Section S1. Materials and methods: regression analysis of the depth ranges of species

Logistic regression takes in a series of predictor (i.e., independent) variables as input, and it quantifies the relationship between those variables and a dichotomous outcome (i.e., dependent) variable. Each species in BioGoMx is described by one value for each predictor variable, as well as by one value of the dichotomous outcome variable, which can either be 0 or 1 (e.g., 0 if the depth range of the species occurring in a given depth zone does not extend into another zone, or 1 if it does extend). To relate predictors to the outcome, logistic regression estimates the values of a series of coefficients, with one coefficient associated with each predictor. The goal in those estimations is to determine the values of the coefficients that, taken together, best fit the following equation across all species:

$$Logit\ Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

Here, *Logit* refers to the log-transformed odds (log-odds) that the dichotomous outcome variable *Y* is equal to 1. Such odds are directly related to probability, as they are calculated as p/(1-p), or the probability that a species has a *Y* value of 1 divided by the probability that it instead has a *Y* value equal to 0. On the other side of the equation, b_0 is the y-intercept of the model, or the log-odds that *Y* is equal to 1 when all other predictor values for a species ($X_1, X_2, \ldots, X_n$) are equal to 0. X_1 is the predictor of interest, which in our case is the benthic versus pelagic lifestyle of a species. Then, b_1 is the estimated coefficient linking X_1 with the log-odds of *Y*. The other terms X_2 to X_n, associated with b_2 to b_n, represent other predictors, or confounding variables, as well as the estimated coefficients linking them to the log-odds of *Y*. Finally, *n* refers to the total number of predictors of the log-odds of *Y* being analyzed. From this formulation, an optimized logistic regression model estimates values for the coefficients $b_1, b_2, \ldots, b_n$, such that species with an actual *Y* value of 1 have a higher log-odds value of *Y* or, equivalently, a probability of $Y = 1$ closer to 1, and vice versa for species with an actual *Y* value of 0. Because $b_1, b_2, \ldots, b_n$ are estimated together, predictors can each account for one another in this optimization process.

The most important output from a logistic regression model for our purposes, which is derived from its estimated coefficients, is the odds ratio. The odds ratio for the predictor of interest X_1, namely the benthic versus pelagic lifestyle of species, is calculated as $e\hat{}b_1$, where *e* is euler's number (2.718) and b_1 is the coefficient associated with the predictor. This ratio denotes, among only the species that occur in a given depth zone, the disparity between the odds that a pelagic species achieves an outcome of $Y = 1$ (e.g., extends from that zone to another), relative to the total number of pelagic species occurring in that zone only, and the corresponding odds in benthic species. Because X_1 is encoded by assigning pelagic species a value of 1 and benthic species 0, an odds ratio > 1 means that the odds of a pelagic species achieving an outcome of $Y = 1$ is greater than the odds of a benthic species achieving that outcome. The opposite would be true for a ratio < 1, and a ratio equal to 1 would signal equal odds between benthic and pelagic species. Further, because every model coefficient is associated with a p-value, and because a coefficient is used to calculate the odds ratio, the statistical significance of each odds ratio may be assessed using that p-value. The odds ratio thus has several advantages in distinguishing between benthic and pelagic species in relation to a given outcome variable *Y*. It circumvents the bias that benthic species are more numerous than pelagic species, and thus have more frequent occurrences of $Y = 1$ based upon their sheer numbers. And it accounts for other relevant predictors, or confounding variables, that also affect *Y*, because it is derived from a coefficient that was calculated together with other variable coefficients.

Using this modeling framework, we optimized a total of six logistic regression models, each discerning if the benthic versus pelagic lifestyle of species (predictor X_1) influences a different dichotomous outcome variable of interest (*Y*) after accounting for confounders. Each of the six models encompassed only the species that occur in a given depth zone, and then addressed the outcome: *Y* = whether or not each species extends from that depth zone into another (one of the other two zones). Using Harrell's 'rms' package in R v3.6.1 (R Core Team 2019, Harrell Jr. 2021), we built each model by employing the equation above upon the predictor values of the species in the BioGoMx dataset. We improved the models by adding statistically significant interaction terms to the equation (terms such as $b_3(X_1X_2)$, if coefficient b_3 has a p-value < 0.05), to address scenarios in which one predictor may influence another. Further, we accounted for potential non-linear relationships between continuous predictors and the log-odds of the outcome *Y* by implementing restricted cubic splines upon those predictors. Splines are breaks that signal the points along the continuous predictor at which its relationship with the log-odds of *Y* may change direction. Once the models were optimized, we obtained each's odds ratio for predictor X_1, as well as its associated p-value. Further, we used the 'rms' package's "Predict" function to plot the probabilities that the benthic versus pelagic species that occur in a given depth zone extend into another, across varying minimum or maximum depths of species.

The confounding variables ($X_2, X_3, \ldots, X_n$) that we considered for all models were those addressing other available and relevant attributes of species (i.e., those recorded in or derivable from the BioGoMx dataset). Specifically, we considered each species' taxonomic group, minimum and maximum depths of vertical range, occurrence or absence in eight geographic octants of the Gulf of Mexico (GoMx) and their endemism, or lack thereof, to the GoMx. All of these confounders are recorded in BioGoMx in their raw form, with the exception of species' taxonomic groups, which was derived. Specifically, we used the raw data of each species' recorded kingdom, phylum, and class to divide them into the following groups of ecological importance (Felder & Camp 2009, Brenner et al. 2010): Annelida (N = 608), Chordata (N = 1670), Ciliophora (N = 490), Cnidaria (N = 784), Crustacea (N = 2503), Echinodermata (N = 566), flatworms (N = 75), Foraminifera (N = 896), macroalgae (N = 669), Mollusca (N = 2567), Plantae (N = 380), Porifera (N = 281), and other smaller taxa (N = 802). We only used species' minimum and/or maximum depths of vertical range as predictors if they were not directly involved in the calculation of the outcome variable *Y*. For example, in the model in which *Y* addresses the extension of species that occur in the mesophotic zone into the shallow zone, the minimum depths of species are used to calculate whether or not that extension occurs. Including minimum depth as a predictor is then inappropriate, as its coefficient would be artificially high, artificially reducing the coefficients of other predictors and rendering the model uninformative. Separately, we performed a sensitivity analysis regarding the minimum and maximum depths of species' vertical range, and we discovered that removing them from all models altogether had negligible effects upon the models' odds ratios and goodness-of-fit.

To assess the goodness-of-fit of the models, we recorded each's C-statistic and Brier score, and plotted a reliability diagram for each. The C-statistic is the probability that a given species with an actual outcome value of *Y* = 1 has a greater log-odds of having that value than a given species with an actual value of *Y* = 0. It is measured on a scale of 0-1, in which a value closer to 1 indicates a more robust model. The Brier score is also measured on a scale of 0-1, but it operates in the opposite direction: a value closer to 0 signals a more robust model. It is a measure of model accuracy, as it quantifies the differences between the log-odds of $Y = 1$ and the actual values of Y across all species. Finally, a reliability diagram provides a means of visualizing comparisons between predicted and actual probabilities of the outcome variable Y. The diagram is constructed using the following steps: (1) all species are divided into K groups (we chose K = 200), based upon their model-predicted probabilities of having an outcome of $Y = 1$. Species with small such probabilities are grouped, and so on up to species with large probabilities. (2) For each group, two quantities are calculated, namely the average model-predicted probabilities described in step 1 across all species in the group, and the actual frequency of species in the group with an outcome of $Y = 1$. (3) The two quantities are plotted against each other, with predicted probabilities of $Y = 1$ on one axis and actual frequencies on the other, and a curve is fit to that plot. A perfect model is one in which that curve is a diagonal line, such that predicted probabilities and actual frequencies of $Y = 1$ are equal across all 200 species groups.

Table S1. C-statistics and Brier scores for all logistic regression models optimized. Each dichotomous dependent variable was predicted from multiple predictor variables, including the benthic versus pelagic lifestyle of species (see Section S1). C-statistics and Brier scores help to quantify the ability of each model to predict each dependent variable, as a form of model validation. A C-statistic closer to 1 and a Brier score closer to 0, both on a 0-1 scale, indicates a model with more predictive power (see Section S1).

Model Dependent Variable	**C-Statistic**	**Brier Score**
(a) *Extension or Non-Extension of Species that Occur in the Shallow Zone into the Mesophotic Zone*	0.858	0.152
(b) *Extension or Non-Extension of Species that Occur in the Shallow Zone into the Deep Zone*	0.792	0.129
(c) *Extension or Non-Extension of Species that Occur in the Mesophotic Zone into the Deep Zone*	0.792	0.186
(d) *Extension or Non-Extension of Species that Occur in the Mesophotic Zone into the Shallow Zone*	0.811	0.152
(e) *Extension or Non-Extension of Species that Occur in the Deep Zone into the Shallow Zone*	0.844	0.149
(f) *Extension or Non-Extension of Species that Occur in the Deep Zone into the Mesophotic Zone*	0.834	0.162

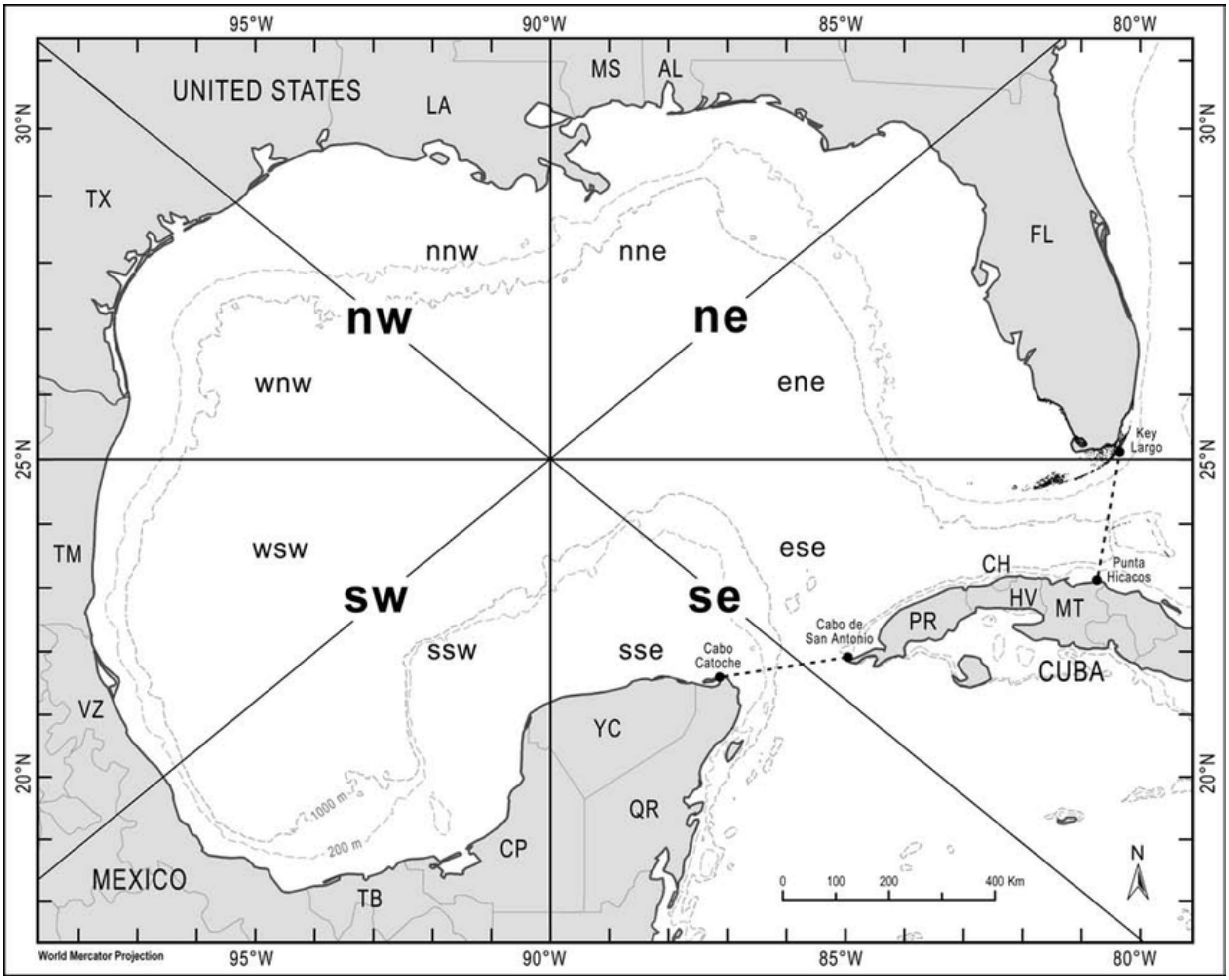

Figure S1. Map of the Gulf of Mexico, representing our study area. Gray areas are terrestrial, while white areas are marine. Lines traversing the map represent the boundaries of geographic octants that divide the Gulf (e.g., "nnw" represents the northern northwest octant). States are abbreviated as follows: MT = Matanzas, HV = La Habana, CH = Cuidad de la Habana, PR = Pinar del Rio, QR = Quintana Roo, YC = Yucatan, CP = Campeche, TB = Tabasco, VZ = Veracruz, TM = Tamaulipas, TX = Texas, LA = Louisiana, MS = Mississippi, AL = Alabama, and FL = Florida. The map was originally created by Fabio Moretzsohn, and we obtained it directly from Felder & Camp (2009).

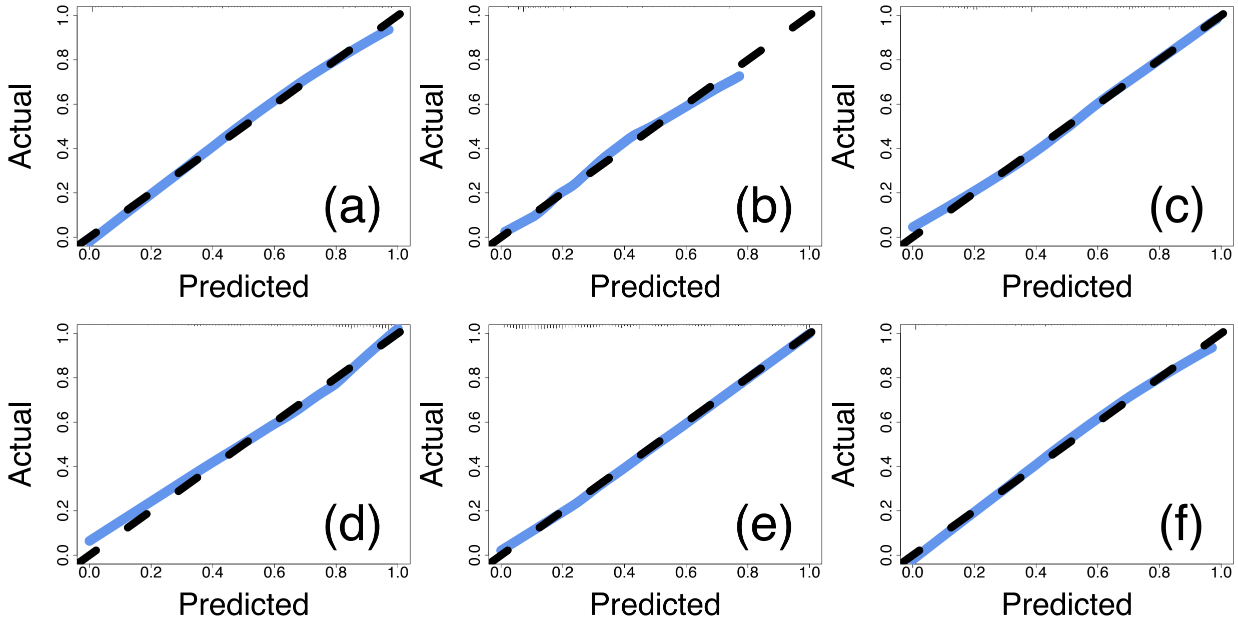

Figure S2. Reliability diagrams for all logistic regression models optimized. Diagrams address models of species' extensions from (a) shallow to mesophotic, (b) shallow to deep, (c) mesophotic to deep, (d) mesophotic to shallow, (e) deep to shallow, and (f) deep to mesophotic zones. Axes refer to actual frequencies of species' extensions between depth zones versus their model-predicted probabilities (see Section S1). Black dashed lines refer to a perfect model, in which predicted probabilities and actual frequencies are equal. Blue solid lines represent the reality of each model, or a curve fitted to actual versus predicted values.

References

Brenner J, Moretzsohn F, Tunnell JW, Shirley T (2010) BioGoMx Database: Biodiversity of the Gulf of Mexico. Proc NatureServe Conservation Conference, Austin, TX

Felder DL, Camp DK (2009) Gulf of Mexico Origin, Waters, and Biota: Volume 1, Biodiversity. Texas A&M University Press, Corpus Christi, TX

Harrell Jr. FE (2021) rms: Regression modeling strategies. R package version 6.2-0. URL https://cran.r-project.org/package=rms

R Core Team (2019) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/