Vol. 458: 199–211, 2012
doi: 10.3354/meps09777

MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser

Published July 3

OPEN ACCESS

Feeding habitat of the whale shark *Rhincodon typus* in the northern Gulf of Mexico determined using species distribution modelling

Jennifer A. McKinney[1,4,*], Eric R. Hoffmayer[2], Wei Wu[3], Richard Fulford[3], Jill M. Hendon[1]

[1]Center for Fisheries Research and Development, University of Southern Mississippi, Gulf Coast Research Laboratory, Ocean Springs, Mississippi 39564, USA

[2]National Oceanic and Atmospheric Administration, National Marine Fisheries Service, Mississippi Laboratories, 3209 Frederick Street, Pascagoula, Mississippi 39567, USA

[3]Department of Coastal Studies, University of Southern Mississippi, Gulf Coast Research Laboratory, Ocean Springs, Mississippi 39564, USA

[4]*Present address:* Louisiana Department of Wildlife and Fisheries, Fisheries Management Division, 2000 Quail Dr., Baton Rouge, Louisiana 70808, USA

ABSTRACT: Whale shark *Rhincodon typus* is a globally distributed species, but there is a lack of knowledge pertaining to their biology, seasonal occurrence, and distribution in the northern Gulf of Mexico (NGOM). Understanding critical habitat for whale sharks is essential on both a regional and global basis for proper management because of their large migratory range. The goal of the present study was to describe the regional distribution of whale shark feeding aggregations in the NGOM by exploiting a presence-only dataset collected as a part of a volunteer sighting survey. Whale shark aggregations have been documented in large numbers in the NGOM since 2003, and species distribution models provide a unique approach to analyzing these presence data. We used maximum entropy and ecological niche factor analysis, 2 algorithms designed for predicting species distribution based only on presence data, to analyze data for the summer period in 2008 and 2009. Cohen's kappa (kappa) and the 'area under the receiver operating characteristic curve' (AUC) were used to evaluate model performance with an external testing dataset. Kappa values ranged from 0.28 to 0.69, and AUC values ranged from 0.73 to 0.80, indicating that the predicted distribution had a fair to substantial agreement with the testing data. Distance to continental shelf edge, distance to adjacent petroleum platforms, and chlorophyll *a* were the variables most strongly related to whale shark sightings, likely due to an association of these variables with high food availability. Suitable habitat was predicted along the continental shelf edge, with the most suitable habitat predicted south of the Mississippi River Delta. The spatial distribution of suitable habitat is dynamic; therefore, a multi-year study is underway to better delineate temporal trends in regional whale shark distribution and to identify consistent areas of high suitability. Presence-only habitat models are a powerful tool for delineating important regional habitat for a vulnerable, highly migratory species.

KEY WORDS: Whale shark · Distribution · MaxEnt · ENFA · AUC · Kappa

INTRODUCTION

Effective management can be challenging for wide-ranging species because often their geographic range and habitat requirements are poorly understood and may include several political jurisdictions (Pearce & Boyce 2006). Species distribution modelling is an effective and useful approach for overcoming logistical constraints on data collection required to determine habitat preferences within a species' range.

*Email: jmckinney@wlf.la.gov

Optimally, species distribution data come from a fully randomized monitoring program, but, in many cases, regional species distributions can only be described with non-random sampling approaches that give only presence information. Such data are not applicable to certain modelling techniques, which require accurate absence locations, such as generalized linear models (GLM). However, presence locations provide viable information on habitat choice, and by quantifying the relationship with predictor variables, presence-only modelling can generate predictions at unsampled locations throughout the study area (Guisan & Thuiller 2005). For a species in which systematic survey data is not available or difficult to obtain, presence-only modelling broadens the range of available data for model building to include museum collections and volunteer stakeholder surveys (Hirzel et al. 2002). Presence-only modelling has been successfully applied to many terrestrial and marine species, including migratory birds (Peterson 2001), marine fishes (Wiley et al. 2003), stony coral seamounts (Tittensor et al. 2009), and *Odontoceti* whales (Praca et al. 2009).

There is also value in developing regional models of population distributions for highly migratory species due to changing habitat preferences between migratory endpoints, as well as separation between habitats during different life-history stages. Global distribution models describing species with large ranges typically demonstrate less accurate predictions when modelled as a single population, but when subpopulations were modelled individually, the model accuracy increased; this was likely due to regional adaptations in habitat quality and preference (Stockwell & Peterson 2002). Therefore, the best approach may be to first gain an understanding of population distributions on a regional scale, which can then be integrated with other regional studies to develop a global habitat model. Meaningfully partitioning datasets based on known seasonal, behavioral, or life-history stages can allow the modeler to gain insight on the critical habitat distribution throughout these times, thereby reducing generalist predictions. The habitat of other large pelagic species, such as Atlantic bluefin tuna *Thunnus thynnus* and swordfish *Xiphias gladius*, has been shown to vary considerably between feeding and spawning periods throughout the year in the Mediterranean (Tserpes et al. 2008, Druon et al. 2011). In the Indian Ocean, seasonal habitat distribution of whale sharks *Rhincodon typus* has been modelled using long-term fisheries-based datasets (Sequeira et al. 2012).

Whale sharks would be an excellent choice for presence-only modelling for multiple reasons. (1) Whale shark aggregations are known to occur in coastal areas with some predictability in several locations circumglobally through the development of directed research and ecotourism industry (Stevens 2007), thereby resulting in a large database of presence locations (though often lacking associated absence data). (2) Whale sharks are currently listed as a 'Vulnerable' species by the International Union for the Conservation of Nature and Natural Resources (IUCN; www.redlist.org) and included in Appendix II of the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES; www.cites.org). These listings are based on their susceptibility to fisheries collapse and slow recovery due to their k-selected life-history characteristics (Stevens 2007). Therefore, baseline ecological data on whale sharks are essential to appropriately assess population risks and measure the success of protective legislature. (3) Remotely sensed imagery and the geographic information system (GIS) can characterize the features that affect the biological productivity that whale sharks are associated with and could therefore be used to provide the broad-scale, continuous environmental data required for modelling whale shark distribution.

Whale sharks have a circumglobal distribution, inhabiting all tropical and warm temperate waters, except the Mediterranean Sea (Compagno 1984). Whale sharks are opportunistic filter-feeders that require large densities of prey to meet their energetic demands (Compagno 1984). They are known to aggregate in areas of high biological productivity, such as near plankton blooms, fish spawns, and areas where changes in water temperatures occur (Compagno 1984, Taylor & Pearce 1999, Heyman et al. 2001, Hoffmayer et al. 2005, Taylor 2007). Their movements may correspond to bathymetric and/or oceanographic features, such as thermal fronts, eddies, currents, and zones of high chlorophyll *a* concentration (Taylor & Pearce 1999, Hoffmayer et al. 2005, Hsu et al. 2007, Kumari & Raman 2010), likely due to the presence of zooplankton assemblages and fish populations that are known to accumulate near these features (Balch & Byrne 1994). Their distribution is also believed to be linked to specific environmental conditions, such as narrow temperature ranges and areas of upwelling (Colman 1997, Sequeira et al. 2011). In the northern Gulf of Mexico (NGOM), over 4000 rigs function as the largest artificial reef system worldwide, which is inhabited by many reef and pelagic species (Franks 2000), including whale sharks (Hoffmayer et al. 2005).

The present study aims to describe the spatial distribution of whale shark feeding aggregations in the NGOM during their peak season by employing 2 species distribution modelling approaches. We hope to (1) understand the key environmental factors associated with whale shark feeding aggregations, (2) derive distribution and habitat preference maps for the region, and (3) evaluate the models' predictive performance.

METHODS

Two modelling approaches that are effective at modelling with presence-only data were applied to describe the spatial distribution of whale sharks *Rhincodon typus* in the NGOM: maximum entropy (MaxEnt) (Phillips et al. 2006) and ecological niche factor analysis (ENFA) (Hirzel et al. 2002, 2007). MaxEnt is a machine-learning technique that attempts to fit a probability distribution of species occurrence over the entire study area using a complex suite of transformations on the environmental variables believed to be important to the target species (Phillips et al. 2006, Elith et al. 2006). On the other hand, the ENFA approach is similar to a principal component analysis, in which a series of orthogonal factors are determined based on linear combinations of the environmental variables that describe the relationship between the values found at the locations where species are present with those of the entire study area (Hirzel et al. 2002). The ENFA approach takes into account the mean and standard deviation of all environmental variables for the entire study area and for the cells in which the species is present. The ENFA output allows one to look at the relationship between these key factors using marginality and specialization scores. Marginality is the standardized difference between the species mean and the study area mean, whereas specialization is the ratio of the study area's and species' standard deviations. The predictive model is then constructed so that the first factor accounts for all of the species' marginality and the majority of the specialization is accounted for by the subsequent factors (Hirzel et al. 2002). A suitability score for each cell within the study area is then assigned, based on the associations between the presence records and the environmental variables.

Using both model types, each year of the study (2008 and 2009) was modelled separately, due to temporal variability in whale shark distribution and environmental variables. Each modelling approach was implemented with its stand-alone software, available for download online. The Biomapper toolkit (www.unil.ch/biomapper) includes a set of GIS and statistical tools to develop an ENFA model for any target species and creates a habitat-suitability map based on optimal model parameterization (Hirzel et al. 2007). MaxEnt software (www.cs.princeton.edu/~schapire/MaxEnt) uses a deterministic algorithm along with complex feature classes (linear, quadratic, product, threshold, hinge, and categorical) to converge on the probability distribution of maximum entropy (Phillips et al. 2006, Phillips & Dudík 2008).

Data preparation

Spatial data (environmental predictor variables and whale shark presence data) were prepared using ArcGIS (ESRI Corporation) and IDRISI (Clark Labs at University of Clark) GIS software. The working cell size was determined by the highest resolution of areal predictor data, which was about 1.1 km^2. All spatial data were resampled to the same cell size (0.009° ~ 1 km^2), geographic datum (WGS 1984), and spatial extent (Fig. 1; latitude: 25 to 31°N, longitude: 82 to 98°W). The temporal period of June through September in both 2008 and 2009 was selected so that the environmental data corresponded to the time during which most of the presence data were collected.

Presence data

The University of Southern Mississippi, Gulf Coast Research Laboratory's Northern Gulf of Mexico Whale Shark Sightings Survey (WSSS), originated in 2003 (Hoffmayer et al. 2005), has compiled whale shark sightings throughout the NGOM from Port Aransas, Texas, to Tampa Bay, Florida. Sightings data include: date, time and duration of the encounter, location (GPS coordinates), approximate size and number of individuals, observed behavior, and, if available, photographs. Validity of reports was confirmed via photographs or verbal descriptions of observations. Reports deemed as duplicate were removed from the analysis. The survey received a total of 70 reports in 2008 and 176 reports in 2009. For the present study, any report that included 2 or more individuals was considered an aggregation. For analysis, each 1 km grid cell in the study area was either considered present or absent, based on the WSSS results; therefore, even if several aggregation

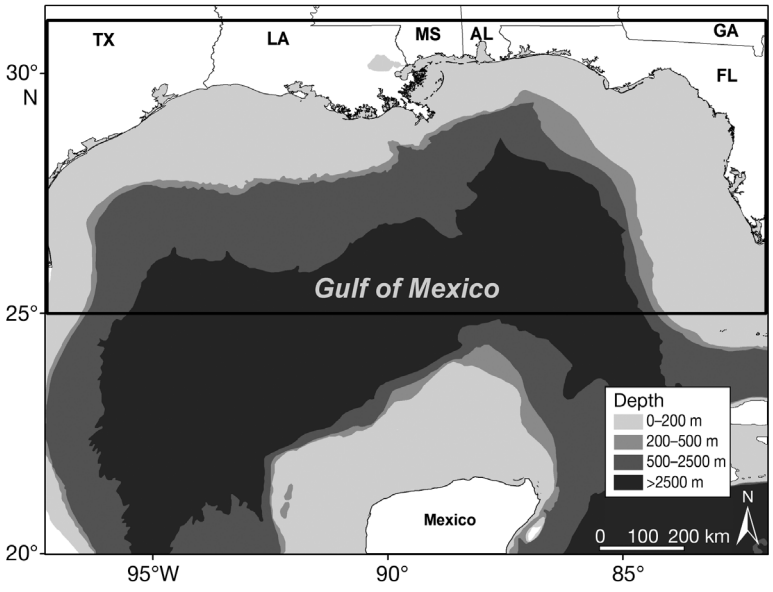

Fig. 1. Map of the Gulf of Mexico basin depicting bathymetry (gray shading) and delineating the extent of the study area (black box)

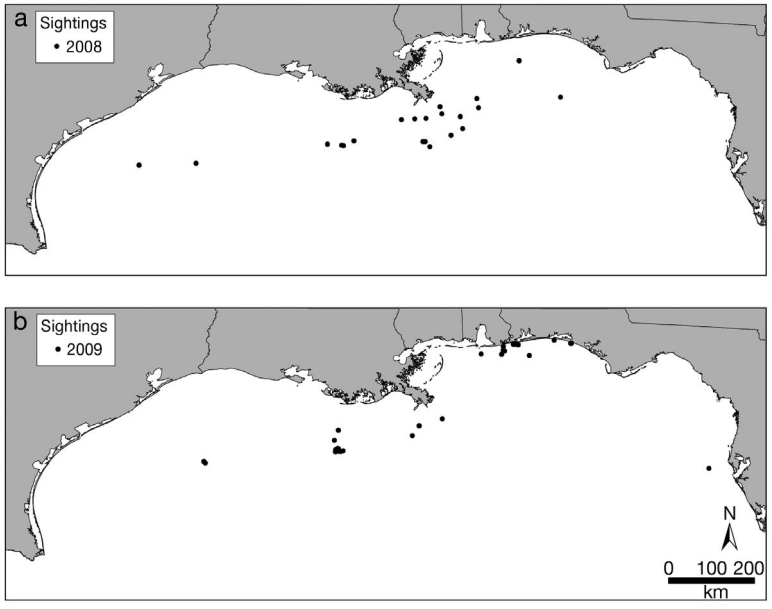

Fig. 2. Maps depicting the sightings data used to generate *Rhincodon typus* distribution models for (a) 2008 and (b) 2009

reports occurred within the same grid, it was only counted once. The number of records used in our study was similar between years (2008, n = 21; 2009, n = 24), though the spatial distribution varied greatly (Fig. 2). Herein, models will be referenced by the algorithm and year (i.e. MaxEnt 2008, ENFA 2008).

Environmental data

Environmental variables used in the present study consisted of a suite of spatial data derived from remotely sensed images and GIS data (Table 1). All variables were selected because they serve as a proxy for food availability or characterize the environmental habitat associated with whale shark distribution described in the primary literature. A preliminary analysis was conducted to ensure multi-collinearity was not present amongst variables, using an unweighted pair-group method using arithmetic averages (UPGMA) (Sneath & Sokal 1973), the results of which are shown in Appendix 1 (Fig. A1). The following independent variables were included in the final models: bathymetric slope (Slope), distance to continental shelf-edge (Shelf), distance to nearest petroleum platform (Drig), density of petroleum platforms within 1 km radius (Rigden), chlorophyll *a* concentration (Chl), sea-surface temperature (SST), and sea-surface height (SSH) (Table 1).

Test data

An independent dataset of whale shark distribution was developed to test the predictive capacity of all models. The testing dataset consisted of 80 presence and 80 pseudo-absence locations derived separately from the dataset used to build the model (Fig. 3). The presence locations during the same season (June to September) were extracted from aerial survey data collected by the National Marine Fisheries Service, Mississippi Laboratories, over the entire NGOM slope waters from Tampa Bay, Florida, to Brownsville, Texas, between 1989 and 1998 (Burks et al. 2006). Pseudo-absence locations were randomly generated in ArcGIS in areas outside of the observed locations. This dataset could not be used for model building due to the lack of availability of remotely sensed variables from these years; however, it did provide an opportunity to utilize and independently test the data, rather than

Table 1. Predictor variables used in whale shark *Rhincodon typus* species distribution models

Predictor variable	Abbreviation	Unit	Source	Source resolution
Bathymetric slope	Slope	% rise km^{-2}	ArcGIS derived[a]	1 km
Distance to shelf	Shelf	m	ArcGIS derived[a]	–
Distance to rig	Drig	m	ArcGIS derived[b]	–
Rig density	Rigden	no. of rigs km^{-2}	ArcGIS derived[b]	–
Chlorophyll *a*	Chl	mg m^{-3}	SeaWIFS[c]	4 km
Sea-surface height	SSH	cm – mean sea level	Aviso[d]	7 km
Sea-surface temp.	SST	°C	AVHRR[e]	1.1 km

[a]Bathymetry data obtained by National Geophysical Data Center (www.ngdc.noaa.gov/mgg/mggd.html) GEOphysical Data System
[b]Petroleum platform data obtained by the Minerals Management Service (www.gomr.mms.gov/homepg/pubinfo/repcat/arcinfo/index.html)
[c]Giovanni (GES-DISC Online Visualization and Analysis Infrastructure) web portal, as part of the National Aeronautics and Space Administration Goddard Earth Sciences Data and Information Services Center
[d]www.aviso.oceanobs.com/en/data/products/sea-surface-height-products/regional/index.html using the 'Download Aviso SSH dataset to ArcGIS Rasters' tool in the 'Marine Geospatial Ecology' Toolbox (Roberts et al. 2010)
[e]National Oceanic and Atmospheric Administration's advanced very high resolution radiometer (AVHRR) (ftp://eclipse.ncdc.noaa.gov/pub/OI-daily/NetCDF/)

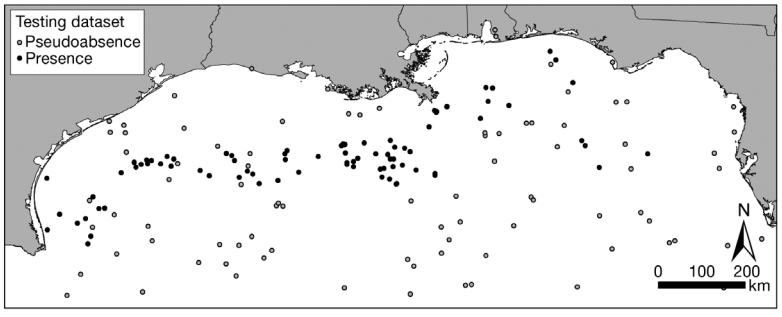

Fig. 3. Map depicting the testing dataset used to evaluate model performance. Presence data (black dots) provided by the National Marine Fisheries Service, Mississippi Laboratories, were collected during aerial surveys targeting marine mammals in the northern Gulf of Mexico from 1989 to 1998. Pseudo-absence data (grey dots) were generated using a random point generator in ArcMap 9.3 (ESRI Corp.)

splitting them up to conduct cross-validation. The main advantage of independent testing was that the data were collected using a stratified-sampling methodology, rather than through volunteer surveys, therefore avoiding potential observer bias. Although cross-validation may be the most commonly used testing approach, we did not employ this test, as it requires splitting the presence data into 2 separate datasets (model building and model testing); the sample size for the presence data used to develop the models was too low to be appropriate for a cross-validation approach. Additionally, we believe that the external testing dataset used can determine the robustness of the predictive models relative to the historically known whale shark distribution throughout the NGOM.

Model evaluation

All models were calibrated using whale shark sighting locations as a response variable and evaluated using the independent testing dataset. Model performance was based on 2 statistics: the 'area under the receiver operating characteristic curve' (AUC) (Wiley et al. 2003) and Cohen's kappa (Cohen 1960). The AUC represents the probability that a random positive instance and a random negative instance are correctly classified. In presence-only modelling, the technique is applied to distinguish presence from random occurrance, rather than presence from absence (Phillips et al. 2006). This approach defines negative instances as x_{random} so that the AUC is then an indicator of whether the model predicts species distribution better than random, and can be used to prove statistical significance. An AUC value of >0.5 indicates the model performed 'better than random,' of 0.5 indicates 'random,' and a value <0.5 means the model actually performed worse than random (Phillips et al. 2006).

Cohen's kappa statistic (hereafter referred to as kappa) is a popular measure of accuracy because it corrects for the expected accuracy due to chance (Allouche et al. 2006). The kappa score ranges from −1 to +1, where a value <0 indicates the model is performing no better than random and an accepted performance rating is as follows: 0 to 0.2 = slight agreement, 0.21 to 0.4 = fair, 0.41 to 0.6 = moderate, 0.61 to

0.8 = substantial, and 0.81 to 1.0 = near perfect agreement (Cohen 1960, Landis & Koch 1977). It is also sensitive to prevalence (the proportion of presence points) in the testing dataset (McPherson et al. 2004). In the present study, the testing dataset has an equal number of presence and absence locations, in order to eliminate this potential bias.

Although calculated in very different ways, each model predicts a habitat suitability score throughout the entire study area. We considered areas of high suitability to be those areas with suitability scores >75. Spatial prediction of highly suitable habitat, variable contribution to predictions, and predictive accuracies were compared for all model runs to assess the suitable habitat and environmental gradients occupied by whale sharks in the NGOM.

Table 2. Descriptive statistics of environmental data throughout the study area (background) and at whale shark *Rhincodon typus* aggregation locations in 2008 and 2009. For temporally explicit variables, background data was provided for each year of the study (Bkgrnd-08/09). Variable abbreviations as in Table 1

Variable	Dataset	Mean	SD	Min.	Max.
Slope	Background	1.13	3.61	0.00	74.17
	2008	1.75	1.48	0.22	5.78
	2009	0.70	1.09	0.04	4.46
Shelf	Background	173748.58	108443.91	0.00	557343.38
	2008	39644.38	46060.03	5362.05	213216.83
	2009	45901.81	37241.26	0.00	168371.67
Drig	Background	171850.39	170656.17	0.00	718934.25
	2008	15660.11	30374.68	0.00	136548.06
	2009	44731.37	95412.40	0.00	458917.25
Rigden	Background	0.01	0.10	0.00	15.00
	2008	0.10	0.29	0.00	1.00
	2009	0.16	0.37	0.00	1.00
Chl	Bkgrnd-08	1.32	3.74	0.00	52.12
	Bkgrnd-09	1.25	3.81	0.00	49.56
	2008	1.54	2.24	0.18	10.75
	2009	1.31	1.36	0.18	5.04
SSH	Bkgrnd-08	8.40	6.71	−12.67	41.70
	Bkgrnd-09	9.05	6.25	−7.02	26.98
	2008	9.12	0.86	8.05	10.95
	2009	7.93	1.51	6.05	9.94
SST	Bkgrnd-08	28.62	0.27	27.52	29.27
	Bkgrnd-09	29.31	0.38	28.00	29.83
	2008	28.40	0.12	28.27	28.75
	2009	28.90	0.53	28.00	29.48

RESULTS

Model results indicate that whale shark feeding habitat in the NGOM is comprised of a subset of the habitat available; for nearly all environmental variables modeled, the range of data values where whale sharks were present was restricted compared to the range of that variable throughout the entire study area, or 'background' (Table 2). SST is the one exception, which had a very limited range available (approximately 2°), and thus the whale shark data correspond closely to the background values. The mean values of SSH at whale shark locations were similar to the background values; however, whale sharks were only found in waters that were at least 2 cm above mean sea level (Table 2). In 2008, whale shark aggregations utilized areas in closer proximity to the continental shelf-edge and were therefore characterized by steeper bathymetric slopes, compared to 2009. Furthermore, the mean distance to rigs increased from 16 km in 2008 to 45 km in 2009 (Table 2). In both years, whale shark aggregations were observed in areas with a minimum chlorophyll concentration of 0.18 mg m^{-3}, but ranged as high as 10.75 mg m^{-3} in 2008 and 5.04 mg m^{-3} in 2009.

All model runs for both ENFA and MaxEnt performed better than random. The AUC scores ranged from 0.686 to 0.803 (Table 3). The kappa scores indicated that models had fair to substantial agreement with the testing dataset (0.283 to 0.761; Table 3). The MaxEnt 2008 model produced the highest AUC and kappa scores overall (0.803 and 0.761, respectively). Conversely, the 2009 dataset produced lower evaluation scores in both modelling platforms.

Distance to petroleum platforms and distance to the continental shelf edge were the most influential variables in the MaxEnt models; however, the proportions were nearly inverted between years of the study (Fig. 4). Chl concentration was moderately influential in both years modelled, in relatively equal proportions. Bathymetric slope was the only other variable influential in 2008, but it was not influential in 2009. SST and density of petroleum platforms were influential in 2009 only. SSH was not influential in creating the MaxEnt prediction in either year.

The amount of specialization explained by the first 3 factors of the ENFA models was >90% in both years. The multi-dimensional niche width utilized by whale sharks was narrower in 2008 than in 2009, indicated by a larger amount of specialization

Table 3. Summary of results of model evaluation for whale shark *Rhincodon typus* species distribution models. Statistics include: 'area under the receiver characteristic curve' (AUC) and Cohen's kappa (kappa). Model name abbreviated based on modelling platform (MaxEnt: maximum entropy; ENFA: ecological niche factor analysis) and year (08: 2008; 09: 2009). Marginality and specialization reported for ENFA models only

Model	AUC	Kappa	Marginality	Specialization
MaxEnt08	0.803	0.761	–	–
MaxEnt09	0.686	0.346	–	–
ENFA08	0.745	0.384	1.517	4.887
ENFA09	0.734	0.283	1.615	2.992

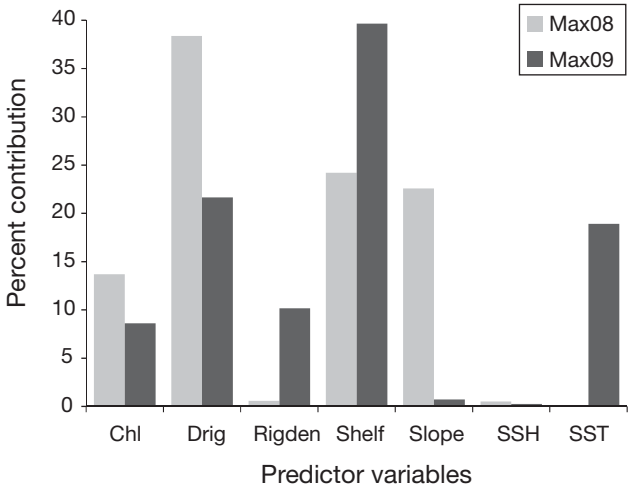

Fig. 4. Heuristic estimate of relative contributions of the environmental variables for whale sharks *Rhincodon typus* calculated from maximum entropy distribution models. Light gray (Max08): 2008 model; dark gray (Max09): 2009 model. Variable abbreviations as in Table 1

accounted for in the first factor (46% in 2008; 13% in 2009) (Table 4). Whale sharks were selecting habitats that had mean platform densities, chl concentrations, and bathymetric slopes higher than mean for the entire study area, as indicated by the positive marginality coefficients in Factor 1. Additionally, negative marginality coefficients indicate whale sharks were selecting areas with lower mean distances to petroleum platforms and the continental shelf edge, and cooler SSTs than the mean for the entire study area. Interestingly, although the means of the background and presence locations were equal for SSH, indicated by the zero marginality coefficient, the high degree of specialization accounted for in Factor 2 indicates that whale sharks were using a very restricted range of SSH values compared to the range of values over the entire study area. Three other variables exhibited a small amount of specialization (distance to platforms, distance to shelf edge and SST) in both years, while bathymetric slope only exhibited specialization in 2008.

Suitable habitat was predicted along the continental shelf edge from the western region near Texas, over to the DeSoto Canyon area in the east, with the most suitable habitat predicted south of the Mississippi River Delta (Figs. 5 & 6). The MaxEnt08 model had the smallest spatial areas predicted as highly suitable habitat (Table 5). At most, suitable habitat (areas with scores >75) was predicted in only 3% of the overall NGOM waters.

DISCUSSION

Whale shark *Rhincodon typus* distribution has been well documented in coastal areas where whale sharks aggregate to feed (Heyman et al. 2001, Wilson et al. 2001, de la Parra Venegas et al. 2011, Rowat et al. 2011); however, this is the first study to spatially quantify the areas of highest suitability and the potential biotic and abiotic drivers associated with their regional distribution. The use of presence-only modelling techniques does limit the discriminatory power of this analysis, as model predictions do not consider avoidance and can only be evaluated against random-distribution rather than against true prediction error (i.e. both false negative and false positive predictions) as is possible with a random survey. Nonetheless, presence-only modelling allows for maximum use of unique data types that are informative regarding species habitat preferences. The present study investigated variables that were thought to serve as a proxy for a potential food source and found that chl concentrations, distance to the shelf edge, and distance to platforms were most influential in predicting the distribution of whale shark feeding aggregations in the NGOM. Along the continental shelf edge, physical properties such as along-shelf currents and eddy/slope interactions can create vertical mixing that brings nutrient-rich waters to the surface (Huthnance 1981, Marra 1990, Zavala-Hidalgo et al. 2006). When deep-water nutrients are brought to the photic zone, primary production is increased, supporting the growth of the entire trophic community (Marra 1990). Even apex predators, such as tunas, sharks, and cetaceans, are attracted to these areas of high productivity along continental shelf edges (Vukovich & Maul 1985, Baumgartner 1997). Whale sharks are likely attrac-

Table 4. Correlation between the ecological niche factor analysis (ENFA) factors and the predictor variables used in the whale shark *Rhincodon typus* species distribution models for 2008 (08) and 2009 (09). Factor 1 explains 100% of the marginality. Percentages indicate amount of specialization accounted for by the factor. Symbology based on Hirzel et al. (2004). The symbol + (Factor 1) indicates *R. typus* prefers locations with higher values than the mean of the study area for that variable. The inverse is true for the symbol –. The greater the number of symbols, the higher the degree of separation from the available data. 0 indicates no difference in mean values. The symbol * (Factors 2 and 3) means that *R. typus* was found occupying a narrower range of values than available. The greater the number of symbols, the narrower the range of habitat used. 0 indicates a wide range of habitat use (very low specialization). Variable abbreviations as in Table 1

Variable	Factor 1		Factor 2		Factor 3	
	ENFA08 (46%)	ENFA09 (13%)	ENFA08 (45%)	ENFA09 (77%)	ENFA08 (4%)	ENFA09 (5%)
Chl	++	++	0	0	*	*
Drig	– – –	– – –	*	*	*	****
Rigden	+++	++++	0	0	0	0
Shelf	– – –	– – –	*	*	**	****
Slope	++	+	*	0	**	**
SSH	0	0	*****	*****	*	*
SST	– –	– –	*	*	*****	*

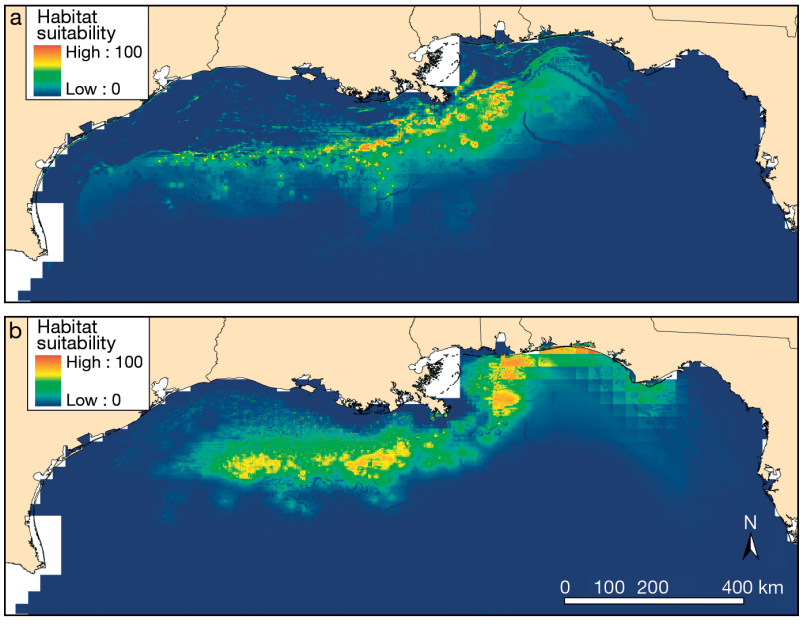

Fig. 5. Habitat suitability maps for whale shark *Rhincodon typus* maximum entropy distribution models for (a) 2008 and (b) 2009. Warmer colors indicate higher suitability

ted to these regions for the same reason and have been observed feeding on baitfish among schools of tunas (Springer 1957, Hoffman et al. 1981). In other regions, such as the southern Gulf of Mexico and Djibouti, it has been suggested that regional upwelling along steep, shelf edges may also be the most important driver influencing the presence of whale shark aggregations (de la Parra Venegas et al. 2011, Rowat et al. 2011). In Ningaloo Reef, Western Australia, whale shark abundances have been correlated with environmental variables, such as currents, water temperatures, and the Southern Oscillation Index, which has been suggested by the authors to also serve as a proxy for food availability (Taylor & Pearce 1999, Wilson et al. 2001, Sleeman et al. 2010). Whale shark distribution has been linked to nutrient-rich waters with elevated chlorophyll levels in other regions, including, but not limited to, Japan, Western Australia, India, and the Galapagos Islands (Iwasaki 1970, Compagno 1984, Arnbom & Papastavrou 1988, Taylor & Pearce 1999, Hsu et al. 2007, Kumari & Raman 2010). Similarly, chlorophyll *a* concentrations have been used to forecast fisheries catches (Solanki 2003) and to delineate migratory corridors and foraging habitat used by highly migratory marine species, such as tunas and turtles (Polovina et al. 2001). At the present time, residency and migratory patterns for whale sharks is poorly understood, but it is well documented that whale sharks are opportunistic filter-feeders that are found in highly productive areas.

There could be mechanistic and ecological explanations as to why distance to the rig was one of the influential variables identified in the whale shark distribution models. There are many structural and functional differences amongst platform types in the NGOM; however, for the present study, all were treated as equal. The significance of Drig could be an artifact of observational bias as a result of increased activity around these platforms. Since a significant amount of the offshore recreational fishing in the NGOM is associated with platforms (Franks 2000), fishermen may encounter whale sharks while en route to or while actively fishing near these platforms. Also, the petroleum industry employs 1000s of personnel on platforms, as well as in marine and air transit to and from the platforms. This increased activity around

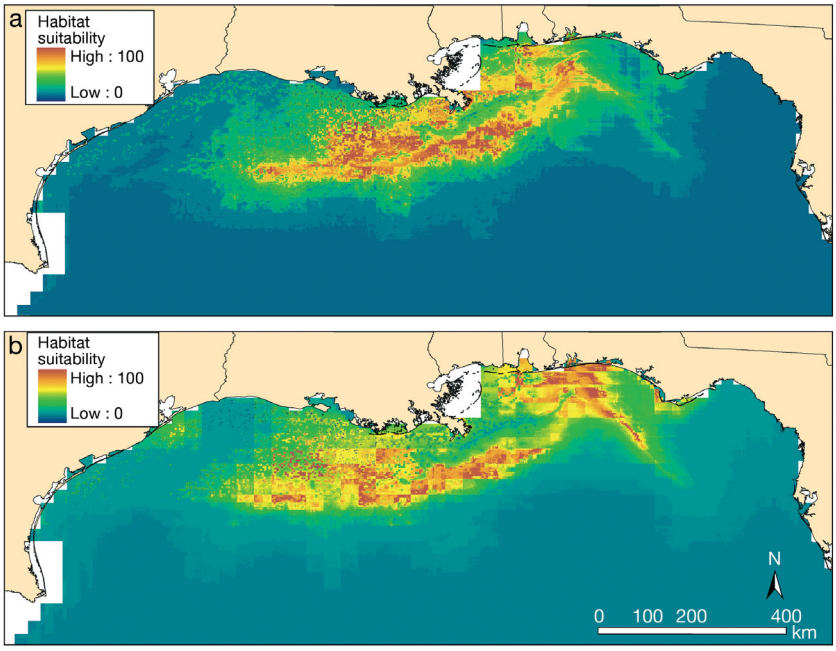

Fig. 6. Habitat suitability maps for whale shark *Rhincodon typus* ecological niche factor analysis distribution models for (a) 2008 and (b) 2009. Warmer colors indicate higher suitability

Table 5. Spatial area (km^2) with habitat suitability scores >75, 90, and 95% thresholds for whale shark *Rhincodon typus* species distribution models. Model name abbreviated based on modelling platform (MaxEnt: maximum entropy; ENFA: ecological niche factor analysis) and year (08: 2008; 09: 2009)

Model	Threshold		
	75%	90%	95%
MaxEnt08	2318	259	118
MaxEnt09	3901	442	162
ENFA08	31071	15915	10016
ENFA09	30711	12729	10056

platforms could have introduced some observational bias, which is one of the unavoidable limitations when dealing with a voluntary observational dataset rather than stratified sampling survey methods. However, based on their increased presence near these structures, there must be some ecological or behavioral association. Petroleum platforms are highly productive artificial structures that have complex food webs and attract many reef and pelagic species (Franks 2000, Stanley & Wilson 1997), including whale sharks (Hoffmayer et al. 2005). Based on the high-performance scores calculated using a dataset comprised of systematically collected presence locations, it seems as though any observational bias with platforms may only affect the variable contribution, but not the habitat suitability predictions.

Interpretation of the results of the present study has raised further questions about the association of whale sharks with platforms in the NGOM. Unfortunately, our study did not take into account multiple visits to the same location; however, there have been numerous reports of daily whale shark presence at the same platform during the same time of day for a period of up to 2 wk (WSSS unpubl. data), suggesting that at least some attraction to these platforms is occurring. Furthermore, despite the numerous platforms covering the continental shelf, whale sharks are found only near those on the shelf edge and beyond (Appendix 1, Fig. A2). The present analysis was unable to account for these potentially important observations. This type of survey provides a wealth of information, and creative modes of analysis must be explored in order to apply meaning to these observational data. The association of whale sharks with platforms, which was observed in the present study, warrants further investigation in order to elucidate whether these sharks are being attracted to such structures.

Differences in the computation of the 2 modelling algorithms used in the present study resulted in considerable differences in spatial output and the interpretation of environmental contribution. MaxEnt has been shown repeatedly to out-perform other modelling algorithms, including ENFA (Phillips et al. 2004, Elith et al. 2006, Hamel et al. 2006, Tittensor et al. 2009), which is likely due to the fact that MaxEnt can create a spatial prediction with a higher fit (as indicated by higher test scores) by implementing different feature classes into more complex functions (i.e. linear, quadratic, product, threshold, hinge, and category indicator) (Phillips & Dudík 2008). The spatial prediction of ENFA may not be as responsive as MaxEnt to the variable relationships because it can only fit linear relationships, unless nonlinear combinations of variables are included as a unique layer (Hirzel et al. 2002). This typically results in higher AUC scores in MaxEnt models when compared to ENFA and less area predicted as highly suitable (Sérgio et al. 2007, Benito et al. 2009, Tittensor et al. 2009, Braunisch & Suchant 2010). One potential problem raised with the

maximum entropy function is its tendency towards over-prediction; MaxEnt software employs a regularization process to avoid over-fitting (Warren & Seifert 2011). Additionally, reducing the complexity of the model and removing correlated variables reduces the tendency towards over-fitting (Elith et al. 2011). In the present study, the smaller spatial area predicted as suitable habitat in the MaxEnt models is likely a result of the differences in computation, rather than due to over-prediction because of the preliminary steps to determine the most parsimonious model.

One should proceed with caution when making comparisons between algorithms, because the environmental variable contribution from each prediction is interpreted differently. However, considering both in conjunction can bolster ecological interpretation of the model results. One advantage of the ENFA approach is that the variable scores are more meaningful and straightforward to interpret compared to MaxEnt, specifically by providing a metric to compare the habitat used by the species in relation to the available habitat (Hirzel et al. 2002, Tittensor et al. 2009). The interpretation of variable contribution in MaxEnt models can be difficult due to the machine-learning technique used and the way the model automatically relates different feature classes (i.e. linear, quadratic, polynomial). The influence of collinearity and interactions between variables is not explicitly reported and can be hard to interpret (Phillips & Dudík 2008). In the present study, preliminary steps were taken to remove collinearity so we could better interpret variable contribution; however, interaction effects may still be present and impossible to discern. Studies that report the variable contribution from both algorithms have found that the predominant variables remained the same in both approaches (Tittensor et al. 2009, Braunisch & Suchant 2010), which was also observed in the present study.

In August 2009, the WSSS received nearly 60 reports (including 13 aggregations) of whale sharks within 10 nautical miles (~18.5 km) of the coast from Mobile Bay, Alabama, to Panama City, Florida. It is possible that the lower test scores directly resulted from this increased inshore activity, which was not observed in the historical dataset used to test the model. Although these inshore sightings were uncharacteristic of historical records, past aerial surveys, and 8 yr of sightings data (Gudger 1939, Hoffmayer et al. 2005, Burks et al. 2006), it is clear that whale sharks were drawn to this region to feed. During scientific encounters at the time, whale sharks were observed ram filter-feeding in surface waters (J. McKinney pers. obs.), and based on the abundance of moon jellies *Aurelia aurita*, a planktivorous competitor, there was probably an abundant planktonic prey source attracting these and other filter-feeders to the region. Plankton samples taken during the event were dominated by a planktonic shrimp species, *Lucifer faxoni*, with a minimal presence of unidentified fish eggs. Since 2009, only 4 whale shark sightings have been reported from this region, during 2010 and 2011, further supporting our claim that something unusual occurred during 2009 that impacted whale shark distribution in the NGOM. Additional years of data will be needed to obtain a more complete understanding of whale shark distribution in the NGOM. Since the WSSS is an on-going project that continually collects sightings data, the methods presented in the current study will be used in the future, in order to delineate longer term trends in distribution, better understand inter-annual variability, and identify consistent areas of high habitat suitability for whale sharks in the NGOM.

The NGOM contains habitat for whale shark feeding aggregations during the summer months, primarily along the highly productive continental shelf edge region. Observed presence in this region is highly seasonal, and true habitat use likely encompasses a much larger area of the entire water column not observable from the surface. Although whale sharks are a vulnerable species and knowledge of suitable habitat is critical for effective management regimes, due to their behavior, the necessary data are difficult to collect. Since whale sharks have a circumglobal distribution, the variables found to influence their distribution in this regional study need to be applied to whale shark populations elsewhere to determine the utility of developing a global habitat suitability index. The results of the present study demonstrate the value of presence-only modelling as a tool for studying whale shark distribution, because it integrates limited, yet readily available, information in a systematic way and produces meaningful predictions. Our study also demonstrates the benefits of combining multiple modelling approaches when studying a species' ecological preference, as each methodology has different advantages and disadvantages. Modelling is an iterative process, and further examination may elucidate other ecological drivers affecting whale shark distribution that were not included in the present study. The marine environment is very dynamic, and, as a result, suitable whale shark habitat is dynamic as well. It is recommended that ensemble predictions from multiple models, investigated at multiple spatial and temporal scales, be examined to gain further insight into whale shark ecology in the region.

Acknowledgements. The present study would not have been possible without the many submitters to the GCRL whale shark sightings survey, who are too numerous to name; as well as the many people and organizations that helped spread the word about the survey. Specifically, we thank Dan Allen and Allen Veret of the Offshore Operators Committee, and Mark Fontenot of the Helicopter Safety Advisory Conference. We also acknowledge James G. Acker of NASA GES-DISC for his help in navigating through the NASA data visualizers and accessing the most up-to-date data products; Jason Roberts of the MGET Center at Duke University, for his personal assistance with the MGET toolkit; and, lastly, the mission scientists, principal investigators, and associated NASA personnel for the production of the data used in this research effort. We also thank 4 anonymous reviewers for their contributions to the development of the manuscript. William Driggers III, Mark Peterson, and James Franks are also acknowledged for many fruitful discussions on the subject matter.

LITERATURE CITED

Allouche O, Tsoar A, Kadmon R (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J Appl Ecol 43: 1223–1232

Arnbom T, Papastavrou V (1988) Fish in association with the whale sharks *Rhiniodon typus* near the Galapagos Islands. Noticias de Galapagos 46:13–15

Balch WM, Byrne CF (1994) Factors affecting the estimate of primary production from space. J Geophys Res 99: 7555–7570

Baumgartner MF (1997) The distribution of Risso's dolphin (*Grampus griseus*) with respect to the physiography of the northern Gulf of Mexico. Mar Mamm Sci 13: 614–638

Benito B, Martínez-Ortega M, Muñoz L, Lorite J, Peñas J (2009) Assessing extinction-risk of endangered plants using species distribution models: a case study of habitat depletion caused by the spread of greenhouses. Biodivers Conserv 18:2509–2520

Braunisch V, Suchant R (2010) Predicting species distributions based on incomplete survey data: the trade-off between precision and scale. Ecography 33:826–840

Burks CM, Driggers WB III, Mullin KD (2006) Abundance and distribution of whale sharks (*Rhincodon typus*) in the northern Gulf of Mexico. Fish Bull 104:579–584

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46

Colman JG (1997) A review of the biology and ecology of the whale shark. J Fish Biol 51:1219–1234

Compagno LJV (1984) FAO species catalogue. FAO, Rome

de la Parra Venegas R, Hueter R, González Cano J, Tyminski J and others (2011) An unprecedented aggregation of whale sharks, *Rhincodon typus*, in Mexican coastal waters of the Caribbean Sea. PLoS ONE 6:e18994

Druon JN, Fromentin JM, Aulanier F, Heikkonen J (2011) Potential feeding and spawning habitats of Atlantic bluefin tuna in the Mediterranean Sea. Mar Ecol Prog Ser 439:223–240

Elith J, Graham CH, Anderson RP, Dudík M and others (2006) Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129–151

Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ (2011) A statistical explanation of MaxEnt for ecologists Drivers Distrib 17:43–57

Franks JS (2000) Pelagic fishes at offshore petroleum platforms in the northern Gulf of Mexico: diversity, interrelationships, and perspective. Colloque Caraibe Actes de Colloques Ifremer Aquat Living Resour (France) 13: 502–515

Gudger EW (1939) The whale shark in the Caribbean Sea and the Gulf of Mexico. Sci Mon 48:261–264

Guisan A, Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. Ecol Lett 8: 993–1009

Hamel P, Barker S, Benítez S, Baldy J and others (2006) Modeling the South American range of the cerulean warbler. In: Proceedings of the 26th ESRI international user conference. ESRI, San Diego, CA

Heyman WD, Graham RT, Kjerfve B, Johannes RE (2001) Whale sharks *Rhincodon typus* aggregate to feed on fish spawn in Belize. Mar Ecol Prog Ser 215:275–282

Hirzel AH, Hausser J, Chessel D, Perrin N (2002) Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? Ecology 83:2027–2036

Hirzel AH, Posse B, Oggier PA, Crettenand Y, Glenz C, Arlettaz R (2004) Ecological requirements of reintroduced species and the implications for release policy: the case of the bearded vulture. J Appl Ecol 41:1103–1116

Hirzel AH, Hausser J, Perrin N (2007) Biomapper 4.0. Laboratory of Conservation Biology, Department of Ecology and Evolution, University of Lausanne, Lausanne

Hoffman W, Fritts T, Reynolds R (1981) Whale sharks associated with fish schools off south Texas. Northeast Gulf Sci 5:55–57

Hoffmayer ER, Franks JS, Shelley JP (2005) Recent observations of the whale shark (*Rhincodon typus*) in north central Gulf of Mexico. Gulf Caribb Res 17:117–120

Hsu HH, Joung SJ, Liao YY, Liu KM (2007) Satellite tracking of juvenile whale sharks, *Rhincodon typus*, in the northwestern Pacific. Fish Res 84:25–31

Huthnance JM (1981) Waves and currents near the continental shelf edge. Prog Oceanogr 10:193–226

Iwasaki Y (1970) On the distribution and environment of the whale shark, *Rhincodon typus*, in skipjack fishing grounds in the western Pacific Ocean. J Mar Sci Technol Tokai Univ 4:37–51

Kumari B, Raman M (2010) Whale shark habitat assessments in the northeastern Arabian Sea using satellite remote sensing. Int J Remote Sens 31:379–389

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174

Marra JH, Houghton RW, Garside C (1990) Phytoplankton growth at the shelf-break front in the Middle Atlantic Bight. J Mar Res 48:851–868

McPherson JM, Jetz W, Rogers DJ (2004) The effects of species' range sizes on the accuracy of distribution models: Ecological phenomenon or statistical artefact? J Appl Ecol 41:811–823

Pearce JL, Boyce MS (2006) Modelling distribution and abundance with presence-only data. J Appl Ecol 43: 405–412

Peterson AT (2001) Predicting species' geographic distributions based on ecological niche modeling. Condor 103: 599–605

Phillips SJ, Dudík M (2008) Modeling of species distributions with MaxEnt: new extensions and a comprehensive evaluation. Ecography 31:161–175

Phillips SJ, Dudik M, Schapire RE (2004) A maximum entropy approach to species distribution modeling. In: Proceedings of the 21st international conference on machine learning. ACM, Banff, Alberta

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. Ecol Model 190:231–259

Polovina JJ, Howell E, Kobayashi DR, Seki MP (2001) The transition zone chlorophyll front, a dynamic global feature defining migration and forage habitat for marine resources. Prog Oceanogr 49:469–483

Praca E, Gannier A, Das K, Laran S (2009) Modelling the habitat suitability of cetaceans: example of the sperm whale in the northwestern Mediterranean Sea. Deep-Sea Res I 56:648–657

Roberts JJ, Best BD, Dunn DC, Treml EA, Halpin PN (2010) Marine Geospatial Ecology Tools: an integrated framework for ecological geoprocessing with ArcGIS, Python, R, MATLAB, and C++. Environ Model Softw 25: 1197–1207

Rowat D, Brooks K, March A, McCarten C and others (2011) Long-term membership of whale sharks (*Rhincodon typus*) in coastal aggregations in Seychelles and Djibouti. Mar Freshw Res 62:621–627

Sequeira A, Mellin C, Rowat D, Meekan MG, Bradshaw CJA (2012) Ocean-scale prediction of whale shark distribution. Divers Distrib 18:504–518

Sérgio C, Figueira R, Draper D, Menezes R, Sousa AJ (2007) Modelling bryophyte distribution based on ecological information for extent of occurrence assessment. Biol Conserv 135:341–351

Sleeman JC, Meekan MG, Fitzpatrick BJ, Steinberg CR, Ancel R, Bradshaw CJA (2010) Oceanographic and atmospheric phenomena influence the abundance of whale sharks at Ningaloo Reef, Western Australia. J Exp Mar Biol Ecol 382:77–81

Sneath PHA, Sokal RR (1973) Numerical taxonomy. Freeman, San Francisco, CA

Solanki HU, Dwivedi RM, Nayak SR, Somvanshi VS, Gulati DK, Pattnayak SK (2003) Fishery forecast using OCM chlorophyll concentration and AVHRR SST: validation results off Gujarat Coast, India. Int J Remote Sens 24: 3691–3699

Springer S (1957) Some observations of the behavior of schools of fishes in the Gulf of Mexico and adjacent waters. Ecology 38:166–171

Stanley DR, Wilson CA (1997) Seasonal and spatial variation in abundance and size distribution of fishes associated with a petroleum platform in the northern Gulf of Mexico. Can J Fish Aquat Sci 54:1166–1176

Stevens JD (2007) Whale shark (*Rhincodon typus*) biology and ecology: a preview of the primary literature. Fish Res 84:4–9

Stockwell DRB, Peterson AT (2002) Effects of sample size on accuracy of species distribution models. Ecol Model 148: 1–13

Taylor JG (2007) Ram filter-feeding and nocturnal feeding of whale sharks (*Rhincodon typus*) at Ningaloo Reef, Western Australia. Fish Res 84:65–70

Taylor JG, Pearce AF (1999) Ningaloo reef currents: implications for coral spawn dispersal, zooplankton and whale shark abundance. J R Soc West Aust 82:57–65

Tittensor DP, Baco AR, Brewin PE, Clark MR and others (2009) Predicting global habitat suitability for stony corals on seamounts. J Biogeogr 36:1111–1128

Tserpes G, Peristeraki P, Valavanis VD (2008) Distribution of swordfish in the eastern Mediterranean, in relation to environmental factors and the species biology. Hydrobiologia 612:241–250

Vukovich FM, Maul GA (1985) Cyclonic eddies in the eastern Gulf of Mexico. J Phys Oceanogr 15:105–117

Warren DL, Seifert SN (2011) Environmental niche modeling in MaxEnt: the importance of model complexity and the performance of model selection criteria. Ecol Appl 21:335–342

Wiley EO, McNyset KM, Peterson AT, Robins CR, Stewart AM (2003) Niche modeling and geographic range predictions in the marine environment using a machine-learning algorithm. Oceanography (Wash DC) 16: 120–127

Wilson SG, Taylor JG, Pearce AF (2001) The seasonal aggregation of whale sharks at Ningaloo Reef, Western Australia: currents, migrations and the El Niño/Southern Oscillation. Environ Biol Fishes 61:1–11

Zavala-Hidalgo J, Gallegos-García A, Martínez-López B, Morey S, O'Brien J (2006) Seasonal upwelling on the western and southern shelves of the Gulf of Mexico. Ocean Dyn 56:333–338

Appendix 1. Additional data

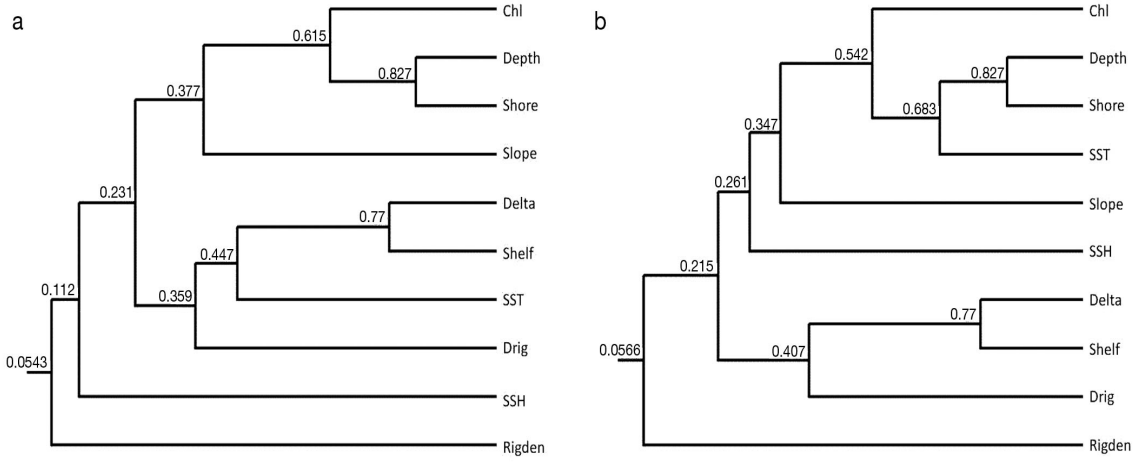

Fig. A1. UPGMA correlation tree for (a) 2008 (b) environmental variables used in the initial model exploration of whale shark *Rhincodon typus* distribution models. Variables with a high degree of multi-collinearity (Shore, Depth and Delta) were not included in the final model presented in this paper. Variable abbreviations as in Table 1

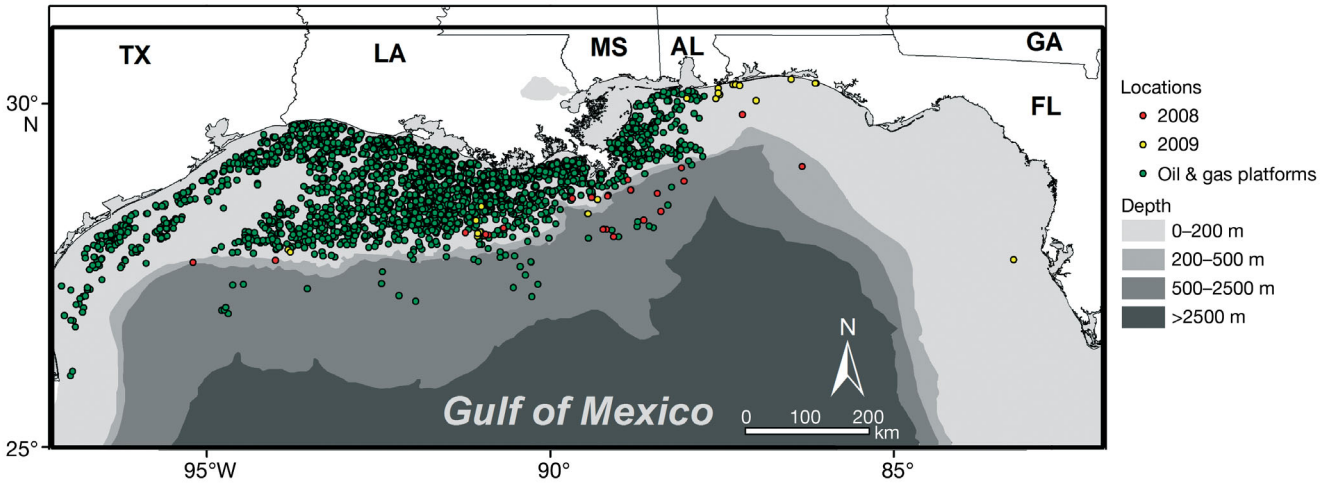

Fig. A2. In the northern Gulf of Mexico, petroleum platforms (green dots) blanket the continental shelf from Texas to the Florida state line. Whale shark *Rhincodon typus* aggregation data used to build species distribution models has been included; 2008 locations indicated by red dots, 2009, by yellow dots

Editorial responsibility: Konstantinos Stergiou, Thessaloniki, Greece

Submitted: September 20, 2011; Accepted: April 21, 2012
Proofs received from author(s): June 20, 2012