Predicting Pelagic Habitat with Presence-only Data using Maximum Entropy
for Olive Ridley Sea Turtles in the Eastern Tropical Pacific

by

Lindsey Peavey
Dr. Larry Crowder, Advisor

May 2010

Masters project submitted in partial fulfillment of the requirements for the Master of Environmental Management degree in the Nicholas School of the Environment of Duke University

Approved:

Dr. Larry Crowder, Advisor
# Table of Contents

List of Figures ......................................................................................................................................................................... 4  
List of Tables ........................................................................................................................................................................... 7  
Abstract .................................................................................................................................................................................. 8  
Introduction ............................................................................................................................................................................... 9  
Olive ridley sea turtles in the eastern tropical Pacific (ETP) .......................................................................................... 9  
Oceanic habitat modeling ..................................................................................................................................................... 15  
Methods ................................................................................................................................................................................... 17  
Study area ............................................................................................................................................................................... 17  
Species data collection ......................................................................................................................................................... 18  
Environmental predictors ..................................................................................................................................................... 20  
Correcting for sample bias ............................................................................................................................................... 24  
Habitat modeling ................................................................................................................................................................. 26  
Results .................................................................................................................................................................................... 29  
GLM .................................................................................................................................................................................... 29  
GAM .................................................................................................................................................................................... 34  
Maxent .................................................................................................................................................................................. 38  
Model evaluation ................................................................................................................................................................. 42  
Discussion ................................................................................................................................................................................. 45  
Picking the best model ......................................................................................................................................................... 45  
Potential role of ETP oceanographic features .................................................................................................................. 48  
Thoughts about Maxent – a viable presence-only oceanic habitat modeling tool? ....................................................... 50  
Future Research .................................................................................................................................................................... 52  
Appendix ................................................................................................................................................................................. 58  
MGET GLM and GAM GIS model builder .......................................................................................................................... 58  
GLM model fit summary statistics ........................................................................................................................................ 59  
GLM model summary ........................................................................................................................................................... 60  
GAM model fit summary statistics ........................................................................................................................................ 60  
GAM model test summary statistics ..................................................................................................................................... 61  
GAM model summary ........................................................................................................................................................... 62  
Maxent data preparation GIS script tool ............................................................................................................................. 62  
Maxent open-source GIS data analysis tool ........................................................................................................................ 63  
Screenshots of the Maxent java interface and the user-specified constraints exactly as the model was run for this analysis .................................................................................................................................................. 66
List of Figures

Figure 1. Olive ridley sea turtles nesting en masse during an “arribada” on Playa Ostional, Costa Rica on 09 September 2004. Photo by Michael Jensen. ............................................................ 12

Figure 2. Plot of 4850 ETP turtle sightings during line transect surveys during STAR ecosystem assessment survey cruises from years 1992, 1998, 1999, 2000, 2003 and 2006; all during the time period of July - December (Figure by Gerrodette, et al. unpub). ............................................. 14

Figure 3. Map of the eastern tropical Pacific (ETP) study area used in this investigation, as designated by the ongoing National Oceanic and Atmospheric Administration (NOAA) Stenella Abundance Research (STAR) ecosystem assessment cruises.................................................. 18

Figure 4. Histogram of all STAR turtle captures years 1989-2006 (n = 1159). Size classes based on straight carapace length (SCL) are indicated. ............................................................................. 19

Figure 5. Seasonal average maps of the four environmental predictors considered in this analysis, from top left to bottom right: chlorophyll(a) (Chl), sea surface temperature (SST), Depth and Distance. ..................................................................................................................................... 22

Figure 6. Scatterplots of the data points for each environmental predictor are displayed in the bottom left, including red trend lines to show general relationships. Frequency histograms (solid blue bars) plotting the distribution of each predictor are shown diagonally in the middle. Pearson correlation coefficients describing the relationship between two predictors are displayed in the upper right; text is sized by magnitude of correlation (larger, more correlated; smaller, less correlated). Note the correlations between ‘Distance’ to shore and ‘Depth’ (0.45), and ‘Distance’ to shore and sea surface temperature (‘SST’) (0.59). This plot was created using MGET tools in ArcGIS and the R Project for Statistical Computing. ............................................. 23

Figure 7. Pseudo-absence points were randomly generated within a convex hull encompassing olive ridley presence points. These pseudo-absence points were used to sample the underlying environmental data, and then utilized as the ‘absence’ data in the GLM and GAM, and the “background” file for the Maxent model (see Appendix for details). .................................................. 26

Figure 8. Simplified schematic of the methodology used in this study. .................................................. 28

Figure 9. GLM terms plots for the three predictors: chlorophyll(a), sea surface temperature, and depth. The solid line is the mean fit; dashed lines represent 2 standard error deviations. Data points are displayed as open circles and presence density for the predictor is shown along the bottom of the figure with the densest areas displayed in continuous black. .............................. 31
Figure 10. (a) The resultant GLM predictive habitat classification map, including the training points plotted in white; and (b) the standard error associated with the predictive map.

Figure 11. GAM terms plots for the three predictors: chlorophyll(a), sea surface temperature, and depth. The solid line is the mean fit; dashed lines represent 2 standard errors deviation from the mean. Data points are displayed as open circles and presence density for the predictor is shown along the bottom of the figure with the densest areas displayed in continuous black.

Figure 12. (a) The resultant GAM predictive habitat classification map, including the training points plotted in white; and (b) the standard error associated with the predictive map.

Figure 13. Maxent response curves for the three predictors: chlorophyll(a), sea surface temperature, and depth.

Figure 14. The same Maxent predicted habitat map is shown twice here with the training points from 2006 plotted in the upper panel and the historical testing points plotted in the lower panel.

Figure 15. GLM ROC curve indicating the cutoff value is 0.532561, which can be used to create a binary habitat map.

Figure 16. GAM ROC curve indicating the cutoff value is 0.566241, which can be used to create a binary habitat map.

Figure 17. Maxent plot of true positive rate, known as “sensitivity”, vs. false positive rate, or (1 – “specificity”). Specificity is the true negative rate. Although Maxent is not logistic regression, it does offer an option to balance sensitivity and specificity, which acts as the decision threshold similar to a cutoff value in GLM and GAM.

Figure 18. Scale for comparing area under the ROC curve values, known as “AUC.”

Figure 19. Schematic diagram of surface water masses and currents in the eastern tropical Pacific. STSW, Subtropical Surface Water; TSW, Tropical Surface Water; ESW, Equatorial Surface Water. Shading represents mean sea surface temperature (darker = colder; Shea et al. 1992) (Figure from Fiedler and Talley 2006).

Figure 20. Adult female olive ridley sea turtle (SCL: 53.3 cm) tracks overlaid on a static image of the surface currents on 2006 Julian Day 333 (29 November). The satellite transmitter (Wildlife Computers SDR T16) was deployed on 2006 November 05 and persisted for 166 days. The light red circle highlights the section of turtle tracks (late November - early December) that generally corresponds with the date of the static surface currents, indicating that the turtle is moving along the edges of the eddy.
Figure 21. Olive ridley sea turtle captures from STAR 2006 mapped based on straight carapace length as noted in the literature: 56-79cm “Adults,” 49-56cm “Sub-Adults,” 6-49cm “Juveniles,” <5cm “Hatchlings.”
List of Tables

Table 1. Population changes for the Olive Ridley Index Sites for the ETP region (Abreu-Grobois and Plotkin 2008). ................................................................................................................................. 13

Table 2. Source, resolution and scale of the remotely-sensed environmental data used in analysis. 21

Table 3. Maxent model parameters used in this study.  See Appendix for screenshots of Java software inputs.  See the Maxent Tutorial (available at http://www.cs.princeton.edu/~schapire/maxent/) or Phillips et al. (2006, 2008 & 2009) for complete definitions and explanations of each parameter.......................................................... 27

Table 4. Area under the ROC curve (AUC) values and AUC percent change from training to testing for the GLM, GAM and Maxent habitat models.  See Appendix for full model statistics for each............................................................................................................................................ 45
Abstract

Little is known about the oceanic distribution of olive ridley sea turtles (*Lepidochelys olivacea*) in the eastern tropical Pacific (ETP), or what governs their offshore movements. In collaboration with NOAA’s Southwest Fisheries Science Center, I opportunistically sampled 350 olive ridley sea turtles in the ETP between August and December 2006. Using these presence-only observations and remotely-sensed oceanographic data, I developed a maximum entropy habitat model using the Maxent software package, and considered the influences of chlorophyll(α) concentrations, sea surface temperature, and bathymetry. I compared Maxent results with two more common approaches used to describe pelagic species distribution and habitat suitability: generalized linear and additive models (GLM and GAM). Statistically, the GLM model performed well (AUC > 0.78), whereas the GAM (highly variable AUC: 0.64 – 0.88) and Maxent (AUC < 0.68) models did not. However, based on expert knowledge the Maxent results were the most reasonable despite low AUC values. For the scope of this study, Maxent was determined to produce a viable species distribution model, although areas of improvement are recommended. Maxent is growing in popularity among marine habitat modelers, but I caution against using the method as a panacea for predictive habitat modeling for open-ocean, rare species, despite its accessible and easy-to-use software. Similar studies should be repeated for other oceanic species (e.g. migratory marine megafauna) and compared with richer datasets (e.g. line transect surveys and telemetry) to gain a better understanding of Maxent’s ability to accurately predict oceanic habitat.
Introduction

Olive ridley sea turtles in the eastern tropical Pacific (ETP)

Pacific olive ridley sea turtle (*Lepidochelys olivacea*) populations are decreasing, and the species is listed as ‘vulnerable’ on the IUCN Red List of Threatened Species (IUCN 2009). However, relative to other, more severely threatened sea turtle populations, olive ridleys are currently the most abundant sea turtle in the world (Plotkin 2007, IUCN 2009).

Eastern tropical Pacific (ETP) olive ridley population abundance was recently estimated as high as 1.39 million (Eguchi 2007). Nevertheless, this is at least an order of magnitude lower than their historical abundance (Abreu-Grobois and Plotkin 2008). In the 1960’s and 1970’s, this species experienced drastic population declines due to the over-exploitation of their meat and eggs (Green & Ortiz-Crespo 1995; Peñaflores et al. 2000; Abreu-Grobois and Plotkin 2008), as well as substantial incidental mortality from commercial fishing gear interactions (Cheng & Chen 1997; Pandav et al. 1998; Lewison et al. 2004; Pindeo & Polacheck 2004). These intense and consistent threats caused this population to collapse to near extinction (Márquez et al. 1997; López-Castro 2005). In 1982, olive ridleys were listed as ‘Endangered’ on the IUCN World Conservation Union Red List, and included in the Convention on International Trade in Endangered Species (CITES) (IUCN 2009). National (United States, USNMFS and USFWS 1998; Mexico, DOF 1990) and international (Convention on Migratory Species; Inter-American Convention for the Protection and Conservation of Sea Turtles) conservation plans were developed and implemented for olive ridleys in many of the countries where adult females return each year to nest (Abreu-Grobois and Plotkin 2008).

As is generally the case with all sea turtle populations, the urgency to save this population from extinction far overshadowed the need to study and learn about its ecology (Prichard 1997;...
Priority and resources were given to the monitoring and protection of nesting beaches, rather than costly and time-consuming ecological research. These efforts seem moderately successful: in 2007, the Pacific olive ridley was demoted to ‘vulnerable’ status on the Red List.

Despite this success, there is still much work to be done not only to learn more about the animal, but also to protect its ecological niche. There are two conservation concerns specific to olive ridleys that will affect the longevity of this realized success. (1) Despite being largely regarded as a ‘conservation success story’, the eastern tropical Pacific (ETP) population is still declining at a rate of 33-39% nesting females over ~40 years (Abreu-Grobois and Plotkin 2008). There are sizeable differences in the population dynamics of the many subpopulations in the region that have yet to be investigated. This instability warrants the attention of the scientific community, but research and conservation efforts for olive ridleys are sub par relative to the efforts for other, more critically endangered sea turtle species. Generally, funding available for the study and conservation of olive ridley sea turtles is nominal relative to others because comparatively, they are not considered a ‘sea turtle conservation priority’ (Plotkin 2007). (2) Subsequently, little research has been published on this species compared to others such as loggerhead (*Caretta caretta*), green (*Chelonia mydas*), and Kemp’s ridley (*Lepidochelys kempii*) sea turtles (Plotkin 2007). As a result, decision-makers are implementing conservation strategies and policies without having critical baseline knowledge of the species or knowing what role it plays in their marine and terrestrial ecosystems.

In general, we know very little about the ecology of ETP olive ridleys—a wide-ranging, pan-tropical species. Population dynamics, life history, and habitat selection of the ETP olive ridley are poorly understood. These knowledge gaps present serious obstacles for species managers who need to know where the animals are, how they select and use habitat, and what dictates their movement patterns in order to effectively protect the species spatially. Historically, rough
population estimates were calculated from counting nesting females and emerging hatchlings—methods that do not account for male counterparts or behavioral variations between coastal and oceanic habitats (Plotkin 1995). Rookery data are collected mostly by nesting beach protection programs in developing countries that are fueled by volunteers and subject to limited funding. Unreliable and/or sparse data has made it nearly impossible to determine accurate population trends for this species (Eguchi 2007).

Since first being described in the 1940’s, the basic morphology and physiology of olive ridleys are well understood, and their feeding ecology has been roughly characterized. Hatchlings measure 38-44mm in straight carapace length (SCL), and adults typically grow to 65-70 cm. It is generally believed that olive ridleys are open-water epipelagic opportunistic foragers, feeding mainly on tunicates, crustaceans and Cnidaria (NMFS & USFWS 1998). However, the following specifics about their life history are practically unknown: ecological roles in nearshore and open ocean ecosystems, population dynamics including connectivity, habitat uses and preferences, their distribution parameters, foraging behaviors, extinction thresholds, and to some extent their mating and nesting patterns. This is especially true of the oceanic ecosystems and habitats utilized by olive ridleys during internesting periods (Plotkin 1995).

Much more is known about their nesting behavior. Genetic studies of ETP olive ridleys have described this large population as a single pan-mictic population (all individuals of the opposite sex are potential partners) ranging from Costa Rica to Mexico (NMFS & USFWS 1998). Recently, solitary nesting has also been documented as far south as Ecuador (Alava 2007) and as far north as Baja California, Mexico, where there has been a distinct colonization event (López-Castro 2005). Nesting in Ecuador has yet to be studied. Based on mtDNA sequencing, the Baja California nesting cohort has been identified as genetically distinct from turtles nesting on continental (Mexico and Central American) beaches. This recent finding suggests that the observed phenotypic variation
may be associated with genetic differentiation and reproductive isolation, and therefore challenges the assumption that a single genetic and conservation unit of olive ridleys exists in the ETP (López-Castro 2005). More research is needed to clarify olive ridley distributions, geographic ranges, genetic variability, and connectivity in the ETP (NMFS & USFWS 1998).

Striking localized differences in population trends based on nest counts are documented in the literature, and these differences indicate far lower survival probabilities in some rookeries than in others, and certainly compared to what regional results suggest (down 33-39%, Table 1). Specifically for olive ridleys, the occurrence of *arribadas*, a collective nesting phenomenon, dramatizes this point. Sea turtles have appeared in the fossil record for over 110 million years (Hirayama 1998). Despite this, researchers only discovered their synchronized nesting behavior, coined with the Spanish word *arribada* (“arrival”), in the 1940’s (Pritchard 2007). Arribada nesting behavior is unique to the genus *Lepidochelys*, which includes two extant species: *L. olivacea* and *L. kempii*. Arribadas usually occur 3-8 consecutive nights at a time, approximately once a month during the peak nesting season, August - October (Hughes and Richard 1974; Cornelius 1986; Dash and Kar 1990; Plotkin 1995; NMFS and USFWS 1998). During these events, 100’s to 10,000’s of female turtles visit the same, usually relatively small, nesting beach to lay eggs *en masse* (Figure 1).

![Figure 1. Olive ridley sea turtles nesting *en masse* during an “arribada” on Playa Ostional, Costa Rica on 09 September 2004. Photo by Michael Jensen.](image)
turtle species, olive ridleys also nest solitarily year-round on dozens of beaches throughout the world’s tropical and subtropical oceans. However, arribadas are currently known to occur at only four major rookeries: La Escobilla beach in Oaxaca, Mexico; Playa Nancite in Costa Rica; Playa Ostional in Costa Rica; and in Orissa, India (Pritchard 2007; Abreu-Grobois and Plotkin 2008). Arribadas present a challenge for managers because the current IUCN methods for determining global conservation status introduces sample bias resulting in a regional status that underestimates the true decline levels for non-arribada rookeries (IUCN 2009; Peavey unpub.). Specifically for ETP olive ridleys, the total regional population decline (33-39%) is dominated by the arribada rookeries’ status (down 32-35%), when in reality, non-arribada rookeries are experiencing much steeper declines of 73-95% (Table 1).

Table 1. Population changes for the Olive Ridley Index Sites for the ETP region (Abreu-Grobois and Plotkin 2008).

<table>
<thead>
<tr>
<th>Region</th>
<th>Past Annual Nesting Female Subpopulation Size</th>
<th>Present Annual Nesting Female Subpopulation Size (2005)</th>
<th>% Change</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EASTERN PACIFIC OCEAN</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arribada rookeries - Playa de Mismaloya, Jalisco, México; Ixtapa, Michoacán, México; Piedra de Teléguenque, Guerrero, México; Chacahua, Oaxaca, México; Escobilla, Oaxaca, México; La Flor, Nicaragua; Nanito, Costa Rica; Ostional, Costa Rica; Isla Gáfas, Panama</td>
<td>1,184,911</td>
<td>765,249</td>
<td>-32% to -35%</td>
</tr>
<tr>
<td>Non-arribada rookeries - El Verde, Sinaloa, México; Platániche, Nayant, México; Cuyúllán, Colima, México; Marata-Cojo, Michoacán, México; Pto Arista, Chiapas, México; Barra de Chapatón - Monterrico, Guatemala</td>
<td>1,117,732</td>
<td>757,081</td>
<td>-95% to -73%</td>
</tr>
<tr>
<td>TOTALS</td>
<td>2,262,643</td>
<td>1,522,329</td>
<td>-39% to -33%</td>
</tr>
</tbody>
</table>

Because mating, foraging, and migratory movements of olive ridleys in the ETP have been primarily studied near shore, this species was originally thought to undergo short migrations and primarily feed in coastal regions. In the last decade however, satellite telemetry has afforded researchers the tools to track the movements of large juveniles and adults on the scale of hundreds of days. Dozens of post-nesting females have been tracked, and it is now well documented that these females can travel offshore thousands of kilometers to presumed epipelagic foraging habitats.
Polovina et al. (2004) studied the foraging and migration patterns of Pacific loggerheads and olive ridleys using satellite telemetry. They concluded that ten tracked olive ridleys (three western and seven eastern Pacific origin individuals) were found primarily in a region bound by 8-31° North, in the center of the Subtropical Gyre, occupying warm water with sea surface temperatures (SST) of 23-28°C. Transect data from the Southwest Fisheries Science Center show a larger range of 30° North to 15° South (Figure 2). Polovina and his colleagues documented that olive ridleys spent approximately 20% of their time at the surface, and 60% of their time in <40m water, but can dive 150m (perhaps deeper), and do so on a daily basis. All of Polovina’s tracked olive ridleys inhabited stratified surface waters with fairly uniform surface chlorophyll of about 0.1 mg m⁻³, and were not associated with any strong surface chlorophyll gradients. This report highlights the fact that olive ridleys can range over several degrees of latitude because wide SST isotherms in this part of the Pacific maintain a relatively constant temperature of ~27°C (Fiedler and Talley 2006).

Figure 2. Plot of 4850 ETP turtle sightings during line transect surveys during STAR ecosystem assessment survey cruises from years 1992, 1998, 1999, 2000, 2003 and 2006; all during the time period of July - December (Figure by Gerrodette, et al. unpub).
Most information about olive ridley mortality at sea comes from fisheries observers positioned on commercial fishing vessels. It is estimated that tens of thousands of olive ridley sea turtles die in commercial fishing gear every year (Arenas & Hall 1992; Wallace et al. 2010). Distribution estimates of olive ridleys have been made based on incidental capture data from bycatch observations on large-scale purse seine fishing vessels in the ETP (Roberts 2006). These estimates do not include or consider turtles that die in small-scale and/or artisanal fishing gear, which has been shown to be quite significant in other species such as the transoceanic Pacific loggerhead (Peckham et al. 2007). Eguchi et al. (2007) published the only oceanic, fisheries-independent population abundance estimates which used ship-based visual survey transect data. Clearly there is a need for continued offshore olive ridley research in the ETP.

Oceanic habitat modeling

Understanding of oceanic species distribution is critical to managing species on the high seas. It is costly and resource intensive to collect adequate data about species occurrence and the surrounding conditions in the open ocean, and therefore most data comes from opportunistic sampling at sea via research cruises and/or commercial fisheries partnerships. Furthermore, species data are most commonly presence-only, which presents yet another challenge for predicting the occurrence, ranges, and niches of species offshore. Predictive habitat modeling has become an increasingly useful tool for marine ecologists and conservation scientists to estimate patterns of species distribution and subsequently develop conservation strategies (Johnson and Gillingham 2005; Tsoar et al. 2007; Edrén et al. 2010; Ready et al. 2010). Sea turtle biologists trying to describe olive ridley population dynamics and their oceanic habitat distribution have stressed the importance of including in situ and remotely sensed oceanographic data to best determine habitat utilization (Plotkin 1995; Polovina et al. 2004; Robert 2006; Eguchi 2007). Most recently, marine ecologists have started to utilize a modeling software called Maxent that is rooted in maximum
entropy (Edrén et al. 2010; Friedlaender in review). Maximum entropy is similar to maximum likelihood, which maximizes a probability density function over all possible values of the parameter(s). However maximum entropy estimates an unknown probability distribution represented by significant predictor variables by optimizing the probability distribution over all samples simultaneously. Maxent is a generative model that fits a distribution of the input data as ‘loose’ as possible while still being consistent with the observed data. The probability distribution is allowed to approach ‘disorder’ only if it still meets the user-specified constraints, and the distribution must be uniform. Maximum entropy represents a density distribution that is “maximally noncommittal with regard to missing information” (Uffink 1995).

Maxent (Phillips et al. 2006) has been used extensively in terrestrial habitat modeling since its release in 2005 (see Graham and Hijmans 2006; Pearson et al. 2007; Ward 2007; Evangelista et al. 2008; Prates-Clark et al. 2008; Loarie et al. 2008; Kuemmerle et al. 2010); but it has only recently become a recognized method to model open-ocean species distributions. Maxent is gaining traction by marine scientists because:

- It does not require absence data.
- It is suitable for smaller samples sizes.
- There are ways to overcome spatial sampling bias.

All of the above are constraints on rare-species open-ocean species data sets such as sea turtles and marine mammals. What makes the software even more appealing is that Maxent is neatly packaged with a user-friendly java interface, with a variety of user-defined parameter options and constraints. The upside of this is that the user does not need to fully understand the complex operations being completed ‘behind-the-scenes,’ but the downside is that since the user may not be familiar with all of the operations, the model output can be difficult to interpret and at times misleading (Dean Urban unpub.). Nevertheless, it is becoming increasingly popular as a way to model large regions of
oceanic habitat using limited species occurrence data and coarse environmental data (see Friedlaender et al. *in review*; Edrén et al. 2010; and Allen et al. 2009), so many marine ecologists are interested in reviewing its application in marine ecosystems.

Prior to the mid-1990s, the most common presence-only habitat modeling techniques were envelope models (Guisan & Zimmerman 2000). More recently, a number of habitat models have been developed that use presence-only data (in addition to Maxent) to predict habitat suitability relative to the available environmental conditions in a region, such as ecological niche factor analysis (ENFA) (Hirzel et al. 2002), genetic algorithm for rule-set prediction (GARP) (Stockwell and Peters 1999, Peterson and Kluza 2003), and logistic regression approaches (Guisan et al. 2002; Keating and Cherry 2004; Ward et al. 2008). In this paper, I describe the oceanic distribution of ETP olive ridleys using two common distribution modeling techniques, generalized linear models and generalized additive models (GLM and GAM) (Yee and Mitchell 1991; Guisan & Zimmerman 2000; Guisan et al. 2002), and Maxent (Phillips et al. 2006). Each model identifies viable oceanic habitat based on the same three environmental parameters (chlorophyll(a) density, sea surface temperature, and depth) by overlaying remotely-sensed oceanographic data with ship-based pelagic occurrence data for both male and female olive ridley sea turtles of all size classes I captured in 2006. The models were validated using historical data from 1989 to 2006, and the results were compared using area under the Receiver Operating Characteristic curve (ROC) values, also known as ‘AUC.’

**Methods**

**Study area**

Since all of the samples used in this study were taken from the *Stenella* Abundance Research Project (STAR) ecosystem assessment cruises, I chose to use NOAA’s designated ETP study area.
as the study area (i.e. the area in which habitat is predicted based on species distribution models) for this investigation (Figure 3). Throughout the remainder of this report, the oceanic ETP study area will be shown without land as a reference.

Figure 3. Map of the eastern tropical Pacific (ETP) study area used in this investigation, as designated by the ongoing National Oceanic and Atmospheric Administration (NOAA) Stenella Abundance Research (STAR) ecosystem assessment cruises.

Species data collection

I hand-captured 350 olive ridley sea turtles in the eastern tropical Pacific (ETP) on the Stenella Abundance Research (STAR) ecosystem assessment research cruise from 8 August to 2 December 2006. The following morphometric data were collected for each sample: weight, straight carapace length (SCL), straight carapace width (SCW), curved carapace length (CCL), curved
carapace width (CCW), tail measurement, head width, body depth, number of scutes, and comments on physical appearance. Historical turtle capture data were obtained from the Southwest Fisheries Science Center from STAR cruises that took place during years 1989, 1990, 1992, 1993, 1998, 1999, 2000, and 2003. All cruises occurred during the same season (i.e. late July through early December). For data exploration purposes, olive ridley turtles for all years were divided into four size classes based on SCL and the commonly accepted reproductive potentials, as noted in the literature: 56-79cm Adult, 49-56cm Sub-adult, 6-49cm Juvenile, <5cm hatchling (Figure 4).

In 2006, three Satellite Depth Recording (SDR) transmitters were deployed in the ETP on one adult male, one adult female, and one sub-adult presumed female with SCLs of 66.5cm, 62.8cm, and 53.5cm respectively. Dates and locations of deployment were 20 September, 10°23 N, 87°41 W (~190km offshore); 9 October, 14°48 N, 94°34 W (~175km offshore), and 5 November, 14°29 N, 97°34 W (~150km offshore) respectively. The tags transmitted between 122 and 194 days. Exploratory maps of individual tracks concurrent with remotely-sensed sea-surface temperature,
bathymetry, chlorophyll and geostrophic currents were generated with the Satellite Tracking and Analysis Tool (STAT) (Coyne & Godley 2005). While these data were too small (n = 3) and spatially biased for incorporation into this analysis, one of the maps provides insight into further study as highlighted in the ‘Future Research’ section.

**Environmental predictors**

Concurrent oceanographic information was recorded as part of the STAR ecosystem assessments. Expendable bathythermographs (XBTs) were conducted three times daily at 900, 1200 and 1500 to record temperature as a function of depth and capture thermocline information. Conductivity, Temperature, Depth (CTD) casts were conducted twice daily at 0600 and 2000, to measure physical properties of seawater (salinity, CO2, temperature, pressure, and density). For each turtle sampled, capture and release GPS coordinates were recorded, and *in situ* sea surface temperature (SST) and bottom depth data were collected.

Turtle presence points (n = 350) were generated in ArcGIS version 9.3 from latitude/longitude information recorded at the time of capture. In order to evaluate and compare associated environmental information between presence points recorded during the fall of 2006, and to ensure spatial consistency among environmental layers, I sampled underlying seasonal average layers of four environmental predictors at each presence location to generate an attribute data table. This attribute table was populated with calculated seasonal average values of remotely-sensed data, in place of *in situ* data.

I obtained monthly averages of remotely-sensed satellite data for two dynamic environmental variables for dates consistent to turtle captures: sea surface temperature (SST) and chlorophyll(a) concentration (Table 2). I also obtained a static representation of ETP bathymetry (Table 2), which I refer to as “depth” throughout the remainder of the paper. These predictors were chosen because they are consistently monitored and are easily accessible through online resources, and due to
existing evidence that olive ridleys occur in warm (24-31°C) and productive (oceanic waters (Beavers and Cassano 1996; Polovina et al. 2004).

Table 2. Source, resolution and scale of the remotely-sensed environmental data used in analysis.

<table>
<thead>
<tr>
<th>Source</th>
<th>Resolution</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>SST</td>
<td>NOAA’s Advanced Very High Resolution Radiometer (AVHRR)(^1)</td>
<td>0.1 deg (8km / pixel) monthly</td>
</tr>
<tr>
<td>Chl(a)</td>
<td>NASA’s Moderate Resolution Imaging Spectroradiometer (MODIS) on Aqua(^1)</td>
<td>0.05 deg (4km / pixel) monthly</td>
</tr>
<tr>
<td>Depth</td>
<td>Global Self-consistent, Hierarchical, High-resolution Shoreline (GSHHS)(^2)</td>
<td>1 arc-minute N/A</td>
</tr>
</tbody>
</table>

Satellite data commonly suffer data gaps due to interference from cloud cover. In order to reduce the number of ‘no data’ grid cells, seasonal averages (August - November 2006) were calculated for the two dynamic predictors, SST and chlorophyll(a). Further averaging using a 3x3 cell moving window with the ‘Focal Statistics’ tool in ArcGIS generated the final seasonal average SST and chlorophyll(a) predictor layers (Figure 5). I calculated Euclidean distance from shore to develop a raster grid of distance values.

To accurately display and analyze the data in ArcGIS and minimize spatial distortion, I created a unique Albers equal area conic geographic projection centered on the ETP study area with the following parameters:

- Parallel 1 = -7 decimal degrees (dd)
- Parallel 2 = 21 dd
- Central Meridian = -115 dd
- Latitude of Origin = 7 dd

I chose an Albers projection because it is recommended for regions predominately east-west in orientation and located in middle latitudes, not to exceed 35° (see ArcGIS 9.3 Desktop Help “Albers Equal Area Conic”). I projected all data and products using this unique projection throughout my analysis.

---

\(^1\) Downloaded from http://las.pfeg.noaa.gov/oceanWatch/oceanwatch_safari.php.
\(^2\) Downloaded from http://www.ngdc.noaa.gov/mgg/global/etopo1sources.html.
Once the values of all environmental layers were extracted for all points, I explored frequency distributions and correlation amongst the four environmental parameters: SST, chlorophyll$(a)$, depth, and distance from shore. In all analysis, I log transformed chlorophyll$(a)$ to normalize the data. It is important to note that distance from shore and depth are spatial predictors, whereas chlorophyll$(a)$ and SST vary both temporally and spatially. This becomes helpful when choosing predictors for inclusion in habitat models, and when interpreting the results of each model. Based on the frequency histograms in Figure 6, most 2006 turtle captures occurred in water 2,500 - 4,000m deep, temperatures of 27.5 - 31°C, and 0.1 - 0.4 mg m$^{-3}$ chlorophyll$(a)$ concentrations. The Pearson correlation coefficient of 0.59 shows that distance and SST are the only predictors that are loosely correlated (Figure 6).
Figure 6. Scatterplots of the data points for each environmental predictor are displayed in the bottom left, including red trend lines to show general relationships. Frequency histograms (solid blue bars) plotting the distribution of each predictor are shown diagonally in the middle. Pearson correlation coefficients describing the relationship between two predictors are displayed in the upper right; text is sized by magnitude of correlation (larger, more correlated; smaller, less correlated). Note the correlations between ‘Distance’ to shore and ‘Depth’ (0.45), and ‘Distance’ to shore and sea surface temperature (‘SST’) (0.59). This plot was created using MGET tools in ArcGIS and the R Project for Statistical Computing.

Distance to shore was correlated with both depth (0.45) and SST (0.59), and all are surrogates variables for other predictor(s) not considered in this analysis (e.g. thermocline; distribution of prey; currents; etc.). Distance and depth are also inherently biased due to the opportunistic nature of the
sampling effort along survey transect lines and dependency on agreeable weather and sea state, therefore it was appropriate to remove one of the two from the analysis. Since distance was correlated with two other predictors and inherently biased, I removed it from my analysis and built habitat models using the remaining three predictors: SST, chlorophyll(a) and depth.

**Correcting for sample bias**

Sample selection bias affects habitat modeling and should be dealt with prior to building a model, or considered carefully in model interpretation. Occurrence data usually exhibits strong spatial bias because survey effort is fundamentally connected to accessibility (Phillips et al. 2009); this is inherently true of data collected at sea. As previously mentioned, the data used in this study are subject to this type of sample bias. For example, more samples occur off the coast of Acapulco, Mexico because the scientific crew of the R/V David Starr Jordan conducted calibrations of visual observers for ten days in a localized region concurrent with turtle sampling effort. Turtle density was high in the area and I had ample access to small boats for field work, so our sampling effort was higher during these ten days and in this area compared to other regions.

When pseudo-absence data are used in fitting a habitat model, they are commonly generated at random over the entire study area being classified. In traditional regression methods (e.g. GLM and GAM), pseudo-absence data is used to fit the model, whereas in Maxent pseudo-absence data serves as “background” data and is used only for model evaluation (e.g. ROC, AUC). This is the key to what makes Maxent a presence-only modeling technique. Here I generated a single set of pseudo-absence points that serves both purposes: model fitting for GLM/GAM; and model evaluation for GLM/GAM/Maxent.

Biased occurrence data combined with pseudo-absence data that falls outside the region where presence points were sampled can decrease the classification and predictive performance of a habitat model (Phillips et al. 2009). Therefore, constraining pseudo-absence data to a restricted area
relative to your sample bias has been shown to increase predictive performance in generalized linear
or additive models, boosted regression trees, and Maxent (Phillips et al. 2009). To minimize the
effects of sampling bias in this study, I ensured the pseudo-absence data reflected the same bias as
the presence data by limiting the generation of pseudo-absence data to a convex hull around the
presence points (see Phillips et al. 2008).

To generate the convex hull, I used Hawth’s tools in ArcGIS. The convex hull around just
the region where I had occurrence data was a substantially smaller area than my entire ETP study
area. I then generated pseudo-absence points within the convex hull, and then used the ‘Sample
Rasters’ tool within open-source Marine Geospatial Ecology Tools\(^3\) (MGET) to sample all of the
underlying environmental rasters to capture associated predictor values. I used these constricted
pseudo-absence points to build my GLM and GAM models, and serve as the “background” file in
the Maxent model. This technique is illustrated in Figure 7. Although all pseudo-absence /
background data points are generated at random within the specified convex hull, they are notably
more uniform than the true presence points. Presence points are also located directly next to
pseudo-absence points, perhaps within the same grid cell, meaning they could have exactly the
same value for a given predictor. This becomes important when interpreting model results because
each of the three models compared in this study respond differently to discriminating, or not being
able to discriminate, between ‘habitat’ and ‘non-habitat’ based on the inputs. There are several
ways to try and combat these effects, if deemed worthwhile. For example, I could have buffered the
presence points before generating the pseudo-absence points, which may have reduced this
phenomenon, at least nominally. I could have also buffered the entire convex hull by an arbitrary,
reasonable range (e.g. 1km) and generated the pseudo-absence points within the buffered convex

\(^3\) Available at: http://code.env.duke.edu/projects/mget.
hull, which may have minimized the exponential responses at the edges of the term plots (see Figures 9 and 11).

Figure 7. Pseudo-absence points were randomly generated within a convex hull encompassing olive ridley presence points. These pseudo-absence points were used to sample the underlying environmental data, and then utilized as the ‘absence’ data in the GLM and GAM, and the “background” file for the Maxent model (see Appendix for details).

Habitat modeling

Once the observational point data are associated with the environment in space and time, an algorithm is used to model the multivariate relationship between the environment and species presence. The habitat model performance is tested based on a subset of the original presence data
or a relevant independent dataset. Finally, the model is used to predict likelihood of presence over a new region and/or set of environmental conditions (Guisan & Zimmerman, 2000).

**Fitting the model**

I used open-source MGET tools (Roberts et al. *in press*) to complete GLM and GAM predictive models (see Beyer et al. 2010; Appendix).

**GLM:** \( \text{Presence} \sim \log(\text{Chl}) + \text{SST} + \text{Depth} \)

**GAM:** \( \text{Presence} \sim \text{s}(\log(\text{Chl})) + \text{s(SST)} + \text{s( Depth)} \)

I used the open-source Maxent\(^4\) software package to complete maximum entropy predictive models (Phillips, et al. 2006). The parameter settings and applied constraints in the final model are listed in Table 3. I used both regularization (i.e. response curve smoothing) and sample bias mitigation techniques recommended by Phillips (2008, 2009) using a ‘target group’.

Table 3. Maxent model parameters used in this study. See Appendix for screenshots of Java software inputs. See the Maxent Tutorial (available at http://www.cs.princeton.edu/~schapire/maxent/) or Phillips et al. (2006, 2008 & 2009) for complete definitions and explanations of each parameter.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Training Records</td>
<td>292</td>
</tr>
<tr>
<td>No. of Testing Records</td>
<td>850</td>
</tr>
<tr>
<td>Regularization</td>
<td>Default</td>
</tr>
<tr>
<td>linear/quadratic/product (lpq)</td>
<td>0.050</td>
</tr>
<tr>
<td>categorical</td>
<td>0.250</td>
</tr>
<tr>
<td>threshold</td>
<td>2.000</td>
</tr>
<tr>
<td>hinge</td>
<td>0.500</td>
</tr>
<tr>
<td>Replicated Run Type</td>
<td>Crossvalidate</td>
</tr>
<tr>
<td>Random Test Percentage</td>
<td>0</td>
</tr>
<tr>
<td>Test Sample File</td>
<td>TRUE</td>
</tr>
<tr>
<td>Clamping</td>
<td>TRUE</td>
</tr>
<tr>
<td>Fade by Clamping</td>
<td>TRUE</td>
</tr>
<tr>
<td>Threads</td>
<td>2</td>
</tr>
<tr>
<td>Beta Threshold</td>
<td>2.0</td>
</tr>
<tr>
<td>Beta lqp Threshold</td>
<td>0.05</td>
</tr>
<tr>
<td>Apply Threshold Rule</td>
<td>FALSE</td>
</tr>
<tr>
<td>Bias File</td>
<td>FALSE</td>
</tr>
</tbody>
</table>

\(^4\) Available at: http://www.cs.princeton.edu/~schapire/maxent/.
Model Validation

Habitat models generalize environmental conditions in order to characterize habitat suitability. In order to make predictions with new data for applications beyond the model training, such as predicting species distribution, the model should be carefully tested for reliability (Guisan & Zimmerman et al. 2000; Vaughan and Ormerod 2005; Elith et al. 2006). Ideally, an independent testing data set represents a range of conditions over which you intend to draw inferences from the model is best, and the data should be drawn from independent sites (Vaughan and Ormerod 2005). For rare species and hard-to-reach places, this is often impossible to achieve. Therefore, a common approach in presence-only modeling is to cross-validate the model by withholding a certain percentage of training data when fitting the model to then later test the accuracy of the model (e.g. training data = 80%; testing data = 20%). Since I have access to historical data within the same study area (1989-2003), I can use all of the 2006 presence data for training, and validate the model using the historical presence points, or vice versa. Neither method is perfect and it is at the discretion of the user and/or model interpreters to decide which method is more useful. Here I employed both techniques, and due to the determined usefulness of the results I am reporting on the models that use the latter approach.

To validate the GLM and GAM models, I report here on models that were trained using the 2006 data set and validated using the historical data set. For the Maxent models, I am reporting

Collate environmental layers (GIS)

| Process environmental layers into predictor variables | Map species occurrence data |

Apply modeling algorithm

| Model calibration (e.g. adjust model parameters, eliminate predictors) |

Predict habitat (in different region / time period)

| Statistically test predictive performance (e.g. AUC) | Validate with observed data |

Figure 8. Simplified schematic of the methodology used in this study.
on two models: (a) which used the 2006 data set for training and the historical data set for testing; and (b) which used the historical data set for training and the 2006 data set for testing. An illustrative summary of my entire methodology is shown in Figure 8.

**Results**

**GLM**

GLMs and GAMs produce terms plots that help the user interpret the relationship between the predictor and the response. When modeling habitat the response is binary: presence or absence. Figure 9 shows the term plots of all three predictors in the GLM. The solid line represents the mean fit, and the dotted lines indicate the confidence intervals at 2 standard errors. A significant response is seen when the confidence intervals of the fit are not overlapping the x-axis, for either a positive (above) or negative (below) response. The closer the lower and upper confidence thresholds are to each other, the tighter the fit. Because the absence points were generated within a convex hull encompassing the presence points, many of the environmental variable values were similar for a presence point vs. a pseudo-absence point. This highlights one of the drawbacks of using discriminative models like GLM and GAM when the user only has access to presence data. Generating pseudo-absence points to sample the background environmental layers introduces bias. For example in this case, the GLM is having a hard time distinguishing if a chlorophyll(a) density value of 0.2 mg m\(^{-3}\) is preferable or avoided by an olive ridley, since there are both many presence and absence data points associated with each response. This is shown by the fit and confidence intervals spanning the x-axis between 0 and 0.4 mg m\(^{-3}\), and crossing at about 0.2 mg m\(^{-3}\) (p < 3.71e-09; Figure 9a). As the data points become more sparse at concentrations greater than 0.5 mg m\(^{-3}\), the confidence intervals widen and are interpreted as unreliable.
SST is very stable in the ETP, and therefore all the presence and absence points fall within a small range of 26-31 °C. Figure 9(b) indicates that as surface temperature (p < 2e-16) increases, the likelihood of olive ridley presence increases, and the sampled turtles prefer warm temperatures of 29-31 °C within the studied space and time.

The majority of presence and pseudo-absence points were in water 2,500-4,500 m deep, however depth was not a significant contributor to the model (p < 0.911). This makes it hard for an discriminatory model like a GLM to find a relationship between depth and the response, therefore the term plot in (Figure 9(c)) shows the fit basically on top of the x-axis, indicating no relationship, and the upper and lower confidence thresholds on either side of the x-axis, meaning the model has no idea if a certain depth will correspond with a presence or absence response.
Figure 9. GLM terms plots for the three predictors: chlorophyll(a), sea surface temperature, and depth. The solid line is the mean fit; dashed lines represent 2 standard error deviations. Data points are displayed as open circles and presence density for the predictor is shown along the bottom of the figure with the densest areas displayed in continuous black.
The GLM predictive habitat model (Figure 10) shows a band of continuous suitable habitat, indicated in warm colors (i.e. red), along the coastal areas of Mexico and Central America, however virtually no suitable habitat along the coast of South America or in open ocean, indicated in cooler colors (i.e. blue) (Figure 10 (a)). The white circles are the 2006 presence points used to fit the model. Figure 10(b) shows the standard error for the model. Where there is suitable habitat predicted there is also high associated uncertainty, or error regarding that prediction.
Figure 10. (a) The resultant GLM predictive habitat classification map, including the training points plotted in white; and (b) the standard error associated with the predictive map.
\textit{GAM}

The smoothed term plots for the GAM are displayed in Figure 11. These are very different relationships than seen in the GLM term plots (Figure 9), as none of the predictors have an obvious consistent relationship with the response but all were significant contributors to the model (see Appendix):

\begin{align*}
\log(\text{Chl}) & \quad p = 8.09 \times 3.26 \times 10^{-15} \\
\text{SST} & \quad p < 2 \times 10^{-16} \\
\text{Depth} & \quad p = 5.89 \times 10^{-6}
\end{align*}

This is illustrated by the fit line hovering around the x-axis where the data are rich, and the confidence intervals growing large when deviating away from the axis in data-poor parameter space. Basically the term plots tell the user that the model is very certain that it cannot distinguish between habitat and non-habitat.
Figure 11. GAM terms plots for the three predictors: chlorophyll(a), sea surface temperature, and depth. The solid line is the mean fit; dashed lines represent 2 standard errors deviation from the mean. Data points are displayed as open circles and presence density for the predictor is shown along the bottom of the figure with the densest areas displayed in continuous black.
As was indicated by the term plots, the predictive habitat map (Figure 12) is conflicted about what to call habitat vs. non-habitat. The map shows highly suitable habitat in areas where there was no presence data used to fit the model, which is a danger of predicting beyond the scope of originally fitted data. Based on the confidence intervals of term plots (a) and (c) in Figure 11, I would expect to see high error in offshore areas (i.e. deep water, low chlorophyll and average temperature) where high presence is predicted, but instead the GAM shows low standard error in the northwestern parts of the study area (Figure 12).

Because GLM and GAM are discriminative models, I expected higher standard error within the convex hull containing the data used to fit the models. Since many of the presence and pseudo-absence points are associated with similar environmental data, it is hard for a discriminative model to distinguish between the two competing inputs. This holds true for the GLM (Figure 10) but surprisingly not for the GAM (Figure 12). In fact the GAM prediction shows only patchy suitable habitat within the convex hull, with small associated error. This makes the predictions hard to believe.
Figure 12. (a) The resultant GAM predictive habitat classification map, including the training points plotted in white; and (b) the standard error associated with the predictive map.
Maxent produces response curves similar to logistic term plots. Here, “LO” in the plot titles refer to the olive ridley species name, *Lepidochelys olivacea* (LO). These response curves are smoothed via the regularization model parameter settings to avoid over-fitting the data (see Phillips et al. 2008). Figure 13(a) shows a varied response to chlorophyll(*a*) at low levels < 0.2 mg m$^{-3}$, but that most olive ridleys sampled in this space and time prefer 0.2-0.6 mg m$^{-3}$ levels.

The SST response curve illustrates one of the differences between generalized logistic models and maximum entropy modeling. Both Figure 9(b) and 11(b) show that the majority of olive ridleys sampled in this study prefer the ~ 29.5-31.5 °C; Figure 9(b) shows a linear response, but Figure 13(b) shows a more realistic relationship between the data and the response. If the data weren’t regularized (i.e. smoothed), the response curves would be more jagged with exaggerated inflection points (known as “knots” in a GAM) to ensure an accurate fit. This is why Maxent is often criticized for over-fitting the data, and steps such as regularization should be taken to reduce the incidence of overfitting.

The depth response curve shown in Figure 13(c) further shows that depth has no strong relationship to species presence. There is a slight peak in the curve below 1,000 m and above 5,000 m, and this is most likely due to sampling bias. This is also shown in the predictor contribution results: chlorophyll(*a*) 58.2656%, SST 40.2553%, Depth 1.4792%.
Figure 13. Maxent response curves for the three predictors: (a) chlorophyll, (b) sea surface temperature, and (c) depth.
The Maxent prediction map shows a more complex habitat seascape, classifying virtually all of the ETP as somewhat suitable habitat for olive ridleys, with the exception of an area in the center and a small northern section of the study area, and section in the northern part of the study area (Figure 14). Maxent does not output a comparable error surface like the GLM and GAM, so in this figure I have chosen to highlight another part of the analysis, the validation step. The model training data from 2006 is displayed with black circles, and the historical test data are displayed with white circles.
Figure 14. The same Maxent predicted habitat map is shown twice here with the training points from 2006 plotted in the upper panel and the historical testing points plotted in the lower panel.
**Model evaluation**

Receiver operator characteristic (ROC) curves plot the true positive rate (“sensitivity”) against the true negative rate (“specificity”) (Pearce & Ferrier 2000). Points along the ROC curve represent the tradeoff between the two prediction rates, and the point along the curve which the user picks to ‘tune’ the model is known as the “cutoff” value— the decision threshold used to optimize model performance by finding an appropriate balance between maximizing true positives (correctly predicting habitat) and minimizing false positives (incorrectly predicting non-habitat), or a user-specified balance of true positives and true negatives. The cutoff value is also used to create a binary (habitat/non-habitat) prediction map. The user’s choice of an ROC cutoff value is subjective and can vary based on intent, such as whether the user’s study is exploratory or predictive; or when predictive, whether a continuous or binary prediction is desired (see Guisan & Zimmerman 2000).
Figure 15. GLM ROC curve indicating the cutoff value is 0.532561, which can be used to create a binary habitat map.
**GAM**

Figure 16. GAM ROC curve indicating the cutoff value is 0.566241, which can be used to create a binary habitat map.

**Maxent**

Figure 17. Maxent plot of true positive rate, known as “sensitivity”, vs. false positive rate, or (1 – “specificity”). Specificity is the true negative rate. Although Maxent is not logistic regression, it does offer an option to balance sensitivity and specificity, which acts as the decision threshold similar to a cutoff value in GLM and GAM.
Discussion

Picking the best model

Area under the ROC curve (AUC) is a common tool used to evaluate and compare distribution models. Note that it is a single performance evaluation measure that is independent from the cutoff value. The inverted pyramid in Figure 18 shows that an AUC value closest to 1 (1 = perfect prediction of habitat) is considered a better than an AUC value of 0.5, which indicates the model cannot discriminate between habitat and non-habitat. The AUC values for the 2006 presence samples used to fit the model, the historical presence samples used to test the model, and the percent change between the two are listed in Table 4 and color-coded according to Figure 18. The smaller the percent change between the fit and test models, the more robust the model.

Table 4. Area under the ROC curve (AUC) values and AUC percent change from training to testing for the GLM, GAM and Maxent habitat models. See Appendix for full model statistics for each.

<table>
<thead>
<tr>
<th></th>
<th>GLM</th>
<th>GAM</th>
<th>Maxent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fit</td>
<td>0.864</td>
<td>0.642</td>
<td>0.595</td>
</tr>
<tr>
<td>Test</td>
<td>0.783</td>
<td>0.880</td>
<td>0.672</td>
</tr>
<tr>
<td>% Change</td>
<td>-9.38</td>
<td>37.07</td>
<td>12.94</td>
</tr>
</tbody>
</table>

Ideally, the best habitat model would have both “Good” to “Excellent” AUC values and the smallest percent change between the fit and test. My results were not conclusive in this way. Statistically, the GLM was the best performing model (see Table 4), which makes sense based on the exploratory plots and simple linear relationships among the predictors (see Figure 6). However, the suitable habitat that the GLM predicted over the entire ETP study area (see Figure 10) does not
make sense according to what we currently know about olive ridley ecology (Refer to “Introduction”). Based on extended personal experience and discussion with experts (e.g. Pitman, pers. comm.), the GLM predicted habitat is too conservative and therefore discredited. It predicts that there is no suitable habitat in the oceanic areas outside of the convex hull containing the 2006 presence points used to correct for sample bias (see Figure 7), which is wrong.

One reason the AUC values may be low for the Maxent model is due to the restraints applied to avoid over-fitting the data (see Table 3). This is a common complaint discussed in more detail later in this section. Looking at just robustness (i.e. smallest percent change between fit and test), Maxent appears to perform well. If two models have similar AUCs and it’s necessary to compare them, as is the case here with the GLM and Maxent, there are two options: 1) employ complex statistical test (e.g. bivariate statistical analysis), and/or 2) draw from expert ecological knowledge. Here I chose the latter. All things considered, I conclude that despite the least stellar AUC values, the Maxent model produces an intuitively better map of suitable habitat in the ETP. But none of the models are considered acceptable nor should they be used for management purposes. The resultant habitat maps should only be used to compare the three distribution models examined in this study given the available data.

Although all modeling techniques failed in this analysis, they are still valid approaches to modeling species distribution in open-ocean systems. Their failure in this case was due to limited data, discrepancies between the range of environmental conditions of the input data and across the prediction surface, truncated post model processing, and coarse environmental data. Limited data (small sample size, presence-only) is a common problem when modeling rare and/or hard to survey species such as migratory sea turtles. GLM and GAM deal with presence-only data by the incorporation of pseudo-absence data, and there are several ways to generate those data. The generation method chosen can affect model performance. In this case, I chose to generate pseudo-
absence data within a tight (no buffering) convex hull containing the presence data points. Had a
different method been chosen or perhaps a few different methods compared and contrasted, the
GLM and GAM may have performed better. Furthermore, both GLM and GAM are discriminatory
models and it is difficult to discriminate between habitat / non-habitat when similar environmental
conditions are associated with both presence and pseudo-absence data points. This is one reason
why Maxent is appealing— it is generative and uses presence-only data. The pseudo-absence data
(known as “background”) is only used to evaluate model performance. Maxent has also been
proven to work well for small sample sizes, and therefore appropriate for the sample size in this
study (n = 350).

Given this, I expected Maxent to perform better than both the GLM and GAM. Although
Maxent seemed to produce the most sensible prediction map, statistically it performed the worst of
all three models. This could be due to the model constraints applied (see Table 3 and Appendix)
with the intention of smoothing variable response curves and mitigating sample bias. More
exploration of model parameter combinations could have provided a clearer picture of which
parameter constraints caused the model to perform poorly. It is possible that I over-compensated
when smoothing and reducing sample bias.

For all three models, the pseudo-absence data (used to fit the GLM and GAM, and evaluate
all three models) were constrained to the convex hull. This is recommended and appropriate for
Maxent, but perhaps not necessary for the GLM and GAM. Both methods should be compared and
contrasted for more insight.

Only two of the predictors contributed substantially to building all three models
(chlorophyll(a) and sea surface temperature). This makes for an over-simplified model that is hard
to interpret ecologically; it doesn’t tell much about what is really going on because it is not a
substantive multi-variate model. Ways to include more descriptive predictors that have a more
direct and/or a greater influence on olive ridley habitat selection is outlined in the ‘Future Research’ section.

Seasonal averages of remotely-sensed environmental data were used to fit and predict in all three models. Averaging satellite data ‘smears’ the data and results in a loss of precision. Therefore, the models can only assess a coarse set of environmental conditions available to an animal. Incorporating finer-scale data such as a smaller averaging scale (e.g. monthly, weekly) or \textit{in situ} data would provide a more accurate set of environmental conditions available to an animal.

The ETP study area used in this analysis is huge compared to the area where presence data points were available. The offshore areas I was most interested in predicting olive ridley habitat for often had environmental variable ranges outside of the ranges found within the convex hull containing the presence and pseudo-absence / background data points. Therefore, large standard error outside of the convex hull was expected. Incorporating presence data for both nearshore and offshore areas would improve these models and could be accomplished by utilizing ship-based line transect survey data and/or capture/bycatch data from scientific surveys/commercial fisheries.

\textit{Potential role of ETP oceanographic features}

The eastern tropical Pacific (ETP) is the ocean region centered on the eastern Pacific warm pool and includes the equatorial cold tongue and equatorial current system (Ballance et al. 2006; Fiedler and Talley 2006). Figure 19 depicts the complex oceanographic features of the ETP. The region supports a diverse and abundant community of apex predators, including a number of endemic seabird and cetacean species and sub-species, relative to other tropical oceans. The reasons for this are virtually unknown but it has been postulated that its unique thermal structure and other oceanographic characteristics play an important role (Spear et al. 2001; Ballance et al. 2006). The ETP has a strong, shallow thermocline that sustains constant primary productivity in
surface waters (Pennington et al. 2006), therefore the region is moderately productive in coastal upwelling regions and, uniquely, in oceanic regions as well.

Some seabirds in the ETP have been shown to associate with major currents such as the North Equatorial Countercurrent, the South Equatorial Current (Pitman 1986), the Costa Rica Coastal Current, and the Peru Current (Spear et al. 1995). Three physical features have been shown as particularly influential for seabirds and/or cetaceans including the Equatorial Front, the Costa Rica Dome, and the countercurrent thermocline ridge (along 10° N latitude) because they enhance productivity of surrounding waters and attract predators (Ballance et al. 2006). Sea turtles are believed to select suitable habitat based on physical and biological oceanographic conditions (Luschi et al. 2003; Polovina et al. 2004; Seminoff 2008), however conclusions are poorly developed and more research is needed.

Figure 19. Schematic diagram of surface water masses and currents in the eastern tropical Pacific. STSW, Subtropical Surface Water; TSW, Tropical Surface Water; ESW, Equatorial Surface Water. Shading represents mean sea surface temperature (darker = colder; Shea et al. 1992) (Figure from Fiedler and Talley 2006).
The ETP has many distinct, permanent macro- and meso-scale features that vary predictably in time and space (e.g. major surface currents, boundaries between currents, large gyres and eddies, etc.) (Fiedler and Talley 2006). Specifically, fronts are believed to be significant to most marine organisms by aggregating prey and providing strong gradients for visibility, nutrients, temperature and salinity (Olson et al. 1994; Schreiber 2002). It is generally assumed that easily-measured and therefore accessible variables such as depth or sea surface temperature are proxies that directly or indirectly relate to prey abundance or distribution (Ballance et al. 2006). Based on this and what we know about physical and biological oceanographic features influencing apex predators in the ETP, there are likely to be additional, less studied parameters that influence olive ridley habitat beyond chlorophyll and sea surface temperature, such as sea surface currents, mean sea level anomaly (Roberts 2006), and prey density. See the ‘Future Directions’ section for more on this topic, and how telemetry and survey data can contribute to building more accurate and reliable pelagic habitat models.

Thoughts about Maxent – a viable presence-only oceanic habitat modeling tool?

Maxent is unique in that:

1) It is generative, not discriminative (e.g. GLM, GAM)

2) It can handle variables in any state (auto-correlated, interacting, linear, quadratic, etc.)

What is not always understood about maximum entropy modeling is that it is actually a probability density estimation modeling technique, meaning it will find the optimum of all the distributions of the predictors and reconstruct a density distribution across the indicated study area that is consistent with the occurrence data and the maximized entropy of the variables (Brierley et al. 2003, Phillips et al. 2006, Phillips et al. 2008, Phillips et al. 2009).

Limitations

In summary, there are limitations to Maxent that users should be aware of:
1) Maxent is revered because it fits any type of response function (e.g. quadratic, linear, hinge) fairly well, however, that is also one of the common criticisms of the approach (Vaughan and Ormerod 2005; Peterson et al. 2007; Lozier et al. 2009; Ready et al. 2010). By nature, maximum entropy is “maximally noncommittal with regard to missing information” (Uffink 1995), or in other words, finds a distribution over the predictor variables that describes the observed (i.e. presence) habitat samples with as few additional constraints as possible (Dean Urban, unpub). Generally speaking, Maxent tends to over-fit training data unless proper restrictions are applied during the model fit. Using the Maxent Java interface is quite simple, however knowing which model constraints to apply can be confusing (Haegeman and Etienne 2010). As such, caution should be applied when developing Maxent models (e.g. knowing when and how to apply certain model constraints), and especially when interpreting results.

2) Although Maxent is very effective at fitting complicated environmental covariate distributions with a response, its predictive capability and success is not fully proven (Vaughan and Ormerod 2005; Ready et al. 2010; Song Qian pers. comm.). Presence-only models have numerous applications, but relatively few studies (Elith et al. 2006; Hirzel et al. 2006; Tsoar et al. 2007) have assessed and/or compared the performance and reliability of different methods. Validation (testing model prediction against a subset or independent dataset) and resampling / cross-validation (e.g. bootstrapping) evaluation methods, among others, should be employed. These validation steps are especially important when forecasting climate change scenarios.

3) As stated earlier, Maxent’s easy-to-use software interface can actually be a hindrance as anyone can use it (not just a habitat modeling whiz) without understanding the complicated statistical mechanisms going on ‘behind the scenes.’ It’s unfair to assume that
every user should understand all of the intricacies of the mathematical computations involved, however it is fair to assume that the user understands maximum entropy theory, the objectives of the mathematical processes Maxent employs, and how to properly interpret results based on user-defined inputs and model constraints. As Maxent increases in popularity among marine ecologists, so will user proficiency, but because this is still a newer method for marine ecosystems, users should ensure they understand the basics before applying Maxent results.

Conclusion

For the scope of this study, Maxent was determined to produce a practical species distribution model given the available data. However the Maxent model did not statistically perform better than GLM or GAM, or well enough to be used beyond this exploratory study. This study should be replicated using the directives outlined in the ‘Future Research’ section below. Similar studies should be repeated for other oceanic species (e.g. migratory marine megafauna) and compared with richer datasets (e.g. line transect surveys and telemetry) to gain a better understanding of Maxent’s ability to accurately predict oceanic habitat. I caution marine habitat modelers against using Maxent as a panacea for open-ocean, rare species predictive habitat modeling problems, despite its accessible and easy-to-use software, until we have a better handle on its reliability.

Future Research

1) STAR ecosystem assessments are conducted approximately every three years. These research cruises are extremely expensive endeavors, therefore it is imperative that NOAA maximizes time and resources by completing as much data collection as possible through partnering with various researchers and institutions. I aim to continue to use historical and
Additionally, ongoing commercial fisheries observer programs organized by the Inter-American Tropical Tuna Commission (IATTC) have collated large datasets regarding sea turtle occurrence and their interactions with commercial fisheries in the ETP. Such information would be extremely useful in not only developing species distribution models similar to in this report (see Roberts 2006), but also in directing strategic data collection in the future and building key partnerships to effectively implement fisheries bycatch mitigation across international borders. Moreover, combining the fisheries-dependent (i.e. IATTC) and fisheries-independent (i.e. STAR) data to facilitate comprehensive species-habitat and fisheries bycatch interactions and would be hugely beneficial to all parties. This would be an unprecedented step in the right direction for collaborative marine resource management, and therefore a very worthwhile pursuit.

2) Some, if not all, of the predictors used in this study are actually surrogates for other predictors like prey distribution and thermocline (Ballance et al. 2006). Developing a more robust and representative model by incorporating more sophisticated and descriptive predictors as described in the discussion would result in an improved, more accurate habitat distribution model. Parameters such as major currents (Figure 19), mean sea level anomaly (Roberts 2006), and prey density have been noted by others as influencing species occurrence and/or movements (Luschi et al. 2003; Polovina et al. 2004; Swimmer et al. 2009). Ballance et al. (2006) specifically stated that future work in the ETP should focus on the integration of biological variables (e.g. prey availability) into species-habitat
investigations. The ETP is a unique oceanographic region with complex physical interactions such as multiple surface currents, which are hypothesized to influence the distribution and movements of apex predators. In support of this point, Figure 20 suggests that dynamic surface currents influence the movements of olive ridleys. Environmental relationships derived from telemetry data such as this have yet to be robustly integrated into models using population-wide opportunistic observations.

Figure 20. Adult female olive ridley sea turtle (SCL: 53.3 cm) tracks overlaid on a static image of the surface currents on 2006 Julian Day 333 (29 November). The satellite transmitter (Wildlife Computers SDR T16) was deployed on 2006 November 05 and persisted for 166 days. The light red circle highlights the section of turtle tracks (late November - early December) that generally corresponds with the date of the static surface currents, indicating that the turtle is moving along the edges of the eddy.
Sea turtles have a unique life history where they spend their earliest years in the open ocean and undergo ontogenetic shifts from oceanic to neritic habitats during early life stages. For example, North Atlantic loggerheads, Caretta caretta, spend their first 10 years in oceanic habitats, and then recruit to nearshore feeding grounds (Bjorndal et al. 2003; Bolten 2003). However less is known about other species, such as the Atlantic green turtle (Chelonia mydas). The leatherback (Dermochelys coriacea) and olive ridley turtles spend almost their entire life in pelagic systems. Therefore not only are their first ±10 years “lost” at sea, but virtually their entire lives. This presents added challenges for fisheries and conservation managers.

I am most interested in creating accurate and robust pelagic sea turtle habitat models in order to examine how habitat selection and habitat parameters vary with sea turtle size. Different size classes of turtles carry different reproductive potentials, and are therefore more ‘valuable’ from a conservation perspective. Stage-based population models have revealed that large juveniles, sub-adult and adult annual survival is more important to the annual growth of a sea turtle population than egg or hatchling survival (Crouse et al. 1987; Crowder et al. 1994; Heppell et al. 1996). Knowing the spatial dynamics of sea turtles will better inform conservation management. Therefore, replicating this study with a focus on developing distribution models and identifying differences in habitat selection between size classes, species and sex is only logical (see Figure 3 and 21).
4) There are five species of sea turtles that occur in the ETP. Looking at one species such as the olive ridley can be useful and interesting, however multi-species, ecosystem-based studies have been shown to be more effective conservation management strategies (Botsford et al. 1997; Hughes et al. 2005; Crowder and Norse 2008). This study should be replicated and expanded upon for all five species in the ETP in order to develop a reliable understanding of the distributions and overlap of distributions, species-specific and intra-specific habitat selection criteria, of all species to inform conservation and fisheries management.
The strongest inter-annual variability of preferred cetacean habitat distributions in the ETP was attributed to the El Niño-Southern Oscillation (ENSO) cycle (Fiedler and Reilly 1994). The underlying reason for this is that species track preferable habitat, and the distribution of preferable habitat, and perhaps other parameters, changes at a variety of temporal scales (Ballance et al. 2006). Ballance et al. (2006) reported that a number of seabirds and cetaceans species distribution patterns co-vary with patterns of preferred habitat in the ETP.

The biology of many animals cycles seasonally, often correlated to reproduction activities, and sea turtles are no exception. Ballance et al. (2006) suggest that seasonal patterns in biology have evolved to take advantage of oceanographic conditions in both the feeding and breeding grounds for migratory marine species. Inter-annually, oceanographic variation is dominated by the ENSO in 2-7 year cycles (Wang and Fiedler 2006). El Niño effects such as distributional shifts, diet changes and reproductive failures have been noted for apex predators in the California and Peru Current systems, but less is known about the effects in the ETP (Ballance et al. 2006). ENSO is known to have catastrophic effects on sea turtle nesting frequency and abundance (Seminoff et al. 2008), but little is known about effects on oceanic activities. Therefore, more research is needed to understand the effects of El Niño sea turtles in the ETP, and specifically how ENSO events impact preferable habitat, and in effect species distributions.
Appendix

MGET GLM and GAM GIS model builder:

[Fit GLM model: Presence = log(chl) + SST + Depth]

[Fit GAM model: Presence = s(log(chl)) + s(SST) + s(Depth)]
GLM model fit summary statistics:

Area under the ROC curve (auc) = 0.864244
Mean cross-entropy (mxe) = 0.506113
Precision-recall break-even point (prbe) = 0.780000
Root-mean square error (rmse) = 0.411006

Contingency table for cutoff = 0.296735:

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Actual</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>P</td>
<td>726</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>199</td>
</tr>
<tr>
<td></td>
<td></td>
<td>925</td>
</tr>
<tr>
<td>N</td>
<td>P</td>
<td>124</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>692</td>
</tr>
<tr>
<td></td>
<td></td>
<td>816</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>850</td>
</tr>
<tr>
<td></td>
<td>Actual</td>
<td>891</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1741</td>
</tr>
</tbody>
</table>

Accuracy (acc) = 0.814474
Error rate (err) = 0.185526
Rate of positive predictions (rpp) = 0.531304
Rate of negative predictions (rnp) = 0.468696

True positive rate (tpr, or sensitivity) = 0.854118
False positive rate (fpr, or fallout) = 0.223345
True negative rate (tnr, or specificity) = 0.776655
False negative rate (fnr, or miss) = 0.145882

Positive prediction value (ppv, or precision) = 0.784865
Negative prediction value (npv) = 0.848039
Prediction-conditioned fallout (pcfall) = 0.215135
Prediction-conditioned miss (pcmiss) = 0.151961

Matthews correlation coefficient (mcc) = 0.631838
Odds ratio (odds) = 20.359540
SAR = 0.696575

GLM model test summary statistics:

Area under the ROC curve (auc) = 0.782810
Mean cross-entropy (mxe) = 0.572139
Precision-recall break-even point (prbe) = -0.000000
Root-mean square error (rmse) = 0.435240

Contingency table for cutoff = 0.532561:

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Actual</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>P</td>
<td>251</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td></td>
<td>327</td>
</tr>
<tr>
<td>N</td>
<td>P</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>273</td>
</tr>
<tr>
<td></td>
<td></td>
<td>371</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>349</td>
</tr>
<tr>
<td></td>
<td>Actual</td>
<td>349</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>698</td>
</tr>
</tbody>
</table>

Accuracy (acc) = 0.750716
Error rate (err) = 0.249284
Rate of positive predictions (rpp) = 0.468481
Rate of negative predictions (rnp) = 0.531519

True positive rate (tpr, or sensitivity) = 0.719198
False positive rate (fpr, or fallout) = 0.217765
True negative rate (tnr, or specificity) = 0.782235
False negative rate (fnr, or miss) = 0.280802
Positive prediction value (ppv, or precision) = 0.767584
Negative prediction value (npv) = 0.735849
Prediction-conditioned fallout (pcfall) = 0.232416
Prediction-conditioned miss (pcmiss) = 0.264151
Matthews correlation coefficient (mcc) = 0.502432
Odds ratio (odds) = 9.200188
SAR = 0.656255

GLM model summary:

Formula:
Presence ~ log(Chl) + SST + Depth, family = binomial(link="logit")

Deviance Residuals:
       Min         1Q      Median         3Q         Max
-2.74857   -0.90151    0.07662    0.82813    2.81120

Coefficients:
             Estimate    Std. Error  z value   Pr(>|z|)  
(Intercept)  -2.212e+01   2.574e+00   -8.596  < 2e-16 ***
log(Chl)     1.294e+00   2.194e-01    5.897   3.71e-09 ***
SST          8.248e-01   8.540e-02    9.658  < 2e-16 ***
Depth       -9.459e-06   8.475e-05   -0.112     0.911

---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 967.63  on 697  degrees of freedom
Residual deviance: 798.71  on 694  degrees of freedom

AIC: 806.7
Number of Fisher Scoring iterations: 4
Analysis of Deviance Table
Model: binomial, link: logit
Response: Presence

Terms added sequentially (first to last):

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance Resid.</th>
<th>Df Resid.</th>
<th>Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>NULL</td>
<td>967.63</td>
<td>697</td>
<td>967.63</td>
</tr>
<tr>
<td>log(Chl)</td>
<td>51.62</td>
<td>696</td>
<td>916.01</td>
</tr>
<tr>
<td>SST</td>
<td>117.29</td>
<td>695</td>
<td>798.72</td>
</tr>
<tr>
<td>Depth</td>
<td>0.01</td>
<td>694</td>
<td>798.71</td>
</tr>
</tbody>
</table>

GAM model fit summary statistics:

Area under the ROC curve (auc) = 0.641622
Mean cross-entropy (mxe) = 0.853399
Precision-recall break-even point (prbe) = -0.001412
Root-mean square error (rmse) = 0.527586
Contingency table for cutoff = 0.399335:
<table>
<thead>
<tr>
<th></th>
<th>Actual P</th>
<th>Actual N</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted P</td>
<td>557</td>
<td>365</td>
<td>922</td>
</tr>
<tr>
<td>Predicted N</td>
<td>293</td>
<td>526</td>
<td>819</td>
</tr>
<tr>
<td>Total</td>
<td>850</td>
<td>891</td>
<td>1741</td>
</tr>
</tbody>
</table>

Accuracy (acc) = 0.622056  
Error rate (err) = 0.377944  
Rate of positive predictions (rpp) = 0.529581  
Rate of negative predictions (rnp) = 0.470419  

True positive rate (tpr, or sensitivity) = 0.655294  
False positive rate (fpr, or fallout) = 0.409652  
True negative rate (tnr, or specificity) = 0.590348  
False negative rate (fnr, or miss) = 0.344706  

Positive prediction value (ppv, or precision) = 0.604121  
Negative prediction value (npv) = 0.642247  
Prediction-conditioned fallout (pcfall) = 0.395879  
Prediction-conditioned miss (pcmiss) = 0.357753  

Matthews correlation coefficient (mcc) = 0.246005  
Odds ratio (odds) = 2.739558  
SAR = 0.597088

GAM model test summary statistics:

Area under the ROC curve (auc) = 0.880494  
Mean cross-entropy (mxe) = 0.435228  
Precision-recall break-even point (prbe) = 0.785100  
Root-mean square error (rmse) = 0.375005

Contingency table for cutoff = 0.566241:

<table>
<thead>
<tr>
<th></th>
<th>Actual P</th>
<th>Actual N</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted P</td>
<td>267</td>
<td>52</td>
<td>319</td>
</tr>
<tr>
<td>Predicted N</td>
<td>82</td>
<td>297</td>
<td>379</td>
</tr>
<tr>
<td>Total</td>
<td>349</td>
<td>349</td>
<td>698</td>
</tr>
</tbody>
</table>

Accuracy (acc) = 0.808023  
Error rate (err) = 0.191977  
Rate of positive predictions (rpp) = 0.457020  
Rate of negative predictions (rnp) = 0.542980  

True positive rate (tpr, or sensitivity) = 0.765043  
False positive rate (fpr, or fallout) = 0.148997  
True negative rate (tnr, or specificity) = 0.851003  
False negative rate (fnr, or miss) = 0.234957  

Positive prediction value (ppv, or precision) = 0.836991  
Negative prediction value (npv) = 0.783641  
Prediction-conditioned fallout (pcfall) = 0.163009  
Prediction-conditioned miss (pcmiss) = 0.216359  

Matthews correlation coefficient (mcc) = 0.618335
Odds ratio (odds) = 18.597326
SAR = 0.687840

GAM model summary:

Family: binomial
Link function: logit

Formula:
Presence ~ s(log(Chl)) + s(SST) + s(Depth)

Parametric coefficients:

| Parameter  | Estimate | Std. Error | z value | Pr(>|z|) |
|------------|----------|------------|---------|----------|
| (Intercept)| -0.2045  | 0.1200     | -1.704  | 0.0884   |

---

Signif. codes:  0 **** 0.001 *** 0.01 ** 0.05 * 0.1 '. ' 1

Approximate significance of smooth terms:

<table>
<thead>
<tr>
<th>Smooth Term</th>
<th>edf Ref.</th>
<th>df</th>
<th>Chi.sq</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>s(log(Chl))</td>
<td>8.847</td>
<td>8</td>
<td>8.09</td>
<td>3.26e-15 ***</td>
</tr>
<tr>
<td>s(SST)</td>
<td>7.523</td>
<td>110.40</td>
<td>&lt; 2e-16 ***</td>
<td></td>
</tr>
<tr>
<td>s(Depth)</td>
<td>7.990</td>
<td>38.55</td>
<td>5.89e-06 ***</td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 **** 0.001 *** 0.01 ** 0.05 * 0.1 '. ' 1

R-sq.(adj) = 0.417 Deviance explained = 37.2%
UBRE score = -0.056877 Scale est. = 1 n = 698

Maxent data preparation GIS script tool:

# MaxentPrep.py
# Created December 2009
# Created by Lindsey Peavey
# MaxentPrep.py
##
## Description: Resamples and clips environmental rasters by a defined mask
## raster so that all rasters have the exact same extent and cell
## size. Creates a feature class of species observation data points.
## The resultant feature class and rasters layers can be used
## as inputs into ArcMaxEnt.py, ArcMaxEnt_edited1.py, or
## ArcMaxEnt_edited2.py.
##
## Requirements: Workspace path
## Path to a mask file
## Environmental raster layers
## Output file paths
## Species Observation xy data in csv file
##
## (C) December 2009 Lindsey Peavey
Maxent open-source GIS data analysis tool:

# Import system modules
import sys, string, os, arcgis scripting

# Create the Geoprocessor object
gp = arcgis scripting.create()

# Set geoprocessing properties
gp.OverwriteOutput = 1   # Enables files to be overwritten

# Check out any necessary licenses
gp.CheckExtension("spatial")

# Load required toolboxes
gp.AddToolbox("C:\Program Files\ArcGIS\ArcToolbox\Toolboxes\Spatial Analyst Tools.tbx")
gp.AddToolbox("C:\Program Files\ArcGIS\ArcToolbox\toolboxes\Data Management Tools.tbx")

# User-defined variables
gp.workspace = sys.argv[1]   # set the geoprocessing environment
mask = sys.argv[2]          # set mask raster
str_envrasters_in = sys.argv[3]   # define input rasters
str_envrasters_out = sys.argv[4]   # define output paths
in_obspoints = sys.argv[5]   # provide a csv file with xy location data

# check spatial reference and describe the extent and cell size of the mask using the describe function
desc = gp.describe(mask)
spatialReference = desc.SpatialReference.Name
gp.extent = desc.Extent
gp.cellsize = str(desc.MeanCellWidth)
gp.SnapRaster = mask

# iterate through all environmental rasters to snap and clip to the mask

# set variables
rasters_in = str_envrasters_in.split(';')
rasters_out = str_envrasters_out.split(';')

# Loop through Geoprocesses
for i in range(len(rasters_in)):
    in_raster = rasters_in[i]
    out_raster = rasters_out[i]
    gp.ProjectRaster_management(in_raster, "env_project", desc.SpatialReference)   # Project rasters
    gp.Resample_management("env_project", "env_resample", gp.cellsize, "NEAREST")   # Resample rasters
    gp.ExtractByMask("env_resample", mask, out_raster)   # Extract by mask

# generate a feature class containing observation point data

### ArcMaxent.py
###
### Description: Runs MAXENT species modeling software from ArcGIS for
### a set of species locations and environmental layers. This
### script converts the points to a csv file in MaxEnt format
### and converts the environmental raster layers to ASCII format
### datasets. It then forms a MaxEnt batch file that can either
### be run immediately or run later manually.
###
### Requirements: Species Observation points
### Environmental raster layers
### Path to MaxEnt JAR file
### Results folder location.
###
### (C) 2008 John Fay
### updated 2009 Lindsey Peavey
### Nicholas School of the Environment, Duke University
###
```python
import sys, os, arcgisscripting, time
gp = arcgisscripting.create()

# Input variables, if running in Pythonwin input entire paths
obsPointsFC = sys.argv[1]
envrasters = sys.argv[2]
maxEntDir = sys.argv[3]
outFolder = sys.argv[4]
runnow = sys.argv[5]

# script variables, user-defined spname
spname = sys.argv[6] # must put in quotes, ex. "olivenidley"
maxEntDir = os.path.dirname(maxEntDir)

mempAlloc = str(312)
samplesFile = os.path.join(outFolder, 'samples.csv')

def msg(msgText, severity=0):
    now = '%s' % time.ctime()
    print now, msgText
    if severity == 1:
        gp.AddWarning(msgText)
    elif severity == 2:
        gp.AddError(msgText)
    else:
        gp.AddMessage(msgText)
    return

# Create the output folder
if os.path.exists(outFolder):
    msg('Output folder already exists.',1)
else:
    msg('Creating output folder %s' % outFolder)
    os.mkdir(outFolder)

# Convert GRIDS to ASCII
for raster in envrasters.split(' '):
    raster_catPath = gp.describe(raster).CatalogPath
    raster_basename = os.path.basename(raster_catPath)
    ascii = os.path.join(outFolder, raster_basename + '.asc') # must be Arc GRID...
    if not os.path.exists(ascii):
        msg('Converting %s to ASCII %s
    gp.RasterToASCII_conversion(raster, ascii)
    else:
        msg(' %s already exists. skipping %s
```
## Initialize the Maxent formatted samples file

```python
# create Maxent formatted samples file
samples = open(samples_file, 'w+)
samples.write('Species, xcoord, ycoord\n')
```

## Convert observation points to Maxent formatted samples file

```python
# convert observation points to Maxent formatted samples file
recs = gq.searchCursor(obspointsRC)
rec = rec.next()
while rec:
    shape = rec.shape.getPart(0)
    x = shape.x
    y = shape.y
    samples.write('X: %f, Y: %f, %s
' % (x, y, shape.crs))
    rec = rec.next()
samples.close()
```

## Generate Maxent command

```bash
# generate Maxent command
batchstring = java -jar maxent.jar -N outfolder
batchstring += -n -s samplesfile
batchstring += -x outputdirectory
batchstring += --noclasses
batchstring += --parameters
batchstring += --setparam
batchstring += writeoutputdata
```

## Run Maxent

```python
if run:
    batchstring += '-a'  # run Maxent
    batchstring += ' -n'  # no warnings
    os.chmod(maxentBin, 'x')
    print('Running Maxent with options: %s' % batchstring)
    os.system(batchstring)
    print('Processing complete. Preparing results')
    os.system('calcStatOut %s %s %s' % (outfolder, samplesfile, maxentBin))
```

## Calculate statistics

```bash
# calculate statistics
batchstring = os.path.join(outfolder, 'RUNMAXENT.bat')
bash = open(batchfile, 'w')
bash.write(batchstring)
bash.close()
```

## Maxent processing complete

```python
print('Processing complete for %s' % samplesfile)
```
Screenshots of the Maxent java interface and the user-specified constraints exactly as the model was run for this analysis:

```
<table>
<thead>
<tr>
<th>Samples</th>
<th>Environment layers</th>
</tr>
</thead>
<tbody>
<tr>
<td>File: Peavy/Maxent_run/inutsamples.csv</td>
<td>Directory: Peavy/Maxent_run/inutsbackground.csv</td>
</tr>
<tr>
<td>LO</td>
<td>chl</td>
</tr>
<tr>
<td>depth</td>
<td>Continuous</td>
</tr>
<tr>
<td>distance</td>
<td>Continuous</td>
</tr>
<tr>
<td>sst</td>
<td>Continuous</td>
</tr>
</tbody>
</table>
```

- **Linear features**
- **Quadratic features**
- **Product features**
- **Threshold features**
- **Hinge features**
- **Auto features**

**Create response curves**
**Make pictures of predictions**
**Do jackknife to measure variable importance**

**Output format** Logistic
**Output file type** col

**Output directory** F:\Dupe files\Peavy\Maxent_runs
**Projection layers directory/file** \w\Peavy/Maxent_runs\inputs\env_asc_oddist
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threads</td>
<td>2</td>
</tr>
<tr>
<td>Log to log threshold</td>
<td>30</td>
</tr>
<tr>
<td>Linear to log threshold</td>
<td>10</td>
</tr>
<tr>
<td>Hinge threshold</td>
<td>15</td>
</tr>
<tr>
<td>Beta threshold</td>
<td>-1</td>
</tr>
<tr>
<td>Beta categorical</td>
<td>-1</td>
</tr>
<tr>
<td>Beta log</td>
<td>-1</td>
</tr>
<tr>
<td>Beta hinge</td>
<td>-1</td>
</tr>
<tr>
<td>Default nodata value</td>
<td>-9999</td>
</tr>
</tbody>
</table>
References


Pitman, R.L. 1986. Atlas of seabird distribution and relative abundance in the eastern tropical Pacific. Administrative Report LJ-86-02C. Available from Southwest Fisheries Science Center, 8604 La Jolla Shores Drive, La Jolla, California, 92037, USA.


Acknowledgements

My data collection would not have been possible without the support of the Southwest Fisheries Science Center (SWFSC) and the entire crew of the NOAA R/V David Starr Jordan (may she rest in peace), specifically Chico Gomez, Jose Coito, Joaoh, and Victor. I would like to sincerely thank all of the many scientists who assisted me in the field, especially Bob Pitman, Rich Pagen and Chris Cutler. I owe special thanks to Dr. Jeffrey Seminoff for his perpetual training and advice, the entire Marine Turtle Research Program at the SWFSC, Tomo Eguchi, Tim Gerrodette, and of course Annette Henry and Dr. Lisa Ballance for making it all happen. Thank you to Wallace “J.” Nichols for his continual moral support, encouragement and inspiration.

My study at Duke University was supported by the Doris Duke Conservation Fellowship and the George Burlew Scholarship Programs. I am extremely grateful to Dr. Dean Urban, Dr. Song Qian, and Ben Best for generously advising my data analysis.

Thank you to Dr. Larry Crowder, Dr. Andy Read, and Ben Best for their editorial contributions. Dr. Dean Urban has provided me with invaluable professional guidance and mentorship. I could not have survived the past two challenging years without my many cherished friends, my supportive family (Heather, Chris, Isabella, Sophia, Mom and Clair), and my daily solace Ben Best.