Using machine learning models to predict the distribution of a cryptic marine species: the sperm whale

Philippine Chambault[1], Sabrina Fossette[2], Mads Peter Heide-Jørgensen[3], Daniel Jouannet[4], and Michel Vély[5]

[1]Greenland Institute of Natural Resources Climate Research Centre
[2]Biodiversity and Conservation Science, Department of Biodiversity
[3]Greenland Institute of Natural Resources
[4]EXAGONE réseauTERIA
[5]Megaptera

August 4, 2020

Abstract

Implementation of effective conservation planning relies on a robust understanding of the spatio-temporal distribution of the target species. In the marine realm, this is even more challenging for cryptic species with extreme diving behaviour like the sperm whales. Our study aims at investigating the movements and predicting suitable habitat maps for this species in the Mascarene Archipelago in the South-West Indian Ocean. Using 21 satellite tracks of sperm whale and 8 environmental predictors, 14 supervised machine learning algorithms were tested and compared to predict the whales' distribution during the wet and dry season, separately. Fourteen of the whales remained in close proximity to Mauritius while a migratory pattern was evidenced with a synchronized departure for 8 females that headed towards Rodrigues Island. The best performing algorithm was the random forest, showing a strong affinity for Sea Surface Height during the wet season and for bottom temperature during the dry season. A more dispersed distribution was predicted during the wet season whereas a more restricted distribution to Mauritius and Reunion waters was found during the dry season. The results of our study fill a knowledge gap regarding seasonal movements and habitat affinities of this vulnerable species, for which IUCN regional assessments are still lacking in the Indian Ocean. Our findings also confirm the great potential of machine learning algorithms in conservation planning and provide concrete tools to support dynamic ocean management.

1. INTRODUCTION

Implementation of effective conservation planning relies on a robust understanding of the spatio-temporal distribution of the target species. In the marine realm, this is even more challenging for cryptic species that are rarely seen at the sea surface due to their extreme diving behaviour such as beaked whales or sperm whales (Perrin et al. 2009). Among these deep diving predators, the sperm whale (*Physeter macrocephalus*) that can display long (˜45 min) and deep dives (up to 1860 m) with short surface intervals (˜9 min) (Watwood et al. 2006, Teloni et al. 2008), is listed as 'vulnerable' on the IUCN classification redlist. Depletion of this species' global population is the result of excessive historic hunting and the lack of complete recovery of the population worldwide (Whitehead 2002). Although numerous studies have focused on sperm whales' spatial ecology and habitat selection (Jaquet 1996, Watkins et al. 1999, Gannier et al. 2002, Whitehead & Rendell 2004, Gannier & Praca 2007, Pirotta et al. 2011, 2020), regional assessments are still limited to the North-East coast of Europe and the Mediterranean Sea (Gannier et al. 2002, Laran & Drouot-Dulau 2007,

Laran et al. 2017b, Taylor et al. 2019, Virgili et al. 2019) despite the presence of sperm whales in the Pacific (Davis et al. 2007, Whitehead et al. 2008) and Indian Oceans (Laran et al. 2017a, Huijser et al. 2020).

Since the establishment of the Indian Ocean Whale Sanctuary by the International Whaling Commission in 1979 (Holt 1983), an increasing number of surveys focussing on the distribution of cetaceans (including sperm whales) in this region have been conducted (Mannocci et al. 2014b, Laran et al. 2017a). Recent aerial surveys conducted in the South-West Indian Ocean confirmed the presence of sperm whales around Reunion and Mauritius Islands (Mannocci et al. 2014b, Lambert et al. 2014, Laran et al. 2017a), but in surprisingly low densities. Low densities may be the result of biased predictions generated by a large amount of false absences (Virgili et al. 2017) due to deep divers like sperm whales spending a small amount of time at the sea surface, i.e. 16-21% (Jaquet et al. 2000, Hooker & Gerber 2004). In addition, although aerial surveys have significantly improved our understanding of the habitat use of marine megafauna in this region, this methodology can only provide a static picture of a species distribution unless surveys are regularly repeated throughout the year which is unlikely due to the cost of field campaigns and logistical difficulties (e.g. bad weather conditions). Satellite telemetry by tracking animals individually provides another way to assess deep divers' movement patterns, and fine-scale habitat affinities through generating animal's trajectories in space and time, but is often biased towards small sample sizes.

Species Distribution Models (SDMs) have been largely used to predict potentially suitable habitats of marine species based on the relationships between the animal's occurrences and its environment (Austin 2002, Elith & Leathwick 2009). In conservation spatial planning, the potential distribution of a species is a powerful information tool to delineate protected areas in a more efficient way. However, SDMs are usually based on well-established regression methods (e.g. GLM, GAM), incorporating only a limited number of algorithms for model comparison. In contrast to these classical methods, machine learning offers a wide range of algorithms to address ecological questions and provide robust and accurate predictions, being therefore a promising tool in species distribution modelling and conservation planning (Elith et al. 2006).

Using data from the first satellite tags (n=21) deployed on both male and female sperm whales inhabiting Mauritius waters (south-west Indian Ocean, Fig. 1), the predicted distribution of this cryptic species was modelled using a series of machine learning algorithms. By combining the individual satellite tracks with eight oceanographic variables (physical, surface and in-depth predictors), our study aims at (i) investigating this species' seasonal pattern and residency behaviour in the Indian Ocean, (ii) accurately predicting its distribution and (iii) assessing the diel pattern in its diving behaviour. Since deep divers such as sperm and beaked whales might show a weak dependence on surface oceanographic characteristics (Mannocci et al. 2014b), we also included two new covariates describing the vertical characteristics of the water column, i.e. the mixed layer depth and the bottom temperature. By combining machine learning, cutting-edge oceanographic variables and the first tracking dataset around Mauritius, our results provide a first robust baseline needed to assess the spatio-temporal distribution of this vulnerable species in a poorly known region: the South-West Indian Ocean (SWIO).

2. METHODS

2.1 Study area and tag deployment

Field work was conducted in the southwestern part of Mauritius Island in 2014, 2016 and 2018 (Fig. 1). Sperm whales (n=22) were instrumented with Wildlife Computers SPOT5, SPOT6 and SPLASH10 satellite transmitters (*http://wildlifecomputers.com*) and modified for deployment and use on whales by Mikkel Villum Jensen (http://mikkelvillum.com). The tags were deployed using the ARTS, a modified pneumatic air gun, at about 8 to 10 m from the whale set at pressure of 11 bars (Heide-Jørgensen et al. 2001). This is a standard procedure commonly used in tracking projects of large whales (Andrews et al. 2019). Both transmitters consisted of a stainless-steel cylinder (SPOT5: 22x110 mm SPLASH10 24 mm x 155 mm) that contained the electronics and one lithium AA cell. A 38mm stopplate mounted 3 cm from the rear end of the tag stopped

the tag at the surface of the skin and prevented the tag from penetrating deeper into the blubber/muscle layer. The rear end of the steel tube had an antenna (160 mm length) and a salt water switch that ensured that transmissions were only conducted when the rear part of the tag was out of the water. A pressure transducer was positioned just below the stop plate on SPLASH10 tags. In the front, the tags were equipped with a stainless-steel anchor spear with a sharp pointed triangular tip and foldable barbs (40–50mm) to impede expulsion from the blubber-muscle layer. The total length of the SPOT5 and SPOT6 from the stop plate to the tip of the anchor was 170 mm and the mass of the instrument with attachment spear was 133 g. The total length of the SPLASH10 tag was 215 mm and the mass of the instrument with attachment spear was 250 g.

The SPLASH10 tags collected summarized dive data in bins where dives to different depths and time spent at the same depths were binned into 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, >1300 m. The duration of dives was summarized in these bins: 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65 and >65 mins. In addition to that the maximum depth of dives was recorded for each 24 hr.

The tags were programmed to make a maximum of 250 transmissions per day between 04:00 and 16:00. The SPOT5 and SPOT6 tags were allowed to transmit every day in November through January and every other day the rest of the year. The SPLASH10 tags were allowed to collect dive data and transmit every day.

The tagging operation in Mauritius was conducted from a rigid hull inflatable boat (24 ft) with a 2 x 90 hp outboard motor, a steering panel, and a maximum speed of 24 knots. The boat was equipped with a crow nest to secure the tagger and provide a stable platform when approaching and tagging the whales. The satellite tags were deployed into the left or right flank of the whales about 1-2 m ahead of the dorsal fin and within 2 m from the midline of the whale's body. Approximate length of the tagged whales was estimated by comparing the size of the whale with the length of the boats involved with the tagging. Based on dimorphic morphology and on the Mauritius Marine Conservation Organization (M2CO) photo ID catalogue, mature males and mature females were also distinguished (Sarano & Sarano 2017).

2.2 Location data processing

Location and dive were obtained through the Argos Data Collection and Location System using the Kalman filter which greatly improves the location data (Lopez et al. 2014). Dive data were decoded in Wildlife Computers portal. All statistical analyses were performed using R software version 4.0.0. We restricted our dataset to positions associated with a travel speed lower than 7 $km.h^{-1}$(Wahlberg 2002). Locations on land were also discarded. Any individuals containing less than 10 locations (over both seasons together) were also discarded from the analysis. In order to assess seasonal patterns and monsoon periods of the Indian Ocean, seasons were discriminated as follows: dry season from April to November and wet season from December to March.

2.3 Kernel density estimation

To investigate the residency pattern of the sperm whales and locate their high-use areas, a kernel utilisation density approach was used for both seasons separately (Worton 1989). Following the method developed in Chambault et al. (2020), the reference bandwidth parameterh_{ref} was first calculated for each season. Then,h_{ref} was sequentially reduced in 0.10 increment (0.9 h_{ref} , 0.8 h_{ref} , 0.7h_{ref} , ...) until 0.1h_{ref} , and the most appropriate smoothing parameter was chosen visually by comparing the kernel density to the original location data (Kie 2013). The core and global home ranges were calculated from the 50 and 90% kernel contours respectively for each season.

2.4 Environmental data

Strong relationships exist between cetaceans distribution and dynamic environmental variables (Mannocci et al. 2014a, b), such as *sea surface temperature* (SST), *sea surface height* (SSH), *ocean currents* (U and V components) and *ocean current velocity*. These variables were therefore tested as potential drivers of sperm whales' movements and to predict their potential distribution in the SWIO. In addition to surface variables, the *mixed layer depth*(MLD) was also considered as this variable is known to be closely related to primary

productivity. However, the deep diving behaviour of sperm whales might also be influenced by temperatures at the bottom of the water column where they mainly forage. Consequently, *bottom temperature* was also considered a likely driver of sperm whales' movements. Bathymetry was also extracted from GEBCO at a spatial resolution of 1 km and the slope was subsequently derived from the bathymetry and expressed in degrees to get a proxy of the seafloor roughness. The dynamic variables were extracted monthly from the products *Global Ocean Physics Reanalysis Glorys S2V4* (PHYS 001-024) and the *Global Ocean Physics Reanalysis Glorys12v1* (PHY-001-030) at a resolution of 0.08° (from E.U. Copernicus Marine Service Information). All variables were then set to the same spatial resolution of 0.08 decimal degree. Monthly grids of each predictor were then averaged for each season: between December and March for the wet season, and between April and November for the dry season.

2.5 Species distribution modelling

In order to identify the environmental drivers of sperm whales' movements and predict their potential distribution, we built a series of species distribution models (SDMs) using multiple algorithms from the *caret* package in R. We first used an environmental background based technique to generate pseudo-absences (Senay et al. 2013, Iturbide et al. 2015, Hattab et al. 2017, Schickele et al. 2020), relying on the assumption that true absences are more likely located in areas that are environmentally dissimilar from presence locations. Following the same procedure as described in Chambault et al (under review), a principal component analysis (PCA) was used to generate a two-dimensional environmental background representing the ordination results of the seven environmental variables available over the study area. One PCA was performed for each season separately. Pseudo-absences were then randomly generated outside environmentally favorable areas for each season and in equal number to the filtered occurrences (e.g. tracking locations). To increase the robustness of the results and assess their sensitivity to the pseudo-absences generation procedure, 10 different sets of pseudo-absences were simulated (i.e. 10 runs for each season). The eight environmental variables were then extracted at each occurrence and pseudo-absence.

In order to find the most adequate model to predict the distribution of sperm whales with the highest accuracy, we tested 14 different algorithms belonging to the following categories:

1. *Ensemble* : Random Forest (RF) and Stochastic Gradient Boosting (GBM);
2. *Regression* : Generalized Additive Model (GAM) and Multivariate Adaptive Regression Splines (MARS);
3. *Bayesian* : Naïve-Bayes (NB) and Bayesian Additive Regression Trees (BayesGLM);
4. *Decision tree* : Logistic Model Trees (LMT) and C5.0;
5. *Instance-based* : K-Nearest Neighbour (KNN) and KKNN;
6. *Dimensionality reduction* : Linear Discriminate Analysis (LDA) and Quadratic Discriminant Analysis (QDA);
7. *Support Vector Machine (SVM)* : SVM with radial kernel (SVMradial) and SVM with linear kernel (SVMlinear).

The 14 algorithms were ran for each simulation run using the presence of sperm whales (1: presence vs. 0: pseudo-absence) as a response variable. The 14 models included the seven predictors mentioned above. All predictors were scaled between 0 and 1, and collinearity was checked using the Variance Inflation Factor (below four). The dataset of each run was first randomly split between the training dataset (80% of the data) and the validation dataset (20% of the data). Each algorithm was run on the training dataset while model evaluation was performed on the validation dataset. Model comparison was based on a 10-fold cross-validation with three repetitions using the following performance metrics calculated for each run on the 20% validation dataset: the accuracy, the Kappa, the sensitivity, the specificity, the True Skill Statistics (TSS) and the F1 score. The best selected model was then tuned by testing several values of the *mtry* argument (the number randomly selected predictors). The "tuned model" was then used to generate ten prediction maps of the sperm whale's distribution (for each of the ten runs) and for each season separately. In parallel, the *caretEnsemble* package was used to generate ten predictions based on the combinations of the 14 algorithms previously tested, hereafter called the "stacking method". The ten prediction maps of each

approach were finally averaged to provide a final map of the potential distribution of sperm whales during the wet and dry season separately. The coefficients of variation were also calculated to provide a map of uncertainty (ratio of the standard deviation over the mean).

3. RESULTS

3.1 General tracking data

Across the 22 individuals equipped in Mauritius waters, three were males (#3963a, #20166b and #50681b) and all remaining whales were females. The number of locations recorded per sperm whale ranged from 7 to 176 (#50678 vs. #93106, respectively) – Table 1. The tracking duration was on average 34±20 days (range: 4-109 days). The total distance travelled varied between 154 (#6337a) and 3112 km (#24642), and the average travel speed was 2.7±0.3 $km.h^{-1}$.

3.2 Seasonal movements

A seasonal pattern in the movements of the whales was observed between the wet and dry seasons. Sixteen individuals were observed during the wet season from December to March (Fig. 2a) and the tracks of 14 whales were available during the dry season from April to November (Fig. 2b). The male tracked in 2018 (#50681b) was the only one migrating southward (Fig. 2a). The kernel densities showed a core of activity in shallow waters near Mauritius for both seasons. During the wet season, the core home range (50% kernel contour) was located south-west of Mauritius (Fig. 2c). During the dry season, the resident whales exhibited strong site fidelity by inhabiting shallow waters along the west coast of the island (Fig. 2d). The core home ranges were also located on variable topography for both seasons, e.g. high slopes (Fig. 2e, f).

3.3 Migratory patterns

Among the 16 individuals tracked during the wet season, eight whales (#20158a, #24642, #26712, #27261, #50678b, #50682b, #7618 and #7926) left Mauritius waters in December heading towards Rodrigues Island (Figs. 2a and 3). Except one, these individuals were all mature females (based on Photo ID, morphology, size and mother-calf association) and showed a synchronized departure from Mauritius. They made a loop eastward of Mauritius before returning at different times. In 2014, only one individual was considered migrant (departure on the 2014-12-25) while in 2018, seven whales initiated their short migration on the exact same date (2018-12-15). The increased distance from Mauritius matched the decrease in SST (extracted at the whales' positions) for all migrant whales (Fig. 3c,d). Unlike the decrease in SST values at the whale's locations between December and January, the SST extracted inside the 50% kernel increased for both years from December to January (Fig. 3e,f).

3.2 Algorithms comparison

The variation of the performance metrics across the ten simulation runs is illustrated by the box plots in Figure 4. Indeed, the small range of each boxplot indicated little sensitivity to pseudo-absence generation across the ten runs for both seasons. The mean values of the six performance metrics (accuracy, kappa, sensibility, sensitivity, specificity, F1 score and TSS) calculated from the 10-fold cross-validation were high (mean range: 0.81-0.99) for the 14 algorithms for both seasons (Fig. 4), showing good predictive performance. Based on the six performance metrics, the best model was the Random Forest (RF) for both seasons, with values ranging from 0.93 to 0.99.

When comparing the tuned RF and the stacking method during the dry season, the tuned RF approach had slightly but significantly better performance metrics compared to the stacking for the accuracy (mean: 0.978 vs. 0.970), the specificity (mean: 0.980 vs. 0.970), the F1 score (mean: 0.978 vs. 0.970) and the TSS (mean: 0.957 vs. 0.941) (Kruskal-Wallis test, $p < 0.05$, Fig. 5a,d,e,f). However, no significant difference was observed between both methods for the wet season (Kruskal-Wallis test, $p > 0.05$, Fig. 5). Given such low differences,

both methods were used to generate predictions of the whales' potential distribution for the wet and dry season separately.

3.4 Predicted distributions

The maps of the predicted distributions of the sperm whales reflected a pronounced seasonal pattern (Fig. 6). Small differences were observed between both approaches, with similar spatial patterns and globally higher probabilities for the stacking method. During the wet season, the potential distribution was widely spread around Mauritius and between 59-62°E, which coincides with the migration of the eight individuals that left the coastal areas of the island (Fig. 6a,b). In contrast, the favourable habitats during the dry season were mostly confined close to Mauritius (mostly west and south-west) and also north of Reunion Island. Some high probabilities of sperm whale presence were also identified on steep sloping habitats during both seasons, i.e. south of Mauritius (Fig. 6c,d). The most important covariates were the SSH and bottom temperature for the wet and dry season (for the tuned RF), respectively (SI Fig. 1). The coefficients of variation were globally low (<3.2%), confirming the low variability between the ten simulations for both approaches (SI Fig. 2).

3.5 Diving behaviour

A total of 500 maximum dive depths and 529 maximum dive durations were recorded from the eight whales equipped with SPLASH10 tags. The distributions of the dive depths were bimodal with mainly shallow (<500 m during the day and <400 m at night) and deep dives, i.e. between 600-1400 m during the day and 400-1400 m at night (Fig 7a). When looking at the deep dives (>200 m), a diel pattern was observed for the maximum dive depth, with significantly deeper dives during the day (mean: 1146 m) compared to night-time dives (mean: 816 m, Kruskal-Wallis test, $p<0.001$) – See Fig. 7a, b.

Similarly, the distributions of the dive durations were also bimodal, with short (<30 min during the day and <25 min at night) and long dives (between 30-70 min during the day and 25-70 min at night, Fig 7b). Dive durations lasted on average 34 min (Fig. 7b). Twenty percent of the dives were short and lasted less than 10 min, and 45% were long (between 40 and 60 min). However, there was no significant difference in terms of dive duration between day (mean=31 min) and night (mean=35 min, Kruskal-Wallis test, $p=0.9322$).

4. DISCUSSION

Using a combination of tracking data, cutting-edge ocean products and an innovative machine learning approach, our study is the first to shed light on the resident behaviour and seasonal patterns of sperm whales off Mauritius Island and to provide predictive maps of their distribution that will support conservation planning.

4.1 High-use area in Mauritius waters

Sperm whales disperse widely in all ocean basins and their global abundance estimate is in the hundreds of thousands (Whitehead 2002). Results from our limited sample size from a localized population in the South-West Indian Ocean may not be representative of the behaviour of all sperm whales but fill a critical gap in our understanding of this cryptic species. The satellite tracked individuals highlighted two critical hotspots close to Mauritius as well as a migratory route between Mauritius and Rodrigues. Among the 21 sperm whales satellite tracked, 14 remained in close proximity to Mauritius up to a maximum of 107 days. The Mascarene islands (Reunion and Mauritius islands) have previously been identified as a suitable habitat for this species (Mannocci et al. 2014b), using aerial survey data. Here, satellite tracking data have allowed both resident and migratory movements of individually tracked whales to be described in this area and quantified. Although the time spent west of Mauritius varied across individuals, the kernel densities showed two clear hotspots located west and south-west of Mauritius. These two core areas might correspond to a breeding and a nursery ground during the wet and dry season, respectively. Despite mature males being observed from September to June in Mauritius waters, a larger proportion of mature males is seen

between October and December in one of the highlighted core areas while more calves are mostly observed between March and April in the second one (M. Vely, unpublished data). The 16 months gestation period of this species (Ohsumi 1965) and a previous study showing that conception takes place in austral summer south-east of South Africa (Findlay & Best 2016) together with observations of sperm whales giving birth in April (Gambell 1966) reinforce the importance of these potential breeding and nursery habitats in Mauritius waters.

4.2 Seasonal migratory patterns

Although the tracked sperm whales showed a strong site fidelity to Mauritius waters, a significant proportion of the individuals (40%) left the island'scoastal waters to perform a short migration towards Rodrigues during the wet season. These migrant whales were all mature females except one, and in 2018, 70% of the tracked whales surprisingly showed a synchronized departure from Mauritius mid-December. These whales belong to two separate clans which are frequently observed interacting with each other (Sarano & Sarano 2017). In addition to social connections, environmental drivers might also explain such a migratory behaviour. As the whales seemed to move into cooler areas, it is possible that an increase in temperature in Mauritius waters may have impacted them either directly, by affecting their physiology (i.e. capacity to dissipate excess heat), or indirectly by impacting the distribution of their prey. Unfortunately, direct data on prey distribution were not available for this region, and proxies of micronekton biomass via mid-trophic level models (e.g. SEAPODYM) did not show temporal differences that could explain the apparent abandonment of coastal areas near Mauritius. Given that female sperm whales generally congregate into large social groups (Best & Folkens 2007), their synchronized departure could also be related to their social structure and the contrasting behaviour between males and females. Sperm whales are considered to be income breeders (Oftedal 1997) and a behavioural dichotomy is generally observed between males and females. In an Australian sperm whale population, Irvine et al. (2017) have shown that the males are present all year round whereas the females are mostly seen between April-May and September-November, suggesting a migratory behaviour similar to the one found in our study. Although some tracked females did not seem to have left Mauritius, it could simply be due to the relatively short tracking duration for these whales. The limited sample size of this study reinforces the need to track more individuals over the entire annual cycle to clarify the distribution and seasonal patterns of this sperm whale population. Although the majority of the mature males tracked from Mauritius remained in close proximity to the island, suggesting a resident behaviour, several studies indicate that mature males move to higher latitudes before and after the breeding season (Mellinger et al. 2004, Wong & Whitehead 2014, Whitehead 2018). Accordingly, the only male that left Mauritius headed southward in a straight line. This male may have headed towards Crozet or Kerguelen Archipelagos, which are famous hotspots for this species where large males are regularly observed feeding on Patagonian toothfish longline fisheries (Tixier et al. 2019).

Even though sperm whales are occasionally found in coastal waters, they must be considered pelagic animals that forage on ephemeral prey resources over large ocean scales (Kawakami 1980). The variable location of their prey resources may translate into seasonal foraging movements to maintain fitness. But to date, little is known about what drives the movements of sperm whales. In particular, nothing is known about their feeding habits in the waters around Mauritius. Their restricted and sinuous movements close to the island however suggest that they are also feeding in these waters, likely on squid, their main prey resource (Kawakami 1980). In this study, the sperm whales performed shallower dives at night, but did not seem to change the duration of their dives. Davis et al. (2007) studied diurnal vertical migrations of sperm whale and squid in the California Gulf and found that the whales followed the vertical excursions of squids in shallower depths during the night (Stewart et al. 2013), which is in agreement with the diel pattern found in our study. Squids are often considered to be sensitive to temperatures at depth and the vertical movement patterns of the sperm whales observed in this study may be in reaction to changes in squid diel vertical migration (Gilly et al. 2006). Although 11 tags were deployed to record dive data, unfortunately only a few dives were transmitted, preventing comparison of the diving behaviour between seasons and between males and females. Deployment of acoustic tags with time depth recorders and 3D accelerometers could confirm if the sperm whales are feeding in this area.

4.3 Predicted distribution and its conservation implications

An important finding from this study is that even a small sample of tracked whales can provide new and important insight into the physical and oceanographic factors that drive the movements of this cryptic species. This is mainly thanks to the novel method used here to compare models for detecting habitat selection using 14 different supervised machine learning algorithms, and to generate site specific insight into sperm whales' behaviour. Rather than using traditional algorithms (i.e. linear and additive models), we therefore recommend to test a minimum of ten different algorithms when trying to predict animal's distribution, in order to increase the predictive power of the model and get the most reliable predictions despite limited sample sizes. Our results show a strong seasonal pattern with more dispersed movements during the wet season (Dec-Mar) and affinity to contrasting environmental predictors according to the season. Our best model during the wet season showed the strongest affinity for SSH, which is in agreement with a previous study that showed higher sperm whales densities in areas of higher Sea Level Anomalies (Mannocci et al. 2014b). This suggests that sperm whales are likely looking for enriched pelagic waters that could be associated with mesoscale features (i.e. eddies, fronts) when departing from Mauritius. However, we did not find any direct relationship between the whales' tracks and eddies or fronts east of Mauritius during the wet season, probably due to the relatively low mesoscale activity in the Mascarene compared to the Mozambique Channel, where sperm whales encounter rates are much higher .(Mannocci et al. 2014b) During the dry season, the most important predictor was the bottom temperature followed by the bathymetry. This highlighted affinity for particular areas close to Mauritius that are likely associated with higher prey densities in colder waters at certain depths.

In our study, the habitat for sperm whales extended over a restricted latitudinal band (19.5-22°S), which contrasts with previous studies showing north-south migrations (Whitehead et al. 2008, Findlay & Best 2016). During the dry season, the predicted distribution was limited to coastal waters of Mauritius and Reunion islands, reinforcing the need to implement conservation measures in these areas, i.e. promote reserve designation, extend the actual MPAs. Currently, Mauritius has eight MPAs including two marine parks and six areas declared as fishing reserves. They are however relatively small (between 3.5-63.4 km^2) and confined close to shore (Francis et al. 2002). Data on animal distribution is often lacking when designing MPAs, and findings like ours are therefore essential to support conservation planning. Our results could also contribute to the regulation of the whale watching industry, which is omnipresent in such touristic areas. Restricting disturbance of animals is of particular importance at breeding sites like Mauritius coastal waters. Rather than static and sometimes inadequate MPAs, here we recommend designing dynamic MPAs based on the seasonal prediction maps of the whales (Maxwell et al. 2015). In addition to filling a gap in our knowledge about the movements and habitats of sperm whales in the South Western Indian Ocean, our study will contribute to the implementation of conservation measures in the waters of Mauritius and Reunion by clearly delineating the breeding and foraging grounds of this vulnerable species.

Data Accessibility Statement. Data available from the Dryad Digital Repository:*https://doi.org/10.5061/dryad.bnzs7h482.*

Competing interest. None declared.

Author Contributions. PC performed the data analysis and wrote the manuscript. MV, SF, DJ and MPHJ designed the experiment, collected the data and supervised the analysis. MV, SF, DJ and MPHJ participated in the field effort. PC, MV, SF, DJ and MPHJ assisted with organizing the data and analysis and interpretation of the results. All the authors shared the responsibility for contributing to the final version of the manuscript.

Funding information. The study was financed by EXAGONE Réseau TERIA who funded the field campaigns logistics, the tags and the expertise cost for the data analysis.

Acknowledgements. The authors would like to thank the Mauritian authorities who allowed us to implement the Maubydick project during three field campaigns in 2014, 2016 and 2019 and issued permits. A special thanks to Dr M. Rezah Badal and Dr Beenesh Anand Motah, Directors of the Hydrocarbon/Mineral

Production Unit Department for Continental Shelf, Maritime Zones Administration & Exploration, Ministry of Defense and Rodrigues**.** We are also grateful to the Ministry of Ocean Economy, Marine Resources, Fisheries and Shipping and especially Dr Khadun, acting as the permanent secretary for providing on board officers of the Fisheries division as well as the Commanding officer of the National Coast Guard. We are also very grateful to our friend Hugues Vitry, chairman of the Mauritius Marine Conservation Organization (M2C0), well known as the man who speaks to Sperm whales, NGO implementing the Maubydick project for its kind and continuous support and advices, for his strong affinity and knowledge about the marine megafauna, and for his active participation to the first field mission in 2014. We also thank the M2CO team and specially Alex Preud'Homm and François Sarano for providing precious information on the photo ID and sexes of the sperm whales individuals we encountered and tagged. We also thank the team of the blue water diving centre and especially the skipper Navin. We are also very grateful to our friends Jean Noel Mamet and Alain Dubois from the DOLSWIM Company as well as all the skippers and team involved, who provided the tagging boat and the skippers at the minimum cost during the three field missions and their valuable field advices. We wish to thank the company EXAGONE Réseau TERIA for the main funding of the field campaigns, the data analysis and the preparation of that current publication, and Mikkel Villum Jensen who did the tagging and prepared the tags and attachments.

References

Andrews RD, Baird RW, Calambokidis J, Goertz CEC, Gulland FMD, Heide-Jorgensen M-P, Hooker SK, Johnson M, Mate B, Mitani Y, Nowacek DP, Owen K, Quakenbush LT, Raverty S, Robbins J, Schorr GS, Shpak OV, Townsend FI, Uhart M, Wells RS, Zerbini AN (2019) Best practice guidelines for cetacean tagging. Journal of Cetacean Research and Management:27–66.

Austin MP (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling 157:101–118.

Best PB, Folkens PA (2007) Whales and dolphins of the Southern African subregion. Cambridge University Press.

Chambault P, Dalleau M, Nicet J-B, Mouquet P, Ballorain K, Jean C, Ciccione S, Bourjea J (2020) Contrasted habitats and individual plasticity drive the fine scale movements of juvenile green turtles in coastal ecosystems. Mov Ecol 8:1.

Davis RW, Jaquet N, Gendron D, Markaida U, Bazzino G, Gilly W (2007) Diving behavior of sperm whales in relation to behavior of a major prey species, the jumbo squid, in the Gulf of California, Mexico. Marine Ecology Progress Series 333:291–302.

Elith J, Graham* CH, Anderson RP, Dudík M, Ferrier S, Guisan A, Hijmans RJ, Huettmann F, Leathwick JR, Lehmann A, Li J, Lohmann LG, Loiselle BA, Manion G, Moritz C, Nakamura M, Nakazawa Y, Overton JMM, Peterson AT, Phillips SJ, Richardson K, Scachetti-Pereira R, Schapire RE, Soberon J, Williams S, Wisz MS, Zimmermann NE (2006) Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129–151.

Elith J, Leathwick JR (2009) Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics 40:677–697.

Findlay KP, Best PB (2016) Distribution and seasonal abundance of large cetaceans in the Durban whaling grounds off KwaZulu-Natal, South Africa, 1972–1975. African Journal of Marine Science 38:249–262.

Francis J, Nilsson A, Waruinge D (2002) Marine Protected Areas in the Eastern African Region: How Successful Are They? ambi 31:503–511.

Gambell R (1966) Foetal growth and the breeding season of sperm whales. Norsk Hvalfangst-tidende.

Gannier A, Drouot V, Goold JC (2002) Distribution and relative abundance of sperm whales in the Mediterranean Sea. Marine Ecology Progress Series 243:281–293.

Gannier A, Praca E (2007) SST fronts and the summer sperm whale distribution in the north-west Mediterranean Sea. Journal of the Marine Biological Association of the United Kingdom 87:187–193.

Gilly WF, Markaida U, Baxter CH, Block BA, Boustany A, Zeidberg L, Reisenbichler K, Robison B, Bazzino G, Salinas C (2006) Vertical and horizontal migrations by the jumbo squid Dosidicus gigas revealed by electronic tagging. Marine Ecology Progress Series 324:1–17.

Hattab T, Garzon-Lopez CX, Ewald M, Skowronek S, Aerts R, Horen H, Brasseur B, Gallet-Moron E, Spicher F, Decocq G, Feilhauer H, Honnay O, Kempeneers P, Schmidtlein S, Somers B, Kerchove RVD, Rocchini D, Lenoir J (2017) A unified framework to model the potential and realized distributions of invasive species within the invaded range. Diversity and Distributions 23:806–819.

Heide-Jorgensen MP, Kleivane L, OIen N, Laidre KL, Jensen MV (2001) A new technique for deploying satellite transmitters on baleen whales: tracking a blue whale (*Balaenoptera musculus* in the North Atlantic. Marine Mammal Science 17:949–954.

Holt SJ (1983) The Indian Ocean Whale Sanctuary. Ambio 12:345–347.

Hooker SK, Gerber LR (2004) Marine Reserves as a Tool for Ecosystem-Based Management: The Potential Importance of Megafauna. BioScience 54:27–39.

Huijser LAE, Estrade V, Webster I, Mouysset L, Cadinouche A, Dulau-Drouot V (2020) Vocal repertoires and insights into social structure of sperm whales (*Physeter macrocephalus*) in Mauritius, southwestern Indian Ocean. Marine Mammal Science 36:638–657.

Irvine LG, Thums M, Hanson CE, McMahon CR, Hindell MA (2017) Quantifying the energy stores of capital breeding humpback whales and income breeding sperm whales using historical whaling records. Royal Society Open Science 4:160290.

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

Jaquet N (1996) How spatial and temporal scales influence understanding of Sperm Whale distribution: a review. Mammal Review 26:51–65.

Jaquet N, Dawson S, Slooten E (2000) Seasonal distribution and diving behaviour of male sperm whales off Kaikoura: foraging implications. Can J Zool 78:407–419.

Kawakami T (1980) A review of sperm whale food. Sci Reports Whales Res:199–218.

Kie JG (2013) A rule-based ad hoc method for selecting a bandwidth in kernel home-range analyses. Animal Biotelemetry 1:13.

Lambert C, Mannocci L, Lehodey P, Ridoux V (2014) Predicting Cetacean Habitats from Their Energetic Needs and the Distribution of Their Prey in Two Contrasted Tropical Regions. PLOS ONE 9:e105958.

Laran S, Authier M, Van Canneyt O, Doremus G, Watremez P, Ridoux V (2017a) A Comprehensive Survey of Pelagic Megafauna: Their Distribution, Densities, and Taxonomic Richness in the Tropical Southwest Indian Ocean. Front Mar Sci 4.

Laran S, Drouot-Dulau V (2007) Seasonal variation of striped dolphins, fin- and sperm whales' abundance in the Ligurian Sea (Mediterranean Sea). Journal of the Marine Biological Association of the United Kingdom 87:345–352.

Laran S, Pettex E, Authier M, Blanck A, David L, Doremus G, Falchetto H, Monestiez P, Van Canneyt O, Ridoux V (2017b) Seasonal distribution and abundance of cetaceans within French waters- Part I: The North-Western Mediterranean, including the Pelagos sanctuary. Deep Sea Research Part II: Topical Studies in Oceanography 141:20–30.

Lopez R, Malarde J-P, Royer F, Gaspar P (2014) Improving Argos Doppler Location Using Multiple-Model Kalman Filtering. IEEE Transactions on Geoscience and Remote Sensing 52:4744–4755.

Mannocci L, Catalogna M, Doremus G, Laran S, Lehodey P, Massart W, Monestiez P, Van Canneyt O, Watremez P, Ridoux V (2014a) Predicting cetacean and seabird habitats across a productivity gradient in the South Pacific gyre. Progress in Oceanography 120:383–398.

Mannocci L, Laran S, Monestiez P, Doremus G, Canneyt OV, Watremez P, Ridoux V (2014b) Predicting top predator habitats in the Southwest Indian Ocean. Ecography 37:261–278.

Maxwell SM, Hazen EL, Lewison RL, Dunn DC, Bailey H, Bograd SJ, Briscoe DK, Fossette S, Hobday AJ, Bennett M, Benson S, Caldwell MR, Costa DP, Dewar H, Eguchi T, Hazen L, Kohin S, Sippel T, Crowder LB (2015) Dynamic ocean management: Defining and conceptualizing real-time management of the ocean. Marine Policy 58:42–50.

Mellinger DK, Stafford KM, Fox CG (2004) Seasonal Occurrence of Sperm Whale (*Physeter Macrocephalus*) Sounds in the Gulf of Alaska, 1999–2001. Marine Mammal Science 20:48–62.

Oftedal OT (1997) Lactation in Whales and Dolphins: Evidence of Divergence Between Baleen- and Toothed-Species. J Mammary Gland Biol Neoplasia 2:205–230.

Ohsumi S (1965) Reproduction of the sperm whale in the North-West Pacific. Scientific Reports of the Whales Research Institute, Tokyo.

Perrin WF, Wursig B, Thewissen JGM (2009) Encyclopedia of Marine Mammals. Academic Press.

Pirotta E, Brotons JM, Cerda M, Bakkers S, Rendell LE (2020) Multi-scale analysis reveals changing distribution patterns and the influence of social structure on the habitat use of an endangered marine predator, the sperm whale *Physeter macrocephalus* in the Western Mediterranean Sea. Deep Sea Research Part I: Oceanographic Research Papers 155:103169.

Pirotta E, Matthiopoulos J, MacKenzie M, Scott-Hayward L, Rendell L (2011) Modelling sperm whale habitat preference:: a novel approach combining transect and follow data. Marine Ecology Progress Series 436:257–272.

Sarano F, Sarano M (2017) Cartes d'identite des cachalots de l'ile Maurice. Mauritius Marine Conservation Organization (M2CO, Mauritius.

Schickele A, Leroy B, Beaugrand G, Goberville E, Hattab T, Francour P, Raybaud V (2020) Modelling European small pelagic fish distribution: Methodological insights. Ecological Modelling 416:108902.

Senay SD, Worner SP, Ikeda T (2013) Novel Three-Step Pseudo-Absence Selection Technique for Improved Species Distribution Modelling. PLOS ONE 8:e71218.

Stewart JS, Field JC, Markaida U, Gilly WF (2013) Behavioral ecology of jumbo squid (*Dosidicus gigas*) in relation to oxygen minimum zones. Deep Sea Research Part II: Topical Studies in Oceanography 95:197–208.

Taylor B, Baird R, Barlow J, Ford S, Mead J, Notarbartolo di Sciara G, Wade P, Pitman RL (2019) *Physeter macrocephalus* (amended version of 2008 assessment).

Teloni V, Mark JP, Patrick MJO, Peter MT (2008) Shallow food for deep divers: Dynamic foraging behavior of male sperm whales in a high latitude habitat. Journal of Experimental Marine Biology and Ecology 354:119–131.

Tixier P, Welsford DC, Lea M-A, Hindell MA, Guinet C, Janc A, Richard G, Gasco N, Duhamel G, Arangio R, Villanueva MC, Suberg L, Arnould JPY (2019) Fisheries interaction data suggest variations in the distribution of sperm whales on the Kerguelen Plateau. http://heardisland.antarctica.gov.au/research/kerguelen-plateau-symposium/the-kerguelen-plateau-marine-ecosystems-and-fisheries (accessed June 1, 2020)

Virgili A, Authier M, Boisseau O, Canadas A, Claridge D, Cole T, Corkeron P, Doremus G, David L, Di-Meglio N, Dunn C, Dunn TE, Garcia-Baron I, Laran S, Lauriano G, Lewis M, Louzao M, Mannocci L, Martinez-Cedeira J, Palka D, Panigada S, Pettex E, Roberts JJ, Ruiz L, Saavedra C, Santos MB, Canneyt OV, Bonales JAV, Monestiez P, Ridoux V (2019) Combining multiple visual surveys to model the habitat of deep-diving cetaceans at the basin scale. Global Ecology and Biogeography 28:300–314.

Virgili A, Racine M, Authier M, Monestiez P, Ridoux V (2017) Comparison of habitat models for scarcely detected species. Ecological Modelling 346:88–98.

Wahlberg M (2002) The acoustic behaviour of diving sperm whales observed with a hydrophone array. Journal of Experimental Marine Biology and Ecology 281:53–62.

Watkins WA, Daher MA, Dimarzio NA, Samuels A, Wartzok D, Fristrup KM, Gannon DP, Howey PW, Maiefski RR, Spradlin TR (1999) Sperm Whale Surface Activity from Tracking by Radio and Satellite Tags1. Marine Mammal Science 15:1158–1180.

Watwood SL, Miller PJO, Johnson M, Madsen PT, Tyack PL (2006) Deep-diving foraging behaviour of sperm whales (*Physeter macrocephalus*). Journal of Animal Ecology 75:814–825.

Whitehead H (2002) Estimates of the current global population size and historical trajectory for sperm whales. Marine Ecology Progress Series 242:295–304.

Whitehead H (2018) Sperm Whale: *Physeter macrocephalus* . In:*Encyclopedia of Marine Mammals (Third Edition)* . Wursig B, Thewissen JGM, Kovacs KM (eds) Academic Press, p 919–925

Whitehead H, Coakes A, Jaquet N, Lusseau S (2008) Movements of sperm whales in the tropical Pacific. Marine Ecology Progress Series 361:291–300.

Whitehead H, Rendell L (2004) Movements, habitat use and feeding success of cultural clans of South Pacific sperm whales. Journal of Animal Ecology 73:190–196.

Wong SNP, Whitehead H (2014) Seasonal occurrence of sperm whales (*Physeter macrocephalus*) around Kelvin Seamount in the Sargasso Sea in relation to oceanographic processes. Deep Sea Research Part I: Oceanographic Research Papers 91:10–16.

Worton (1989) Kernel Methods for Estimating the Utilization Distribution in Home-Range Studies. Ecology 70:164–168.

FIGURES

Fig. 1. Map of the study area located in the South-West Indian Ocean. Panel (b) refers to Mauritius and Reunion Islands with the whales' locations in orange. Panel (c) to the tagging deployment locations along the west coast of Mauritius. The black dotted lines refer to the Exclusive Economic Zone.

Fig. 2. Locations of the sperm whales during the (a) wet and (b) dry seasons. (c,d) Maps of the bathymetry (expressed in m) and (e,f) the slope (expressed in degrees) over the study area. The utilisation distribution (50% and 90% contours) were superimposed for the wet (left panel) and dry (right panel) seasons. MUS refers to Mauritius Island, REU to Reunion Island and RDG to Rodrigues Island.

Fig. 3 . (a,b) Distance to the tagging site over time for the eight migrant whales tracked in 2014 and 2018. The vertical dotted lines refer to the departure date: 2014-12-25 in (a,c) and 2018-12-15 in (b,d). (c,d) SST extracted at the whale's locations in 2014 and 2018. (e,f) Box plots of the SST extracted inside the kernel 50% during the wet season in 2014 and 2018.

Fig. 4. Box plots of the six performance metrics calculated for each of the 14 models and for each season. The values of each box plot include the performance metrics of each of the 10 simulation runs.

Fig. 5. Box plots of the six performance metrics calculated for the tuned Random Forest model (RF tuned) and the stacking method for each season. The values of each box plot include the performance metrics of

each of the 10 simulation runs.

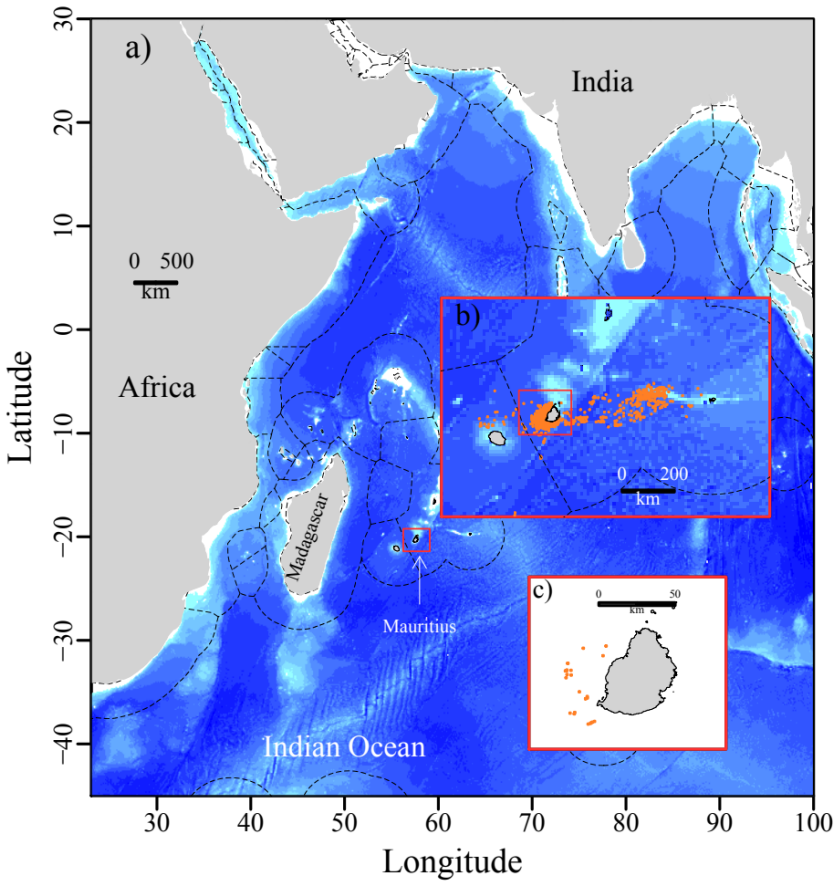
Fig. 6. Averaged prediction maps of the sperm whales' probabilities during the (a, b) wet and (c, d) dry seasons calculated from the tuned Random Forest model (a, c) and the stacking method (b, d). 0 indicates a very low probability to see a whale, and 1 a high probability to see a whale.

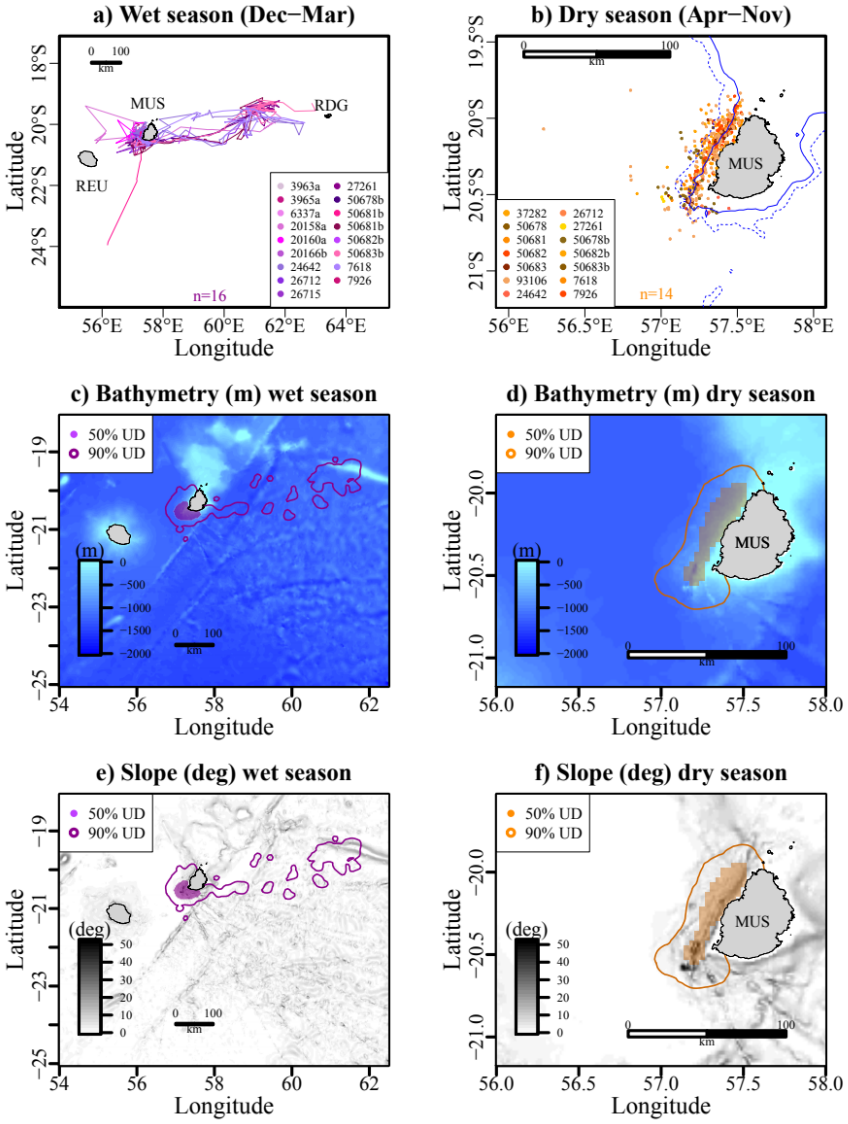
Fig. 7. Kernel densities of the (a) maximum dive depth and the (b) maximum dive duration according to day (in red) and night (in blue).

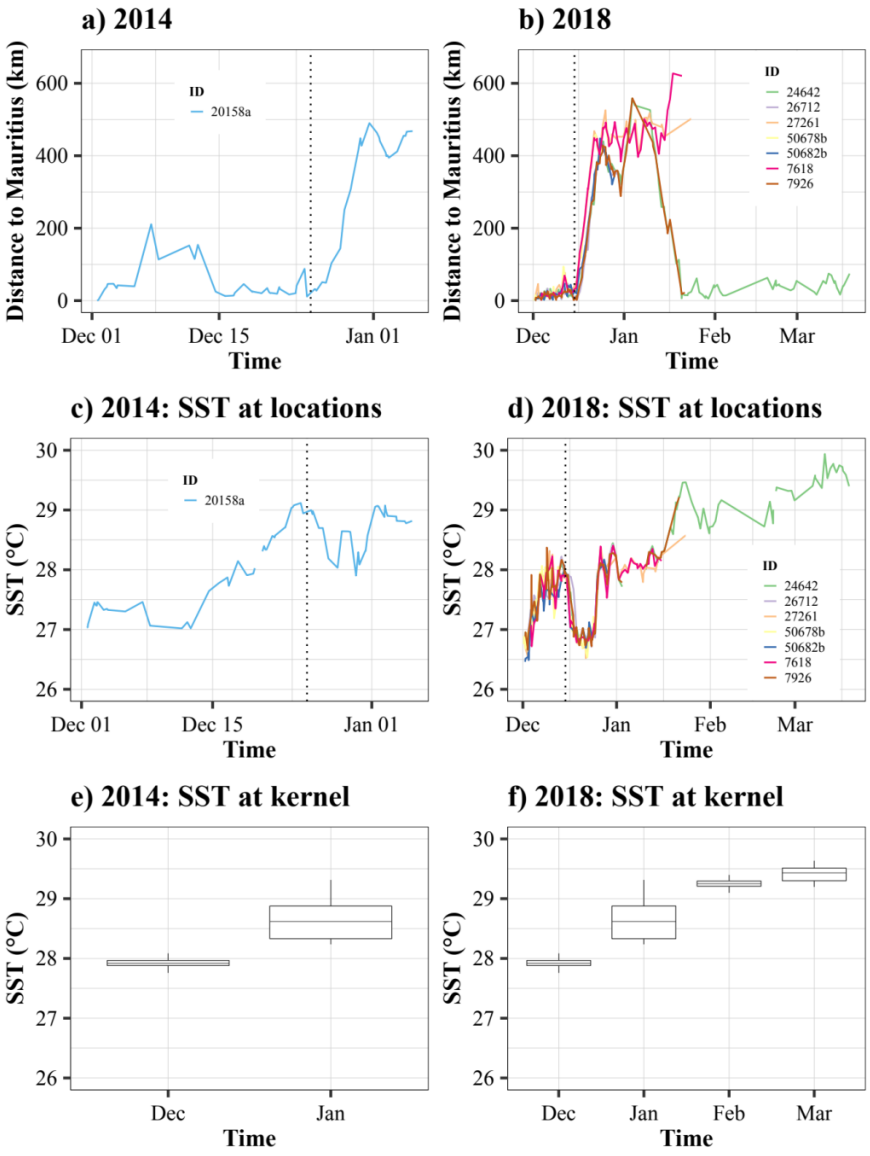
SUPPLEMENTARY INFORMATION

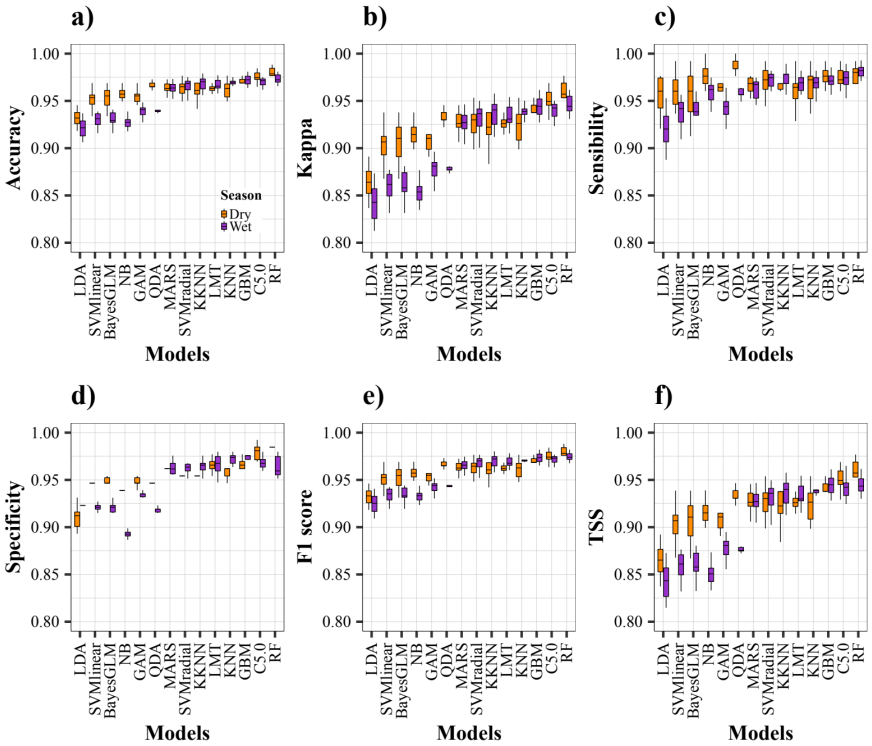
Fig.SI1. Boxplots of the covariates importance for the tuned Random Forest model and each season.

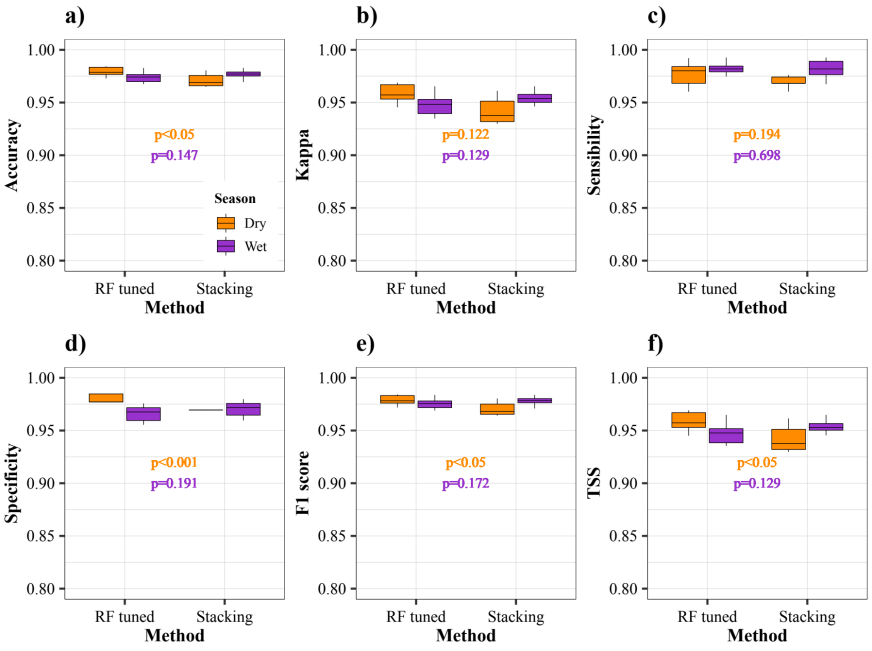
Fig.SI2. Maps of the coefficients of variation (expressed in percentage) during the (a, b) wet and (c, d) dry seasons calculated from the tuned Random Forest model (a, c) and the stacking method (b, d).

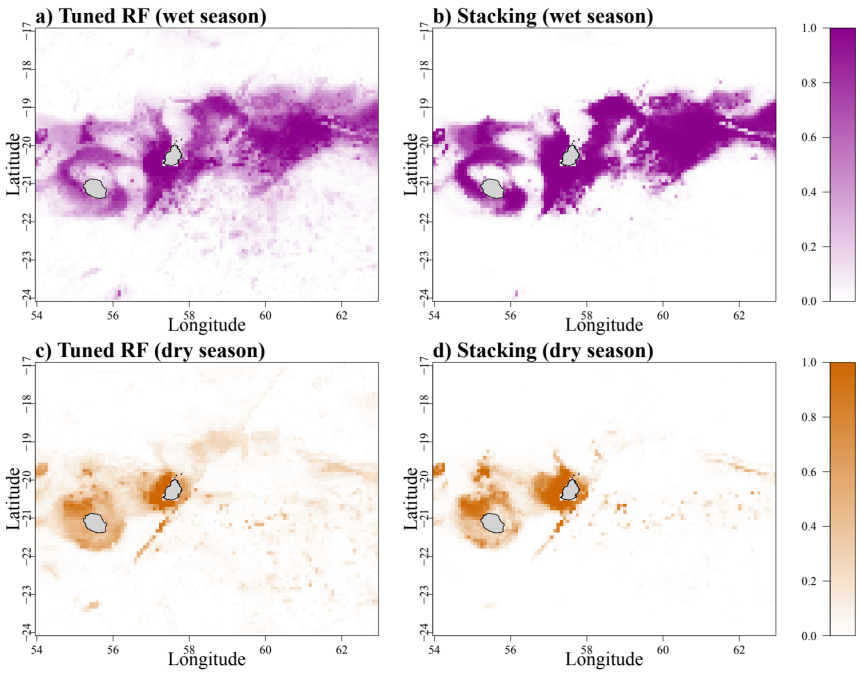


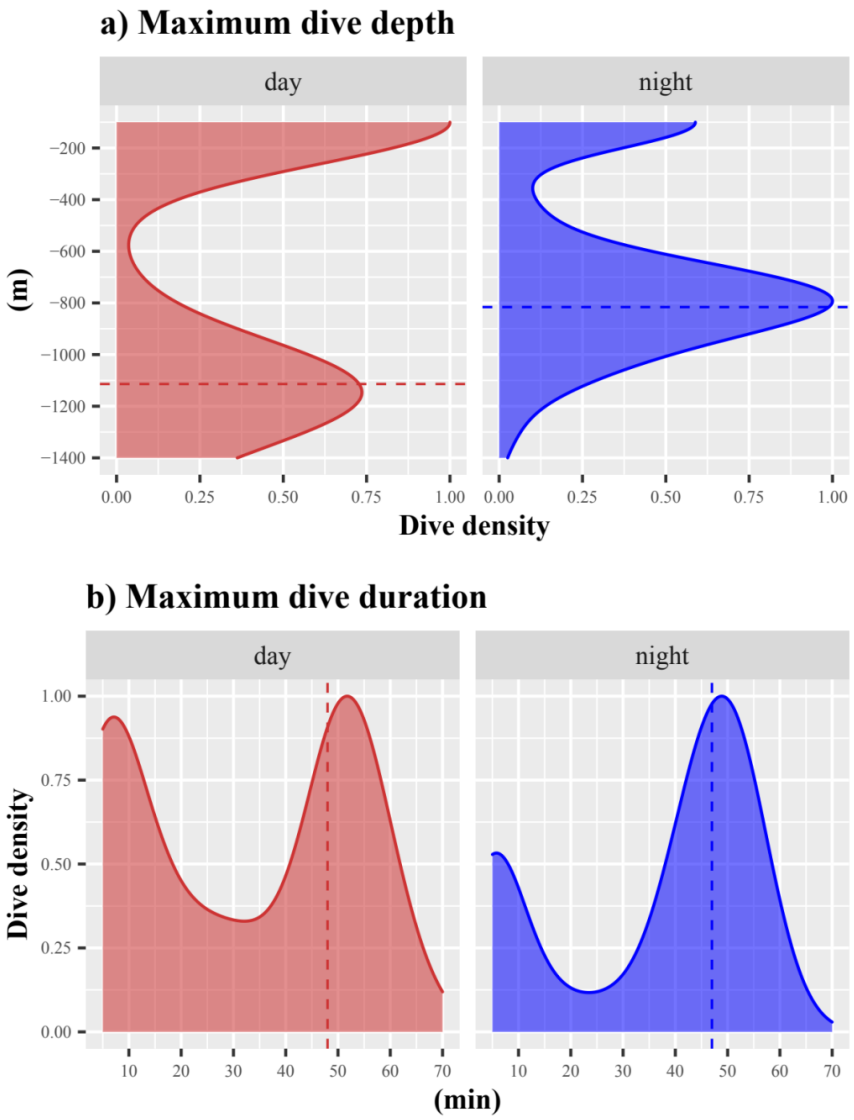












Hosted file

`Table1_EcolEvo.docx` available at https://authorea.com/users/348528/articles/473801-using-machine-learning-models-to-predict-the-distribution-of-a-cryptic-marine-species-the-sperm-whale