GENERALIZED ADDITIVE MODELS FOR PREDICTING THE SPATIAL DISTRIBUTION OF BILLFISHES AND TUNAS ACROSS THE GULF OF MEXICO

Holly A. Perryman[1], Elizabeth A. Babcock[2]

SUMMARY

Generalized additive models (GAMs) were developed to predict the spatial distributions of billfish and tuna species across the Gulf of Mexico. Models were fitted with data from NOAA's Pelagic Longline Observer Program (2005-2010). A delta approach was followed – fitting a Bernoulli GAM with binomial data, and a Gamma GAM with zero-truncated catch rate data [number of organisms per 100 hooks]. Model descriptors included year, season, day/night, sea bottom depth, altimetry, sea surface temperature, and minimum distance from a front. Basis dimensions of smoothing splines were iteratively adjusted. Models fitted for Istiophorus albicans *(sailfish), indicate that both the probability of catching a sailfish and the CPUE are most influenced by sea bottom depth and sea surface temperature. The Bernoulli model explained 14.6% of deviance while the Gamma model explained 48.6% of deviance. Fitted models were used to develop seasonal distribution profiles across the Gulf's open ocean by predicting across grids of environmental data from NCEI and AVISO. Profiles for sailfish indicate a seasonal flux, with increased CPUE between April and September, and higher catch rates associated to fronts.*

RÉSUMÉ

Des modèles additifs généralisés (GAM) ont été élaborés afin de prévoir les distributions spatiales des istiophoridés et des thonidés dans le golfe du Mexique. Les modèles ont été ajustés avec des données obtenues dans le cadre du Programme d'observateurs à bord palangriers pélagiques de la NOAA (2005-2010). On a appliqué une approche delta ajustant un GAM Bernoulli avec des données binomiales et un GAM Gamma avec des données de taux de capture tronquées à zéro [nombre de spécimens / 100 hameçons). Les descripteurs de modèle incluaient : année, saison, jour / nuit, profondeur du fond marin, altimétrie, température de surface de la mer et distance minimale par rapport au front. Les dimensions de base des splines de lissage ont été ajustées de manière itérative. Les modèles ajustés à l'Istiophorus albicans (voilier) indiquent que tant la probabilité de capturer un voilier que la CPUE sont principalement influencées par la profondeur du fond marin et la température de surface de la mer. Le modèle Bernoulli expliquait 14,6 % de l'écart alors que le modèle Gamma expliquait 48,6% de l'écart. Les modèles ajustés ont été utilisés pour élaborer des profils de distribution saisonnière pour l'ensemble du golfe au moyen de prévisions en grilles de données environnementales provenant de NCEI et de AVISO. Les profils dans le cas du voilier indiquent un flux saisonnier, avec une augmentation de la CPUE entre avril et septembre, et des taux de capture plus élevés associés aux fronts.

SUMMARY

Se desarrollaron modelos aditivos generalizados (GAM) para predecir la distribución espacial de las especies de istiofóridos y túnidos en el golfo de México. Los modelos se ajustaron con datos del programa de observadores de palangre pelágico de NOAA (2005-2010). Se siguió un enfoque delta ajustando un GAM Bernoulli con datos binomiales y un GAM Gamma con datos de tasa de captura truncados en cero (número de organismos por 100 anzuelos). Los descriptores del modelo fueron año, temporada, día/noche, profundidad del fondo del mar, altimetría, temperatura de la superficie del mar y distancia mínima desde un frente. Se ajustaron iterativamente las dimensiones básicas de las funciones alisadoras spline. Los modelos ajustados para el pez vela (Istiophorus albicans) indicaban que la profundidad del fondo del mar y la temperatura de la superficie del mar eran los factores que más influían tanto

[1] Rosenstiel School of Marine and Atmospheric Science, 4600 Rickenbacker CSWY, Miami, FL, 33149. hperryman@rsmas.miami.edu
[2] Rosenstiel School of Marine and Atmospheric Science, 4600 Rickenbacker CSWY, Miami, FL, 33149.

en la probabilidad de capturar un pez vela como en la CPUE. El modelo Bernoulli explicó el 14,6 % de la desviación, mientras que el modelo Gamma explicó el 48,6% de la desviación. Los modelos ajustados se utilizaron para desarrollar perfiles de distribución estacional en el mar abierto del Golfo mediante la predicción en cuadrículas de datos medioambientales de NCEI y AVISO. Los perfiles para el pez vela apuntaban a un flujo estacional, con una mayor CPUE de pez vela entre abril y septiembre y tasas de captura más elevadas asociadas con frentes.

KEYWORDS

Mathematical model, Catch/effort, Abundance, Spatial variations

1. Introduction

The 2009 Sailfish Assessment found that stocks may have been reduced below B_{MSY} and there is a higher degree of uncertainty in the status of the western stock (ICCAT 2010). Atlantic sailfish (*Istiophorus albicans*) are an ecologically and economically important species, especially in the Gulf of Mexico. Sailfish were once one of the most sought-after sport fisheries and are still caught by recreational fishers (Prince *et al*. 2007). Sailfish larvae have been found in the Gulf of Mexico at high densities, and research suggests that sailfish frequent the Gulf seasonally for spawning (Wells and Rooker 2009; Simms et al. 2010; Rooker *et al*. 2012). However, the fine-scale spatial and temporal distribution of older individuals is not well understood.

Here we present work done to fit Generalized Additive Models (GAMs) to predict the catch rates of billfish and tuna species within the Gulf of Mexico. Catch rates are used as an indicator of abundance (Hilborn and Walters 1992). Efforts were made to optimize the fits of GAM smoothing splines. Model fits are assessed using residual diagnostics as well as the receiver operating characteristic. Fitted GAMs were used to predict seasonal, Gulf-wide spatial distribution profiles. Profiles were constructed using grids of geographic coordinates representing seasonal averages of environmental conditions. The profiles for sailfish are presented here.

2. Methods and Materials

2.1 Data

Data for model fitting was collected from NOAA's Pelagic Observer Program dataset. The Pelagic Observer Program (Beerkircher *et al.* 2002; Brooke 2012) collects catch data from observers aboard vessels in the U.S. commercial pelagic longline fleet. Vessels suspend longline gear mid-depth, approximately 33-66 m, but the actual fishing depth is unknown due to the influences by currents and environmental conditions (Beerkircher *et al.* 2002). Data were restricted to be within the Gulf of Mexico between 2005 and 2010. Models were fit for the following species identified in the dataset: *Istiophorus albicans*, *Istiophoridae* spp., *Makaira nigricans*, *Tetrapturus albidus*, *Tetrapturus* spp., *Thunnus albacares*, *Thunnus thynnus*, and *Xiphias gladius*. Environmental and temporal variables were used as model descriptors.

Temporal variables included year, season, and daytime (day / night). Season was broken down into four categories: 1 (Jan. - Mar.), 2 (Apr. - Jun.), 3 (Jul. - Sep.), or 4 (Oct. - Dec.). Environmental variables included sea surface temperature (SST) °C, altimetry (SSH) m, sea bottom depth (SBD) m, and minimum distance from a front (MDF) m. The observer dataset includes measurements of SST and SBD throughout the longline set and haul. Single estimates of these variables for each longline event were generated by averaging the multiple measurements. Approximately 5% of the observer records lacked measurements of SST, so the *Interpolate PO.DAAC MODIS L3 SST at Points* tool from the Marine Geospatial Ecology Tools (MGET) toolbox (Roberts *et al.* 2010) was used in *ArcGIS* to generate estimates. Approximately 9% of the observer records were dropped from this analysis since necessary information was missing (e.g., date).

Estimates of SSH and MDF were also used since top marine predators are known to aggregate around oceanographic features (Olson *et al.* 1994). SSH was calculated for individual catch records using the Archiving, Validation and Interpretation of Satellite Oceanographic (AVISO) dataset (Ducet *et al.* 2000). Estimates were the average of the four AVISO records nearest to the catch location, on the same date as the catch. Estimates of MDF were calculated for individual catch records using the *Cayula-Cornillon Fronts in ArcGIS Raster MGET* tool from the MGET toolbox in *ArcGIS.* This tool uses the Cayula and Cornillon (1992) algorithm for the detection and extraction of fronts. Cayula and Cornillon used sea surface temperature to detect fronts but the

Gulf of Mexico is known to have weak sea surface temperature gradients (Legeckis 1978). Thus, AVISO altimetry data were used to derive frontal features. *ArcGIS*'s Model Builder was used to develop a routine that systematically estimates MDF for catch records. The routine loops through each unique date in the AVISO dataset, generates the corresponding frontal features, and calculates the minimum distance between a front and a catch event corresponding to the date.

2.2 Model Description

Generalized additive models (GAMs) for estimating the abundance patterns of individual species were developed using the *gam()* function and the statistical software *R* (Wood 2004; R Development Core Team 2008; Wood 2011; Drexler and Ainsworth 2013; Grüss *et al*. 2014). Abundance was assumed to be proportional to catch-per-unit-effort (CPUE). CPUE was calculated as the total numbers of individuals caught per 100 hooks. A Delta approach was followed to account for the zero-inflated nature of the dataset. The Delta method calls for fitting two statistical models: one predicting the probability of catching an organism with binomial data, the other predicting the CPUE with zero-truncated data. Determining an appropriate error structure for generalized models is an important aspect of model construction (Maunder and Punt 2004). The Bernoulli distribution with a logit link function is used to model the error structure of the binomial data. Preliminary analyses indicated that the gamma distribution is more suitable for the non-zero CPUE data in this investigation. Other studies fitting generalized models have achieved comparable, if not improved, model fits using the gamma distribution (Punt *et al.* 2000; Ortiz *et al.* 2004). A log link function was used to ensure non-negative predictions (Wood 2007). A general form of the fitted GAMs is as follows:

$$\eta = f(year) + f(season) + f(day) + s(SST) + s(SSH) + s(SBD) + s(MDF)$$

where η is either the probability of retaining an organism or the mean non-zero CPUE, *f()* indicates factor descriptors, and *s()* indicate numerical descriptors processed with smoothing splines.

The construction of GAMs requires some attention towards developing robust smoothing splines for each numerical descriptor. This study uses penalized regression splines, which incorporate penalties to the least squares fitting objective based on the flexibility of a spline (Wood and Augustin 2002; Wood 2006). Penalized regression splines are controlled by the smoothing parameter, which controls the tradeoff between the model's fit and smoothness, and the basis dimension (*k*), which defines the maximum possible degrees of freedom (Guisan *et al.* 2002; Wood 2006). Smoothing parameters are automatically calculated by the *gam()* function based on the smoothness selection criterion: generalized cross-validation (GCV) criterion, or the Un-Biased Risk Estimator (UBRE) scores (Wood 2006). We allowed an additional penalty to completely removal of a term if the smoothing parameter equals zero (i.e., if the smoother does not improve model fit). Basis dimensions are not automatically calculated by the *gam()* function and must be specified. Pya and Wood (2016) concluded that the exact setting of a basis dimension is not crucial as long as it large enough to avoid over-smoothing / under-fitting, and highlighted that the simple test for assessing a basis dimension presented by Wood (2006) performs just as well as the complex, time expensive approaches. To improve the fitting of splines, we iteratively adjusted the basis dimension for each numerical descriptor. First, GAMs were fitted with all basis dimensions set to three, the minimum allowable setting. Then, considering smoothing splines one at a time and in sequential order, the basis dimension was tested following the Wood (2006) method to determine if increasing the value might improve the spline's fit. If not, basis dimensions were not adjusted. Otherwise, GAMs was re-fitted with the basis dimension set to the default value suggested by Kim and Gu (2004):$10n^{2/9}$, where *n* is the sample size. The basis dimension was then tested again. If the value did not pass the test then adjusting the basis dimension for that descriptor did not significantly assist in reducing the spline's residual variance, so the basis dimension was re-set to the larger of the following two values: 3, $\lceil edf \rceil + 1$ (where *edf* is the effective degrees of freedom). The latter value accounts for minor improvement in smoother fit. However, if the increase in basis dimension passed the test then the value was iteratively reduced to determine the smallest value for satisfying the test so degrees of freedom can be preserved.

Model fits were assessed with Pearson residuals, which standardize the raw residuals by dividing them by an estimate of the standard deviation of the observed value. Since GAMs allow the use of different error structures it is necessary to standardize residuals, otherwise residuals may be driven by error in the assumed mean-variance relationship instead of the model fit. Residual diagnostic plots considered for Gamma models include normal Q-Q, box-plot, residuals against linear predictor, and residuals against time. Assessing the residuals for Bernoulli models is complicated since a model produces predicted probabilities for a binary response variable. The detection of model inadequacy is often the main focus since non-constant variance is always present, and outliers

can be difficult to diagnose. The logistic model residuals against fitted values can provide some insight into model adequacy. An adequate logistic model should have residuals that produce a LOWESS (locally weighted scatterplot smoother) curve centered along the horizontal line with a zero intercept. The receiver operating characteristic area under the curve (AUC) metric for logistic models was also evaluated, since it is commonly used as a global indicator of logistic model performance (Greiner *et al.* 2000). The AUC is equal to the probability that the model will correctly identify the positive case when presented with a randomly chosen pair of cases in which one case is positive and one case is negative (Hanley and McNeil 1982). This study will follow the arbitrary AUC guidelines suggested by Swets (1988): 0.5 is non-informative, 0.5 - 0.7 are less accurate, 0.7 - 0.9 are moderately accurate, 0.9 - 1 are highly accurate, 1 is perfect.

2.3 Model Prediction

Seasonal, spatial abundance distribution profiles spanning the Gulf of Mexico's open ocean (areas where the bottom depth is >200m) were developed for *Istiophorus albicans* (sailfish) using the fitted GAMs. First, grids for creating distribution profiles were generated in *ArcGIS.* A grid of geographic coordinates (0.1° latitude by 0.1° longitude) spanning the area was created using the *Fishnet* tool. The grid was duplicated to create four versions, one for each season. Then, coordinates were assigned estimates of model descriptors. Seasonal data for the year 2010 were collected from AVISO and NCEI to provide estimates of model descriptors. Data were interpolated into raster files using the *Kriging* tool, and the *Cayula-Cornillon Fronts in ArcGIS Raster MGET* tool was used to develop average seasonal fronts for calculating MDF (**Figure 1**). The *Extract Values to Points* tool (set to bilinear interpolation) was used to assign estimates from raster files to each coordinate in the appropriate seasonal grid. The routine constructed in Model Builder (described in 2.1) was used to calculate MDF for each coordinate in the fishnet. A single bathymetry raster was used to generate SBD estimates for the four seasonal grids. For fishnets, year was set to the year the data came from (2010), season corresponded to the appropriate numerical identification (1-4), and day/night was assumed to be day (D). Lastly, abundance indices for sailfish were predicted at each geographic coordinate in these fishnets using the fitted delta GAMs to develop seasonal distribution profiles.

3. Results and Discussion

Fitted generalized additive models (GAMs) for each species are summarized in **Table 1**. The median model deviance explained by Bernoulli models is 17.3% (ranging from 9.95% to 29.7%). The Bernoulli model for sailfish is one that is explaining a smaller amount of variance (14.6%). Commonly, logistic models struggle to explain deviance since they produce predicted probabilities for a binary response variable. The logistic models for most species were primarily driven by sea surface temperature and sea bottom depth, including the sailfish model. The median model deviance explained by Gamma models is 31.25% (ranging from 8.77% to 52.9%). The Gamma model for sailfish is one that is explaining a higher amount of deviance (48.6%). Most of the Gamma models were primarily driven by sea bottom depth, but sea surface temperature and altimetry were also influential for some models. These results suggest that sailfish are highly influenced by sea bottom depth and sea surface temperature.

The general trends for Bernoulli model diagnostic plots are displayed in **Figure 1**. Most of the logistic models produce positive residuals with heavier tails than the negative residuals, and create LOWESS curves entirely within the negative region (**Figure 2a, b**). This indicates that these models are more successful at estimating low probabilities of catch for non-catch events than at estimating high probabilities of catch for catch events. This includes the logistic model for sailfish (**Figure 2a**). The opposite trend was observed for some species: negative residuals with heavier tails than the positive residuals, and create LOWESS curves almost entirely within the positive region (**Figure 2c**). These models tend to be more successful at estimating high probabilities of catch for catch events than at estimating low probabilities of catch for non-catch events. None of the logistic models produce a LOWESS curve flat and centered on the horizontal line with a zero intercept (which would indicate a good model), but the Receiver Operating Characteristic (ROC) curve and Area Under the ROC curve (AUC) value for the logistic models (**Figure 3**) indicate that the logistic models constructed here are all "moderately accurate". Thus, although there is room to improve the fit of the logistic models, the current fits are reasonable enough to use for predictive purposes.

Residual diagnostics for Gamma models are similar to one another, so residuals for the sailfish model are shown as a general example (**Figure 4**). Diagnostics suggest there are some issues to be aware of. The residual have an obvious right skew in their distribution (**Figure 4a, b**). The heaviness of the tails differs among fitted Gamma models. Thus, models struggle with under-predicting catch rates. There are potentially several outliers in the data

(**Figures 4b**), but some work should be done to determine if these are truly outliers (and should be removed from the statistical analysis) or leverage points (providing important information and should remain in the statistical analysis). The variance of the Pearson residuals appears to quickly increase with increasing linear predictor (**Figure 4c**). The trend is more obvious when considering linear predictors less than then -1.0. This issue may be resolved by using a different error structure and/or link, or by transforming the response variable. Some work has been done with this (Perryman, in progress) with no improvement on residual diagnostics. Fortunately, there is no obvious correlation between error terms over time (**Figure 4d**).

Descriptor fits for the GAMs fitted for sailfish are shown in Figure 5. Since sea surface temperature and sea bottom depth are two of the most influential descriptors for the sailfish delta models, interpretation will focus on smoothing splines for those two numerical descriptors. The probability of catching a sailfish decreases with sea bottom depth (**Figure 5a**). The probability of catching a sailfish increases with sea surface temperature to around 25°C, and then declines with higher temperatures (**Figure 5b**). This is very similar to results reported by Kerstetter *et al.* (2010), and Mourato *et al.* (2010). The positive CPUE of sailfish decreases with depth to about 1000 m (**Figure 5h**). Catch rates don't change much once the sea bottom depth goes beyond about 1000m. Sea surface temperature appears to have a dynamic influence on sailfish catch rates (**Figure 5i**). The most influential ranges are temperatures from 23°C to 25°C (the greatest negative influence on catch rates), and 30°C to 32°C (the greatest positive influence on catch rates). Notice some of the numerical descriptors (e.g., **Figure 5j**) produce nearly flat smoothing splines. This indicates that the descriptor has a limited effect in the model.

Seasonal spatial distribution maps for sailfish are shown in **Figure 6**. The predictions are based on data averaged across season – which likely has a fair amount a variation. Thus, results will be discussed broadly and qualitatively rather than quantitatively. Immediately, a strong seasonal change in abundance is apparent. Sailfish appear to be more abundant in the Gulf of Mexico during season 2 and 3 (i.e., from April to September). Although season was not statistically significant for the fitted sailfish GAMs (**Table 1**), the probability of catching a sailfish is higher for seasons 2 and 3 (**Figure 4f**), and the predicted positive CPUE is higher during seasons 2 and 3 (**Figure 4m**). These results support the argument that sailfish have increased abundance in the Gulf of Mexico from April to September. This could be due to a seasonal migration to Gulf of Mexico spawning grounds (de Sylva and Breder 1997; Simms *et al.* 2010).

The predictions for season 2 and 3 (**Figure 6b, c**) show increased abundance of sailfish around fronts corresponding to the respective season (**Figure 1j, k**). Pelagic predators are often associated to frontal features (Worm *et al*. 2005), and use them as spawning habitat (Bakun 2006). Simms *et al*. (2010) found that the highest density of sailfish larvae is often found within mesoscale frontal features. This could be due to physical processes, but Richardson *et al*. (2009) examined sailfish spawning around a cyclonic, submesoscal Florida Current frontal eddy, and concluded that sailfish spawn at frontal zones. Thus, if sailfish are migrating into the Gulf of Mexico during the spring and summer to spawn at/near frontal features, our delta GAMs are picking up that behavior.

Standard errors for the predictions made with the sailfish models are shown in **Figure 7**. Standard errors are consistently larger in the deepest areas of the Gulf. **Figure 8** shows density plots of the numerical data used to fit and predict with the fitted sailfish GAMs. Notice that the prediction fishnets contain areas exceeding the maximum depth observed in the fitting data. In addition, notice that splines for sea bottom depth (**Figure 1a, h**) have very wide 95% confidence intervals for larger values of sea bottom depth. Since sea bottom depth is a statistically important model descriptor, yet can introduce a large amount of error into model predictions, future use of these models should be limited to predicting across depth ranges comparable to the fitting data.

Statistical models for predicting the horizontal distribution of Atlantic sailfish provided information on environmental drivers, and the use of GAMs allowed us to determine critical ranges of influential descriptors, in addition to spatial patterns in abundance. Although model fits were not ideal, fits and predictions were supported by various findings in the literature. Thus, qualitative trends may be captured by model predictions. Statistical models will likely be improved with additional catch data as well as the inclusion of additional environmental drivers. For instance, research has found that dissolved oxygen may constrain sailfish movement (Ehrhardt and Fitchett 2006; Prince and Goodyear 2006). The Commission encouraged actions to reduce fishing mortality of sailfish from non-industrial, in addition to industrial, fisheries (ICCAT 2010), and this research could provide insight on areas where and when sailfish may aggregate in the Gulf of Mexico.

Acknowledgements

This research was funded by NOAA/SEAGRANT agreement number NA11OAR4170185 and the Cooperative Institute for Marine and Atmospheric Studies (CIMAS), Cooperative Institute of the University of Miami and the National Oceanic and Atmospheric Administration, cooperative agreement number NA10OAR4320143. We acknowledge NOAA's Southeast Fisheries Science Center for providing longline catch data, and Dr. David Gloeckner, Dr. Lawrence Beerkircher, Greta Wells, Charles Weber, and Dr. Lori Hale for assistance. We acknowledge all of the organizations that provided environmental data: Archiving, Validation and Interpretation of Satellite Oceanographic (AVISO), NASA, and NOAA's National Centers for Environmental Information (NCEI). We thank Michael Drexler and Dr. Arnaud Gruss for providing example R code. We thank Dr. Cameron Ainsworth, Dr. Donald Olson, Dr. Joseph Serafy, Dr. David Die, and Dr. Michael Schirripa for their thoughtful comments and feedback.

References

Bakun, A. 2006. Fronts and eddies as key structures in the habitat of marine fish larvae: opportunity, adaptive response and competitive advantage. *Sci. Mar.*, 70(S2): pp. 105-122.

Beerkircher, L.R., C.J. Brown, and D.W. Lee. 2002. SEFSC Pelagic Observer Program Data Summary for 1992-2000. NOAA Technical Memorandum NMFS-SEFC-486: 26 p.

Brooke, S.G. 2012. Federal fisheries observer programs in the United States: over 40 years of independent data collection. *Mar. Fish. Rev.*, 76: pp. 1-38.

Cayula, J.F. and Cornillon, P. 1992. Edge detection algorithm for SST images. *J. Atmos. Oceanic Technol.*, 9(1): pp. 67-80.

de Sylva, D.P. and Breder, P.R. 1997. Reproduction, gonad histology, and spawning cycles of north Atlantic billfishes (Istiophoridae). *Bull. Mar. Sci.*, 60(3): pp. 668-697.

Ducet, N., P. Y. Le Traon, and G. Reverdin. 2000. Global high-resolution mapping of ocean circulation from topex/poseidon and ers-1 and-2. *J. Geophys. Res. C: Oceans.*, 105(C8): pp. 19477-19498.

Drexler, M., and C.H. Ainsworth. 2013. Generalized additive models used to predict species abundance in the Gulf of Mexico: and ecosystem modeling tool. *PLoS one*, 8(5): e6445.

Ehrhardt, N.M. and Fitchett, M.D. 2006. On the seasonal dynamic characteristics of the sailfish, Istiophorus platypterus, in the eastern Pacific off Central America. *Bull. Mar. Sci.*, 79(3): pp. 589-606.

Guisan, A., Edwards, T.C., and Hastie, T. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecol. Modell.*, 157(2): pp. 89-100.

Greiner, M., Pfeiffer, D., and Smith, R.D. 2000. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Pre. Vet. Med.*, 45(1): pp. 23-41.

Grüss, A., M. Drexler, and C.H. Ainsworth. 2014. Using delta generalized additive models to produce distribution maps for spatially explicit ecosystem models. *Fish. Res.*, 159: pp. 11-24.

Hanley, J.A. and McNeil, B.J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1): pp. 29-36.

Hilborn, R., and Walters, C.J. 1992. Quantitative fisheries stock assessment: choice, dynamics and uncertainty. *Rev. Fish Biol. Fish.*, 2(2): pp. 177-178.

ICCAT. 2010. Report of the 2009 sailfish stock assessment. *Collect. Vol. Sci. Pap. ICCAT*, 65(5): pp. 1507-1632.

Kerstetter, D.W., Bayse, S.M. and Graves, J.E. 2010. Sailfish (*Istiophorus platypterus*) habitat utilization in the southern Gulf of Mexico and Florida Straits with implications on vulnerability to shallow-set pelagic longline gear. *Collect. Vol. Sci. Pap. ICCAT*, 65(5): pp. 1701-1712.

Kim, Y.J., and Gu, C. 2004. Smoothing spline Gaussian regression: more scalable computation via efficient approximation. *J. R. Stat. Soc. Series B Stat. Methodol.*, 66(2): pp. 337-356.

Legeckis, R. 1978. A survey of worldwide sea surface temperature fronts detected by environmental satellites. *J. Geophys. Res. C: Oceans.*, 83(C9): pp. 4501-4522.

Maunder, M.N., and Punt, A.E. 2004. Standardizing catch and effort data: a review of recent approaches. *Fish. Res.*, 70(2): pp. 141-159.

Mourato, B.L., Carvalho, F.C., Hazin, F.H., Pacheco, J.C., Hazin, H.G., Travassos, P. and Amorim, A.F. 2010. First observations of migratory movements and habitat preference of Atlantic sailfish, Istiophorus platypterus, in the southwestern Atlantic Ocean. *Collect. Vol. Sci. Pap. ICCAT*, 65(5): pp. 1740-1747.

Olson, D.B., Hitchcock, G.L., Mariano, A.J., Ashjian, C.J., Peng, G., Nero, R.W. and Podesta, G.P. 1994. Life on the edge: marine life and fronts. *Oceanography*, 7(2): pp. 52-60.

Ortiz, M., and Arocha, F. 2004. Alternative error distribution models for standardization of catch rates of non-target species from a pelagic longline fishery: billfish species in the Venezuelan tuna longline fishery. *Fish. Res.*, 70(2): pp. 275-297.

Prince, E.D. and Goodyear, C.P. 2006. Hypoxia-based habitat compression of tropical pelagic fishes. *Fisheries Oceanography*, 15(6): pp.451-464.

Prince, E.D., Snodgrass, D., Orbesen, E.S., Hoolihan, J.P., Serafy, J.E. and Schratwieser, J.E. 2007. Circle hooks, 'J'hooks and drop-back time: a hook performance study of the south Florida recreational live-bait fishery for sailfish, Istiophorus platypterus. *Fish. Manag. Ecol.*, 14(2): pp. 173-182.

Punt, A.E., Walker, T.I., Taylor, B.L., and Pribac, F. 2000. Standardization of catch and effort data in a spatially-structured shark fishery. *Fish. Res.*, 45(2): pp. 129-145.

Pya, N., and Wood, S.N. 2016. A note on basis dimension selection in generalized additive modelling. arXiv preprint arXiv:1602.06696.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN: 3-900051-07-0.

Roberts J.J., Best B.D., Dunn D.C., Treml E.A., Halpin P.N. 2010. Marine Geospatial Ecology Tools: An integrated framework for ecological geoprocessing with ArcGIS, Python, R, MATLAB, and C++. *Environ. Modell. Software.*, 25: 1197-1207. doi: 10.1016/j.envsoft.2010.03.029

Rooker, J.R., Simms, J.R., Wells, R.D., Holt, S.A., Holt, G.J., Graves, J.E. and Furey, N.B. 2012. Distribution and habitat associations of billfish and swordfish larvae across mesoscale features in the Gulf of Mexico. *PLoS one*, 7(4): p.e34180.

Richardson, D.E., Llopiz, J.K., Leaman, K.D., Vertes, P.S., Muller-Karger, F.E. and Cowen, R.K. 2009. Sailfish (*Istiophorus platypterus*) spawning and larval environment in a Florida Current frontal eddy. *Prog. Oceanogr.*, 82(4): pp. 252-264.

Simms, J.R., Rooker, J.R., Holt, S.A., Holt, G.J. and Bangma, J. 2010. Distribution, growth, and mortality of sailfish (*Istiophorus platypterus*) larvae in the northern Gulf of Mexico. *Fish. Bull.*, 108(4): pp.478-490.

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

Wells, R.J.D. and Rooker, J.R. 2009. Feeding ecology of pelagic fish larvae and juveniles in slope waters of the Gulf of Mexico. *J. Fish. Biol.*, 75(7): pp.1719-1732.

Wood, S.N. 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Am. Stat. Assoc.*, 99: pp.673-686.

Wood, S.N. 2006. Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC.

Wood, S.N. 2007. [R] GLM dist Gamma-links identity and inverse. Online posting. https://stat.ethz.ch/pipermail/r- help/2007-June/134424.html

Wood, S.N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Series B Stat. Methodol.*, 73(1): pp. 3-36.

Wood, S.N., and Augustin, N.H. 2002. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecol. Modell.*, 157(2): pp.157-177.

Worm, B., Sandow, M., Oschlies, A., Lotze, H.K. and Myers, R.A. 2005. Global patterns of predator diversity in the open oceans. *Science*, 309(5739): pp.1365-1369.

Table 1. Summary of fitted Generalized Additive Models. Percent deviance explained refers to the whole model, *p* refers to the p-value of the indicated descriptor, and *k* refers to the basis dimension set for the indicated spline.

Bernoulli Model Descriptors	***Istiophorus albicans***	***Istiophoridae***	***Makaira nigricans***	***Tetrapturus albidus***	***Tetrapturus***	***Thunnus albacares***	***Thunnus thynnus***	***Xiphias gladius***
p, Intercept	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
p, Year (2006)	0.927	0.019	0.241	0.024	0.378	0.894	0.100	0.115
p, Year (2007)	0.009	0.936	0.412	< 0.001	0.045	0.690	0.656	0.223
p, Year (2008)	0.495	0.066	0.250	< 0.001	< 0.001	0.293	< 0.001	< 0.001
p, Year (2009)	0.104	0.003	0.222	< 0.001	< 0.001	0.343	0.033	< 0.001
p, Year (2010)	0.597	0.036	0.621	< 0.001	0.005	< 0.001	0.250	< 0.001
p, Season (2)	0.230	0.019	0.317	0.009	0.514	0.007	< 0.001	0.723
p, Season (3)	0.216	< 0.001	0.084	0.084	0.065	0.269	1.0	0.009
p, Season (4)	0.157	0.078	0.939	< 0.001	0.198	0.048	< 0.001	< 0.001
p, Daytime (N)	0.420	< 0.001	0.161	0.037	0.006	< 0.001	0.039	0.021
p, s(SST)	< 0.001	0.006	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
k, s(SST)	5	7	4	4	4	8	7	9
p, s(SSH)	0.008	0.008	< 0.001	0.044	0.025	< 0.001	< 0.001	< 0.001
k, s(SSH)	5	6	7	36	3	3	11	5
p, s(SBD)	< 0.001	0.020	0.005	0.005	0.083	< 0.001	< 0.001	< 0.001
k, s(SBD)	5	9	3	8	8	20	7	14
p, s(MDF)	0.030	0.082	0.091	< 0.001	0.958	< 0.001	0.237	0.592
k, s(MDF)	5	3	3	10	3	6	15	6
% Deviance Explained	14.6	10.7	9.96	19.3	15.7	29.7	20.5	18.9
n	2995	2995	2995	2995	2998	2995	2995	2995
Gamma Model Descriptors								
p, Intercept	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	0.997	< 0.001	0.039
p, Year (2006)	0.999	< 0.001	0.039	0.015	0.809	< 0.001	0.011	0.083
p, Year (2007)	0.762	0.105	0.008	0.210	0.018	0.182	0.101	0.015
p, Year (2008)	0.002	0.002	0.028	0.766	0.005	0.125	0.001	< 0.001
p, Year (2009)	0.070	< 0.001	0.054	0.362	< 0.001	0.111	< 0.001	< 0.001
p, Year (2010)	0.002	0.117	0.047	0.172	0.012	0.731	0.010	< 0.001
p, Season (2)	0.491	0.910	0.240	0.374	0.653	< 0.001	0.002	0.029
p, Season (3)	0.510	0.798	0.250	0.967	0.126	< 0.001	NA	0.237
p, Season (4)	0.513	0.880	0.052	0.724	0.732	0.013	0.356	0.912
p, Daytime (N)	< 0.001	0.115	0.118	0.361	0.779	< 0.001	0.407	< 0.001
p, s(SST)	0.010	0.002	< 0.001	0.007	< 0.001	< 0.001	0.291	< 0.001

k, s(SST)		24	3	5	3	8	7	3	20
p, s(SSH)		1.0	0.344	< 0.001	0.095	0.004	< 0.001	< 0.001	< 0.001
k, s(SSH)		3	12	3	20	7	22	5	9
p, s(SBD)		< 0.001	< 0.001	< 0.001	< 0.001	0.011	< 0.001	0.012	< 0.001
k, s(SBD)		5	21	7	12	3	14	4	45
p, s(MDF)		0.207	0.010	0.097	0.103	0.756	0.004	0.023	< 0.001
k, s(MDF)		3	9	5	3	3	13	3	20
% Deviance Explained		48.6	26.6	27.8	34.7	35.2	16.6	8.77	52.9
n		385	307	564	283	267	2589	605	2407

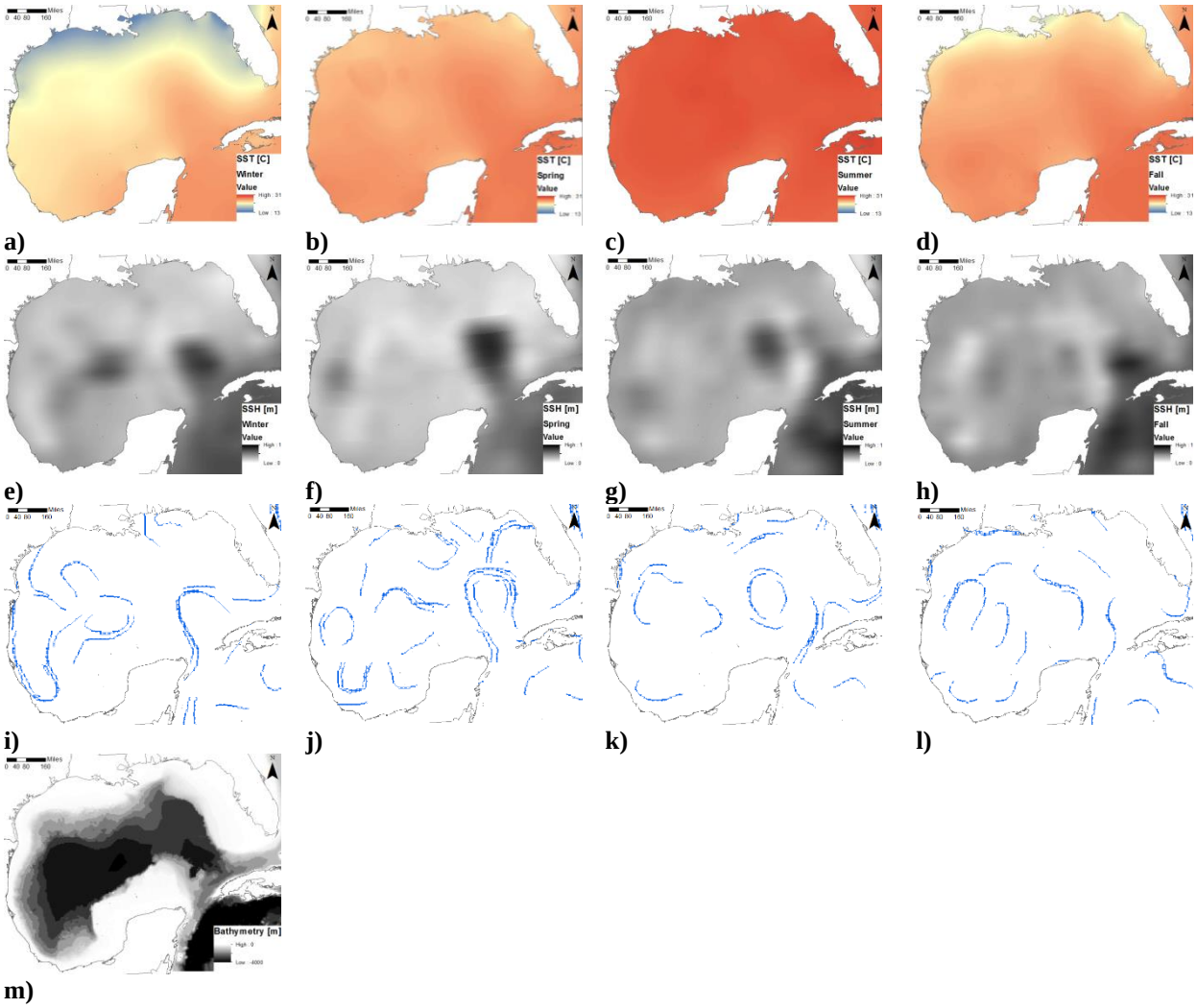
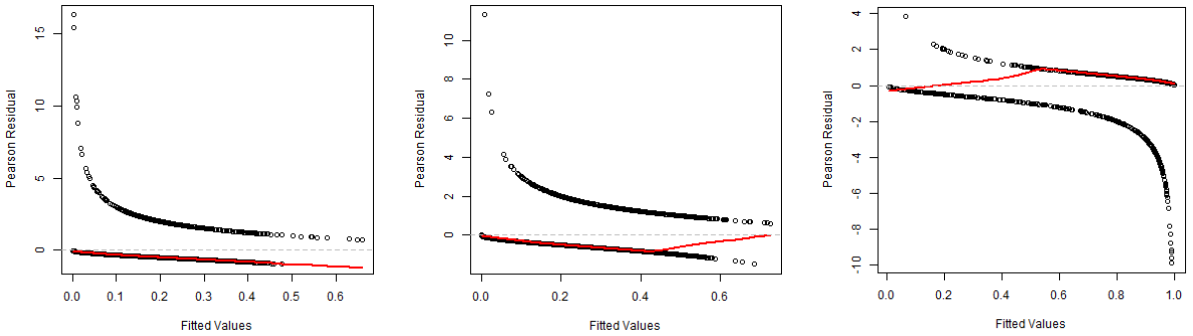


Figure 1. Files developed in *ArcGIS* for creating fishnets to use for model predictions. For panels (a) – (l), columns 1 to 4 corresponds to seasons 1 to 4. Panels (a) – (d) are sea surface temperature (°C), panels (e) – (h) are altimetry (m), panels (i) – (l) display the polyline files for estimating minimum distance from a front (m), and panel (m) displays the bathymetry (m) raster which was used for all fishnets. For fishnets, year was set to the year the data was collected (2010), and day/night was assumed to be day (D).



a) *Istiophorus albicans* **b)** *Thunnus thynnus* **c)** *Thunnus albacares*

Figure 2. General trends of Bernoulli model diagnostic plots: Pearson residuals against fitted values. Model plots for *Istiophorus albicans*, *Istiophoridae* spp., *Makaira nigricans*, *Tetrapturus albidus*, and *Tetrapturus* spp. resemble the plot displayed in panel (a). Panel (b) displays the diagnostic for *Thunnus thynnus*. Model plots for *Thunnus albacares* and *Xiphias gladius* resemble the plot displayed in panel (c). The black circles indicate Pearson residuals, the red line indicates the computed LOWESS curve, and the grey-dashed line indicates the zero horizontal axis (the ideal location of the LOWESS curve).

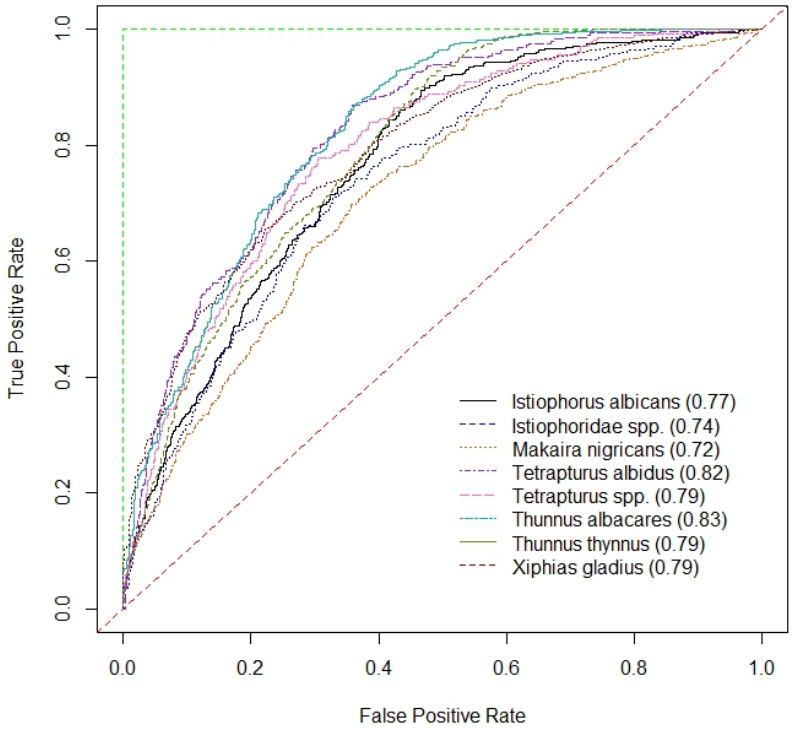


Figure 3. Receiver Operating Characteristic (ROC) curve for all fitted Bernoulli models. Values included in the legend are the corresponding Area Under the ROC curve (AUC) value. The red, dashed line is a hypothetical ROC curve with an AUC approximately equal to 0.5 (a non-informative model), and the green, dashed line is a hypothetical ROC curve with an AUC approximately equal to 1 (a perfect model).

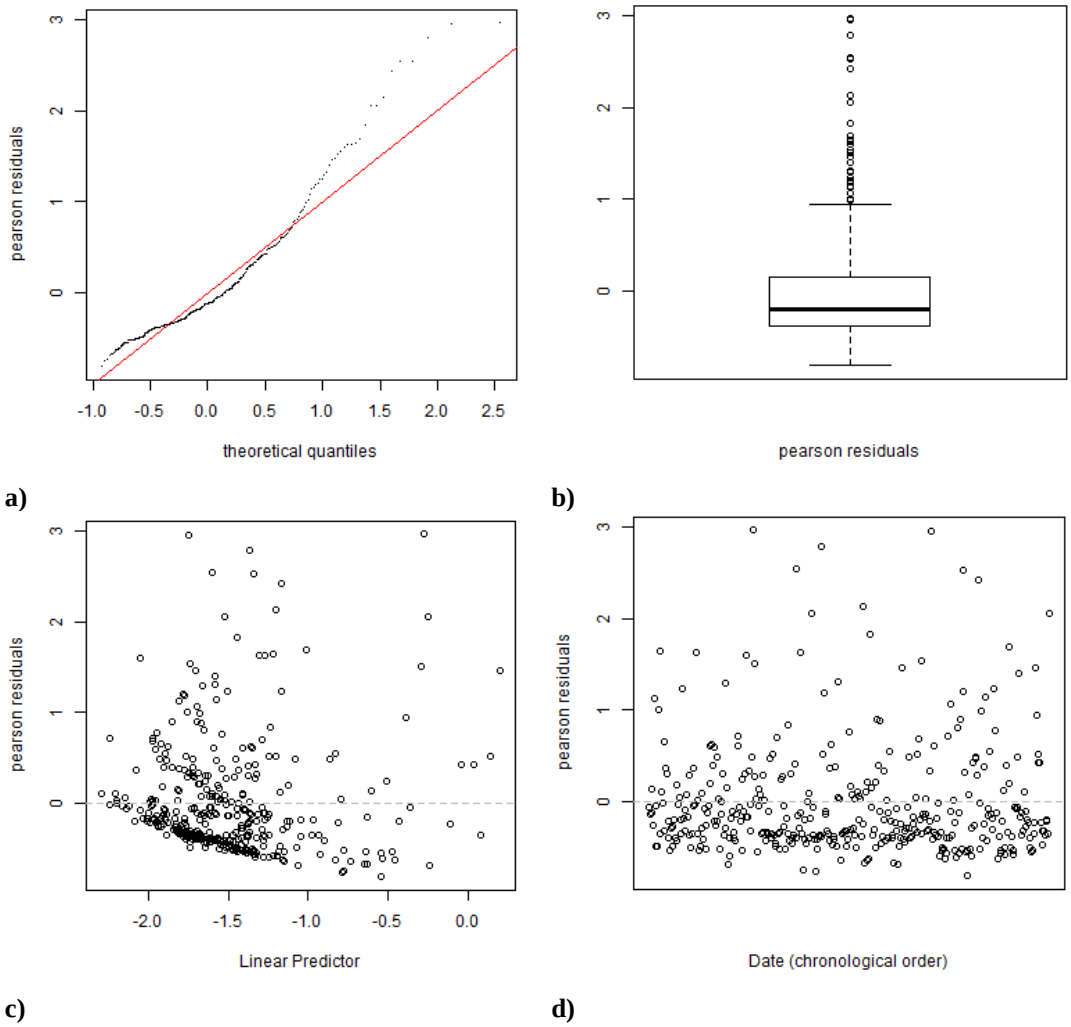


Figure 4. Residual diagnostic plots for the *Istiophorus albicans* Gamma model. Diagnostic plots include the Q-Q plot (a), box-plot (b), residuals against a linear predictor (c), and residuals against time (d).

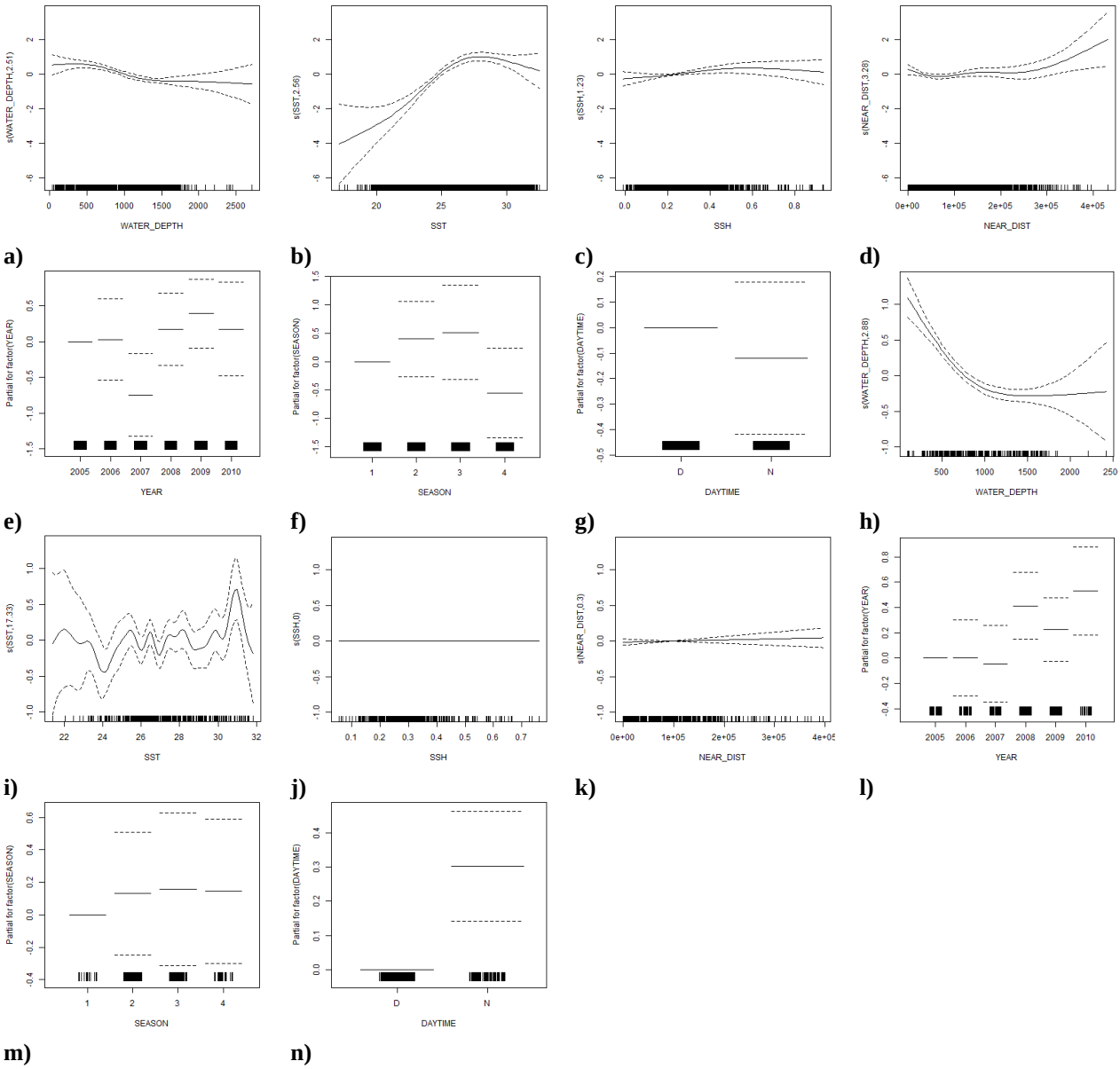


Figure 5. Model descriptor fits for the *Istiophorus albicans* delta generalized additive models. Panels (a) - (g) display descriptor fits for Bernoulli models, and panels (h) - (n) display descriptor fits for the Gamma models. A solid line indicates the fit, dashed lines indicate the 95% confidence interval, and the black dashes along the horizontal axis display the rug plot. The estimated degrees of freedom for spline fits are included in the vertical axis label.

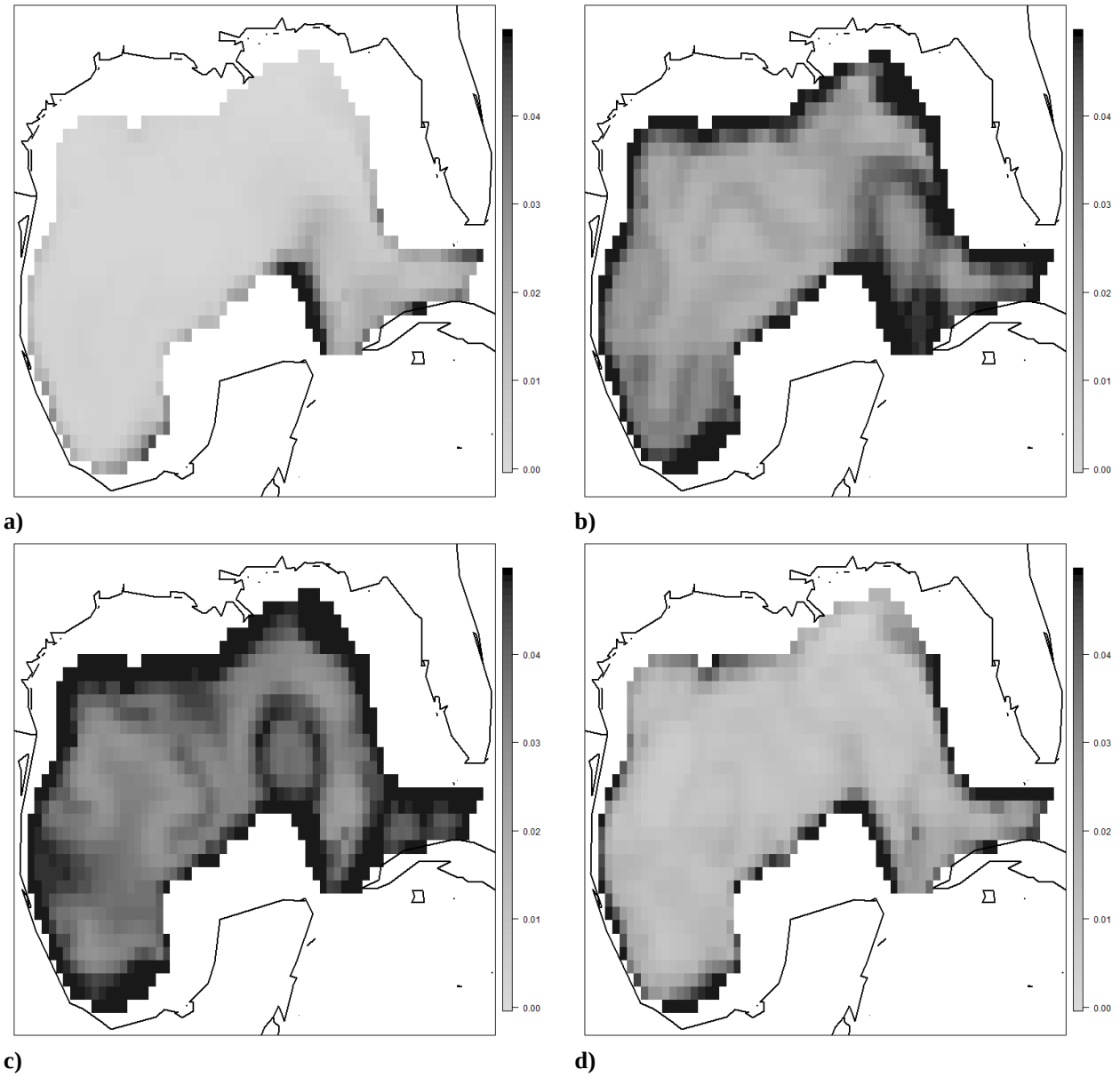


Figure 6. Abundance indices of *Istiophorus albicans* predicted from the fitted delta generalized additive models for Jan. – Mar. (a), Apr. – Jun. (b), Jul. – Sep. (c), and Oct. – Dec. (d). Predictions were generated from data fishnets (0.1° latitude by 0.1° longitude) representing seasonal averages of numerical model descriptors for the year 2010. Fishnets were generated using data from NCEI and NOAA. The plots in this figure were generated from the daytime fishnets (nighttime fishnets displayed similar patterns with slightly larger abundance indices).

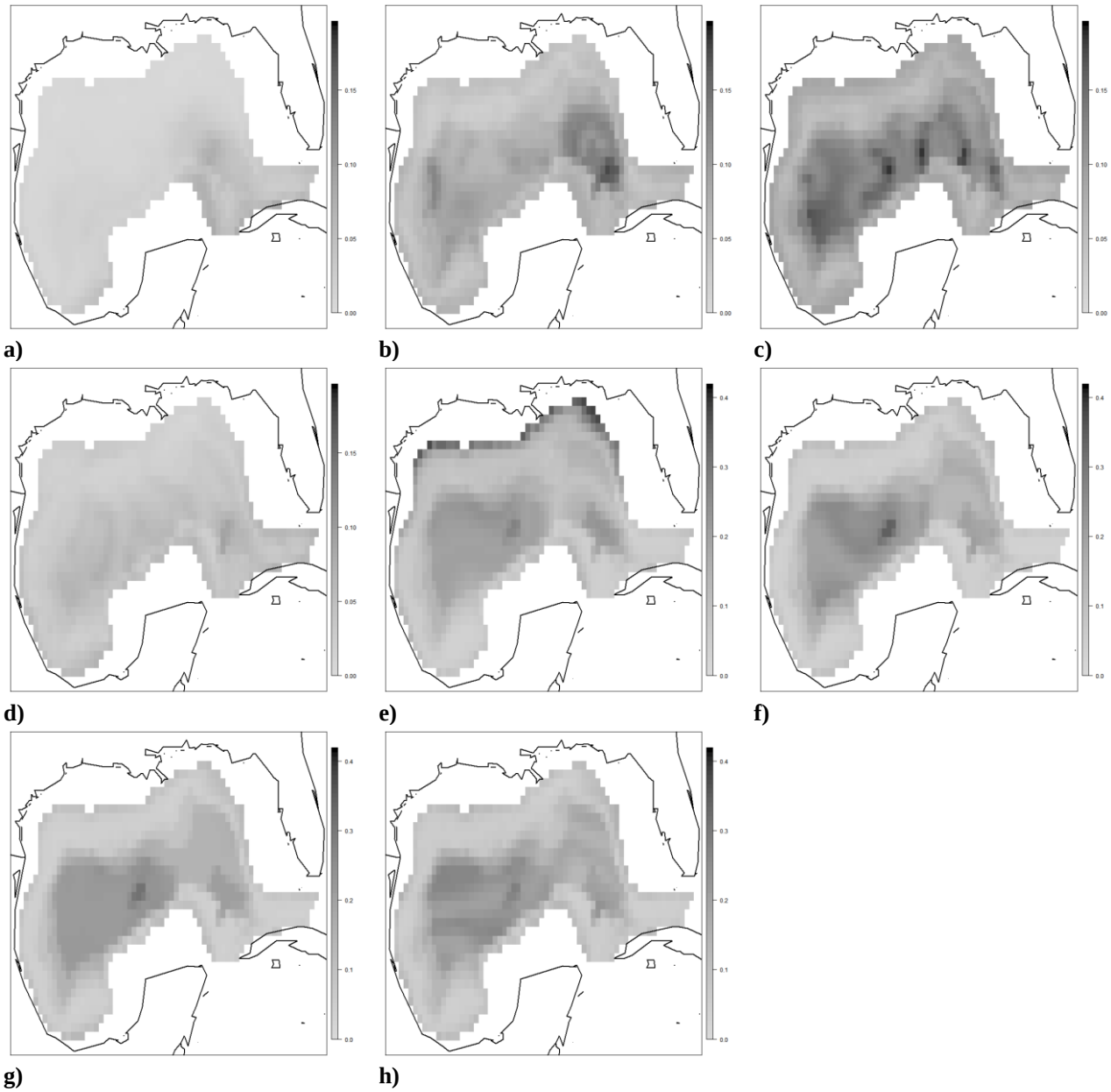


Figure 7. Standard error of predictions made with *Istiophorus albicans* generalized additive models. Panels (a) – (d) are standard errors from Bernoulli model predictions for seasons 1 through 4, respectfully, and panels (e) – (f) are standard errors from Gamma model predictions for seasons 1 through 4, respectfully.

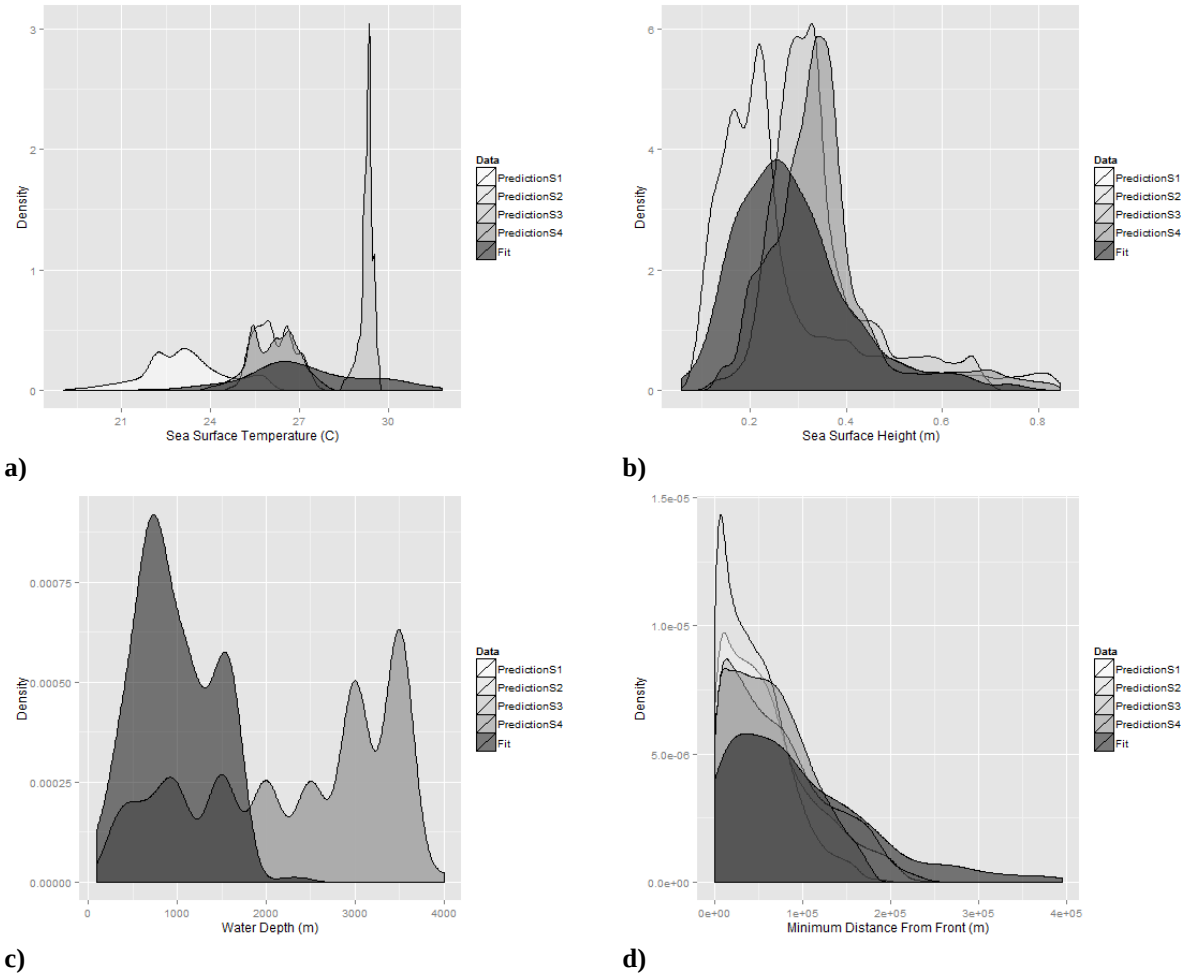


Figure 8. Density curves of data used to fit and predict with *Istiophorus albicans* generalized additive models. Black density curves are derived for data used to fit models, and grey density curves are derived from data used to develop the indicated seasonal prediction. Data used to fit the Gamma model were used to produce the shown curves since data used to fit the Bernoulli model produce similar curves.