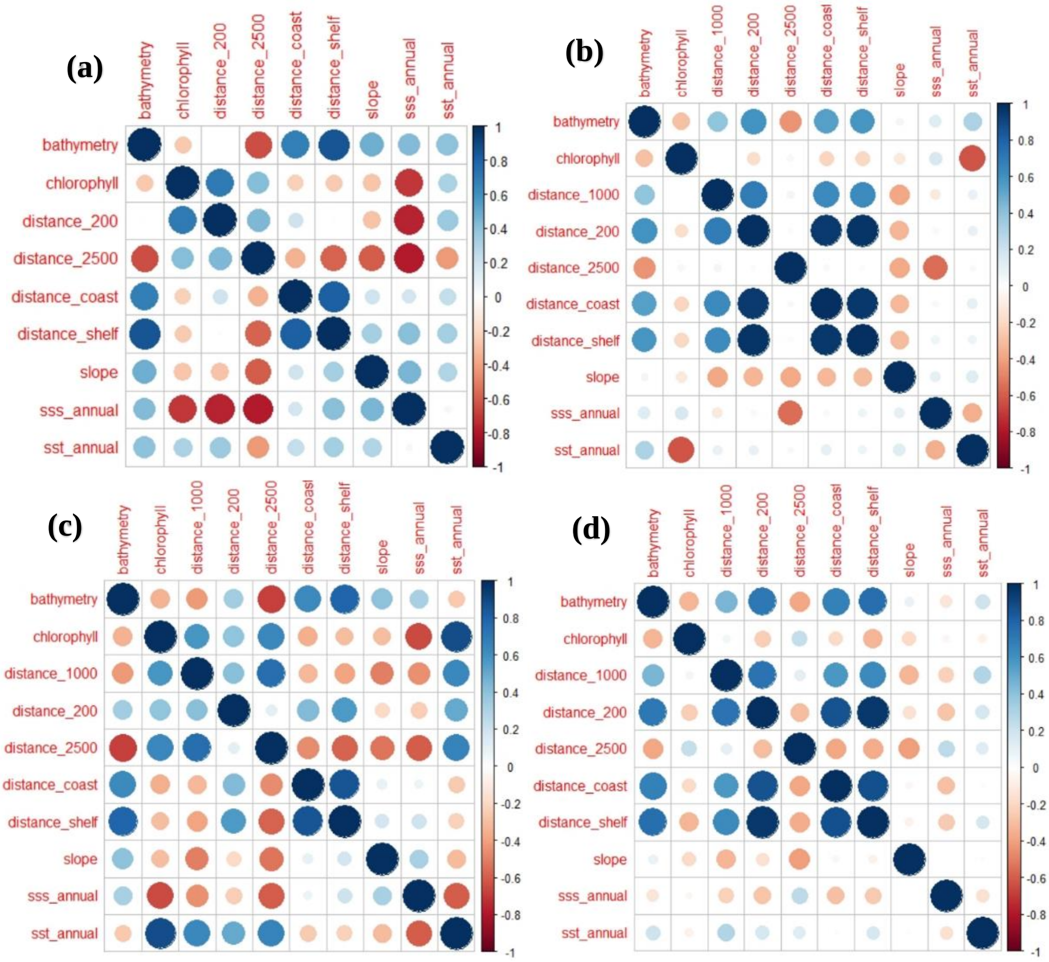

Figure A1. Correlation heatmaps of ten potential environmental predictor variables used for SDM in Bird's Head (a), Lesser Sunda (b), NE Borneo (c), and SE Sulawesi (d). Saturated colors and large dots show a high correlation. In Bird's Head, high correlations occur in: distance_shelf and bathymetry, distance_shelf and distance_coast, SSS and distance_200, as well as SSS and distance_2500. It was thus decided to remove distance_2500, as this isobath barely occurs in Bird's Head region. Distance_shelf was also excluded from the modelling process, as bathymetric is known to be very important in predicting cetacean distributions [11]. In Lesser Sunda, high correlations were found in: distance_1000 and Chl-a, distance_coast and distance_200, distance_shelf and distance_200, distance_shelf and distance_coast, SST and distance_1000, as well as SST and distance_200. It was decided to exclude distance_shelf as distance_coast is very important for cetacean habitat modelling. Distance_200 was also

excluded but as this isobath was close to the coast, this layer was represented by distance_coast. Distance_1000 was excluded as Chl-a is a predictor variable that has shown to be of great importance in cetacean modelling [11]. SST was excluded as two other climatic variables (Chl-a and SSS) already chosen. In NE Borneo, high correlations occur in: distance_2500 and distance_1000, distance_shelf and bathymetry, distance_shelf and distance_coast, as well as SST and Chl-a. Distance_2500, distance_1000, distance_shelf and SST were excluded for the same reason as they were excluded in Bird's Head and Lesser Sunda. In SE Sulawesi, some correlations appear to be high: distance_coast and distance_200, distance_shelf and bathymetry, distance_shelf and distance_200, as well as distance_shelf and distance_coast. Distance_200 and distance_shelf were excluded for the same reasons as for other regions. With this arrangements, five out of ten variables eventually used for SDMs: bathymetry, chlorophyll-a, distance to coast, slope, and sea surface salinity (SSS).

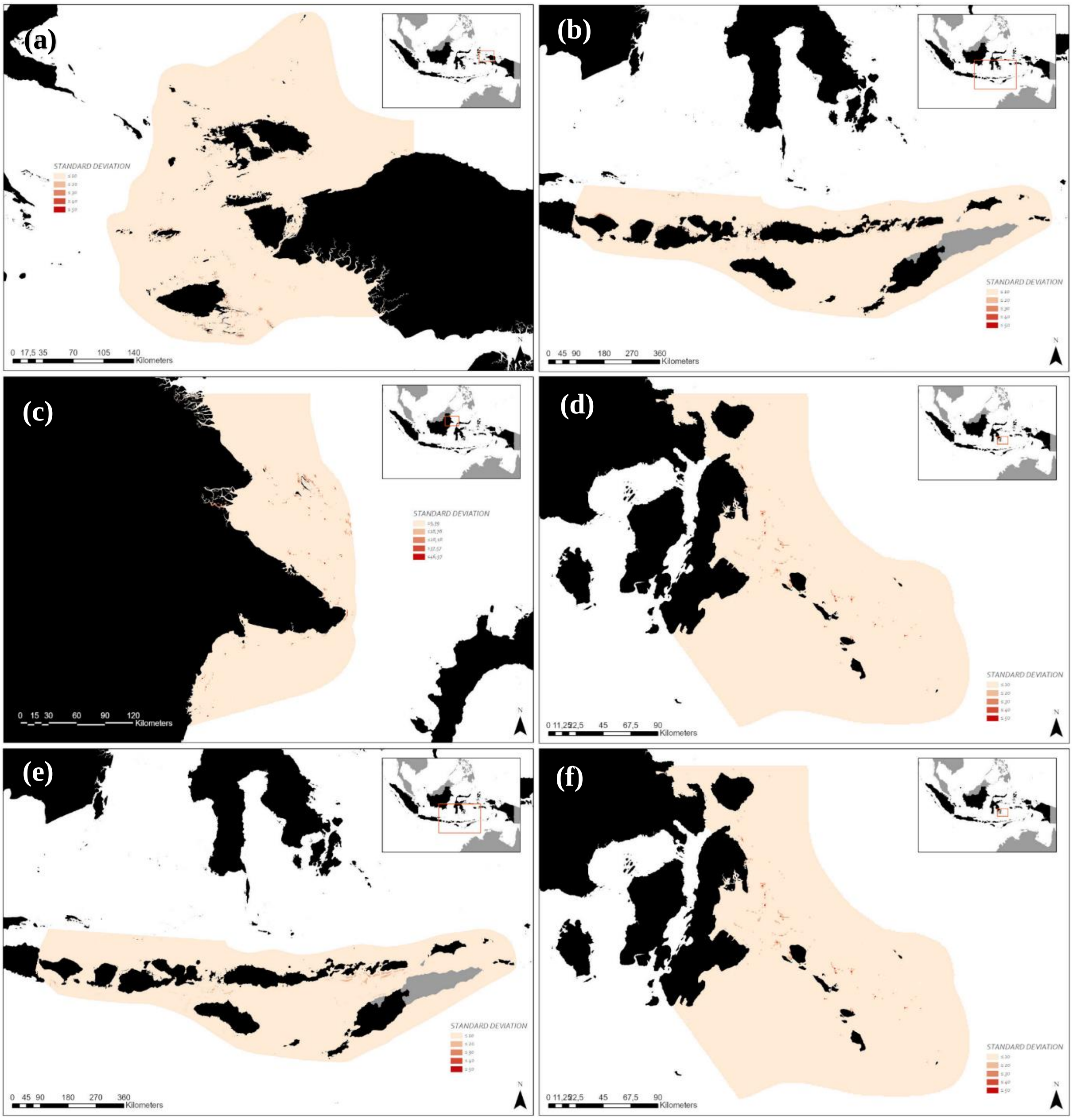

Figure A2. Variability of Random Forest local models per region for common bottlenose dolphin (a-d) and sperm whale (e-f). A light-red coloured cell depicts low variability between the 10 runs and the darker the red colour of the cell the higher the variability.

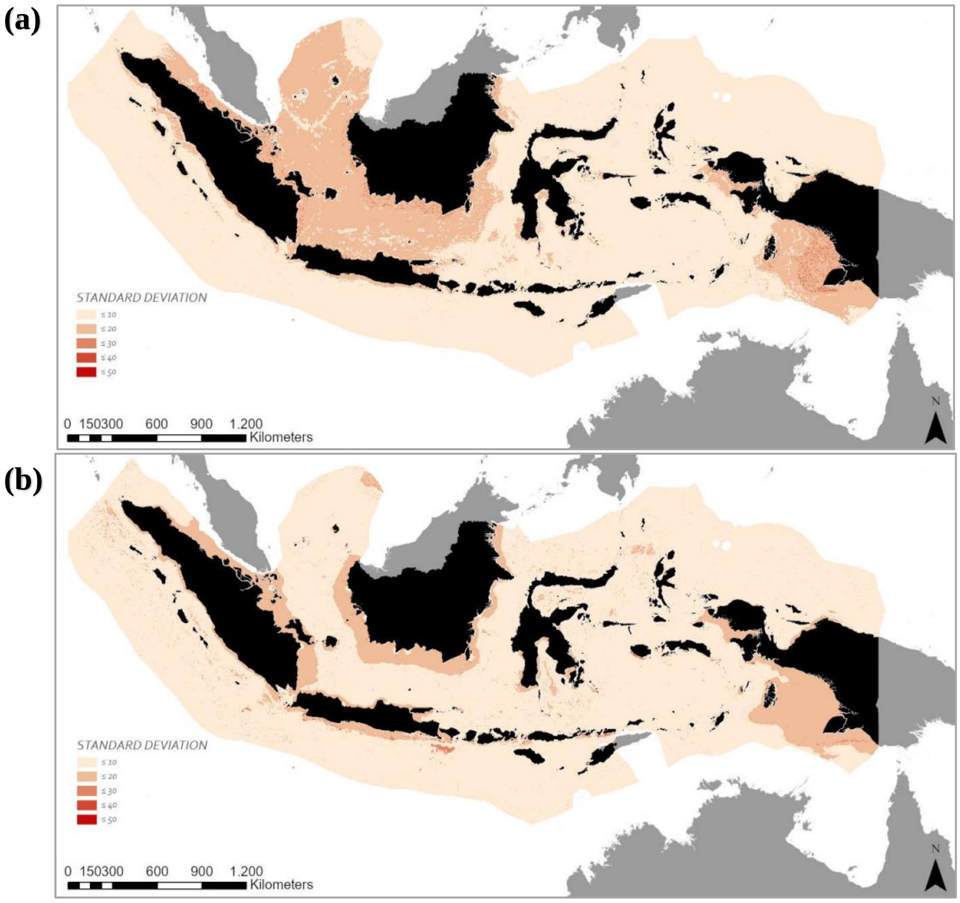

Figure A3. Variability of extrapolated models into unsurveyed Indonesian EEZ for common bottlenose dolphin (a) and sperm whale (b). A light-red coloured cell depicts low variability between the 10 runs and the darker-red the colour of the cell the higher the variability.

References

[1] DeMaster D P, Fowler C W, Perry S L and Richlen M F 2001 *J Mammal* **82** 641–51
[2] Harwood J 2001 *J Mammal* **82** 630–40
[3] Hammond P S et al. 2013 *Biol Conserv* **164** 107–22
[4] https://www.iucnredlist.org/
[5] Carwardine M 2020 *Handbook of Whales, Dolphins and Porpoises* (London: Bloomsbury Publishing Plc) p 528
[6] Kaschner K, Tittensor D P, Ready J, Gerrodette T and Worm B 2011 *PLoS One* **6** e19653
[7] Sahri A, Mustika P L K, Dewanto H Y and Murk A J 2020 *Mar Policy* **117** 103893
[8] https://doi.org/10.1007/3-540-33291-X_3
[9] di Sciara G N and Birkun Jr A 2010 *Conserving whales, dolphins and porpoises in the Mediterranean and Black Seas* (Monaco: ACCOBAMS) p 212
[10] Elith J et al. 2006 *Ecography* **29** 129–51
[11] Mannocci L, Monestiez P, Spitz J and Ridoux V 2015 *J Biogeogr* **42** 1267–80
[12] Alessi J and Fiori C. 2014 *J Coast Conserv* **18** 449–458
[13] Kaschner K, Quick N J, Jewell R, Williams R and Harris C M 2012 *PLoS One* **7** e44075
[14] Braulik G T, Kasuga M, Wittich A, Kiszka J J, MacCaulay J, Gillespie D, Gordon J, Said S and Hammond P S 2018 *Aquat Conserv* **28** 216–30
[15] Davidson A D, Boyer A G, Kim H, Pompa-Mansilla S, Hamilton M J, Costa D P, Ceballos G and Brown J H 2012 *Proc Natl Acad Sci* **109** 3395–400
[16] Sahri A, Mustika P L K, Purwanto P, Murk A J and Scheidat M 2020 *Front Mar Sci* **7** 569936
[17] Marmion M, Parviainen M, Luoto M, Heikkinen R K and Thuiller W 2009 *Divers Distrib* **15** 59–69

[18] Guisan A and Thuiller W 2005 *Ecol Lett* **8** 993–1009
[19] Guisan A, Edwards T C and Hastie T 2002 *Ecol Modell* **157** 89–100
[20] Hengl T, Sierdsema H, Radović A and Dilo A 2009 *Ecol Modell* **220** 3499–511
[21] Pearson R G 2007 *Lessons in Conservation* **3** 54–89
[22] Redfern J V, Moore T J, Fiedler P C, de Vos A, Brownell R L, Forney K A, Becker E and Ballance L 2017 *Divers Distrib* **23** 394–408
[23] Heikkinen R K, Marmion M and Luoto M 2012 *Ecography* **35** 276–88
[24] Green A L and Mous P J 2008 *Delineating the Coral Triangle, its ecoregions and functional seascapes* (Brisbane: TNC Coral Triangle Program) p 44
[25] Sahri A, Putra M I H, Mustika P L K, Kreb D and Murk A J 2021 *Ocean Coast Manag* **205** 105555
[26] Jiménez-Valverde A 2020 *Ecol Indic* **114** 106289
[27] Miller J A 2012 *Progress in Physical Geography: Earth and Environment* **36** 681–92
[28] Brown J L 2014 *Methods Ecol Evol* **5** 694–700
[29] Brown J L, Bennett J R and French C M 2017 *PeerJ* **5** e4095
[30] VanDerWal J, Shoo L P, Graham C and Williams S E 2009 *Ecol Modell* **220** 589–94
[31] Graham C H, Ferrier S, Huettman F, Moritz C and Peterson A T 2004 Trends *Ecol Evol* **19** 497–503
[32] Zaniewski A E, Lehmann A and Overton J M 2002 *Ecol Modell* **157** 261–80
[33] Barbet-Massin M, Jiguet F, Albert C H and Thuiller W 2012 *Methods Ecol Evol* **3** 327–38
[34] Lobo J M and Tognelli M F 2011 *J Nat Conserv* **19** 1–7
[35] Fiedler P C et al. 2018 *Front Mar Sci* **5** 419
[36] Kanaji Y, Okazaki M, Kishiro T and Miyashita T 2015 *Fish Oceanogr* **24** 14–25
[37] Praca E, Gannier A, Das K and Laran S 2009 *Deep Sea Res 1 Oceanogr Res Pap* **56** 648–57
[38] Gomez C, Lawson J, Kouwenberg A L, Moors-Murphy H, Buren A, Fuentes-Yaco C, Marotte E, Wiersma Y and Wimmer T 2017 *Endanger Species Res* **32** 437–58
[39] Harris P T, Macmillan-Lawler M, Rupp J and Baker E K 2014 *Mar Geol* **352** 4–24
[40] Dransfield A, Hines E, McGowan J, Holzman B, Nur N, Elliott M, Howar J and Jahncke J 2014 *Endanger Species Res* **26** 39–57
[41] Alin A 2010 *WIREs Computational Statistics* **2** 370–4
[42] Mukaka M M 2012 *Malawi Med J* **24** 69–71
[43] Mannocci L, Roberts J J, Miller D L and Halpin P N 2017 *Conservation Biology* **31** 601–14
[44] Authier M, Saraux C and Péron C 2017 *Ecography* **40** 549–60
[45] Bearzi G, Agazzi S, Bonizzoni S, Costa M and Azzellino A 2008 *Aquat Conserv* **18** 130–46
[46] Whitehead H 2018 *Sperm Whale: Physeter macrocephalus* (Amsterdam: Elsevier) p 919–25
[47] Wells R S and Scott M D 2009 *Common Bottlenose Dolphin* (Amsterdam: Elsevier) p 249–55
[48] Friedlaender A S, Halpin P N, Qian S S, Lawson G L, Wiebe P H, Thiele D and Read A 2006 *Mar Ecol Prog Ser* **317** 297–310
[49] Marini C, Fossa F, Paoli C, Bellingeri M, Gnone G and Vassallo P 2015 *J Environ Manage* **150** 9–20
[50] Grömping U 2009 *Am Stat* **63** 308–19
[51] Breiman L 2001 *Mach Learn* **45** 5–32
[52] https://cran.r-project.org/package=biomod2
[53] http://dx.doi.org/10.13140/RG.2.2.13774.41289
[54] Pearce J and Ferrier S 2000 *Ecol Modell* **133** 225–45
[55] Jiménez-Valverde A 2012 *Glob Ecol Biogeogr* **21** 498–507
[56] Liu C, White M D and Newell G 2009 *Measuring the accuracy of species distribution models: a review* (Cairns: 18th World IMACS/MODSIM Congress) pp 4241–47
[57] Allouche O, Tsoar A and Kadmon R 2006 *Journal of Applied Ecology* **43** 1223–32
[58] Jiménez-Valverde A 2012 *Glob Ecol Biogeogr* **21** 498–507
[59] Liu C, White M and Newell G 2011 *Ecography* **34** 232–43
[60] Engler R et al. 2011 *Glob Chang Biol* **17** 2330–41
[61] Elith J, Ferrier S, Huettmann F and Leathwick J 2005 *Ecol Modell* **186** 280–9
[62] Horton T W et al. 2017 *Front Mar Sci* **4** 422
[63] Sahri A, Jak C, Putra M I H, Murk A J, Andrews-Goff V, Double M C and van Lammeren R J 2022 *Biol Conserv* **272** 109594
[64] Peters D P C and Herrick J E 2004 *Oikos* **106** 627–36
[65] Kaschner K, Watson R, Trites A W and Pauly D 2006 *Mar Ecol Prog Ser* **316** 285–310