Defining EFH for Alaska Groundfish Species using Species Distribution Modeling

Background: Species distribution models have been widely used in conservation biology and terrestrial systems to define the potential habitat for organisms of interest (Elith et al. 2011, etc.). The models themselves can take a number of forms, from relatively simple frameworks such as generalized linear or additive models to complex modeling frameworks such as boosted regression trees, maximum entropy models, two-stage models or other formulations. The models can be used to predict potential habitat, probability of presence or even abundance, but they all have some features in common.

- the underlying data consists of some type of independent variables (predictors) and a dependent response variable (presence, presence/absence or abundance)
- raster maps of independent variables are used to predict a response map (prediction)
- confidence bounds on the predictions and partitioning of the data can produce test statistics useful for evaluating the model

Approach: We are proposing to use a species distribution modeling framework to refine the descriptions of Essential Fish Habitat for Alaskan groundfish species. This will be attempted for each of the Alaska regions and for all groundfish species. The independent variables will consist of those variables (such as depth, slope, bottom temperature, current speeds, etc.) widely available from remote sensing or long-term monitoring programs at the AFSC. The dependent variables will be survey catches (primarily bottom trawl, but we will attempt to include pelagic surveys and ichthyoplankton surveys where available) of the Alaska FMP species. Where possible, the species will be divided by life history stage into egg/larval, juvenile and adult groups.

Because of the anticipation of variable data distributions (from log-normal to highly zero-inflated) a number of model frameworks will be considered, and the most appropriate for each species/life history stage will be used. An example analysis for Sablefish in the Eastern Bering Sea is shown in Appendix I.

Products: A document (tech memo or part of EIS?) will be delivered that describes the individual species modeling results. ArcGIS coverages will also be provided for each species. Finally, a manuscript describing the general methodology and results will be produced for publication in a peer-reviewed journal.

Timeline: We anticipate beginning this project in December 2014. Data compilation should take 1-2 months, modeling should be completed in 4 months and the document could be produced in October 2015. The final manuscript would be completed by December 2015.

Appendix I: Sablefish Habitat Model in the EBS slope

Here I constructed a distribution model for sablefish (*Anoplopoma fimbris*) for the outer shelf and slope region of the eastern Bering Sea. The model is formulated as a two-stage (or hurdle) model, with the first stage producing estimates for probability of presence or absence and the second stage producing estimates of abundance. Bottom trawl survey data from EBS slope and shelf surveys (2002, 2004, 2008, 2010, and 2012) were used to construct the models.

Habitat Variables. Independent variables for modeling included the standard suite of habitat variables typically collected on the bottom trawl survey as well as a few derived and modeled variables. Haul position and depth were collected during each bottom trawl haul. A start and end position for the vessel during the on-bottom portion of the tow were collected using the vessel-mounted GPS receiver. Vessel position was corrected for the position of the bottom trawl itself by triangulating how far the net was behind the vessel (based on the seafloor depth and the wire out) and subtracting this distance from the vessel position in the direction of the bottom trawl haul. We assumed that the bottom trawl was directly behind the vessel during the tow and that all bottom trawl tows were conducted in a straight line from the beginning point to the end point. The mid-point of the start and end positions of the net was used as the location variable in the modeling. The longitude and latitude data for each tow (and all other geographical data including the raster layers described below) were projected into Alaska Albers Equal Area Conic projection (center latitude = 50° N and center longitude = -154° W) and degrees of latitude and longitude were transformed into 100 m by 100 m square grids of eastings and northings for modeling.

The depth for each tow was estimated from a SeaBird SBE-39 microbathythermograph attached to the headrope of the net plus the measured net height. Mean depth during the tow was calculated for inclusion as an explanatory variable in the modeling. A bathymetry raster for the entire Aleutian Islands region was also produced for this analysis This raster was used for prediction, but not for parameterizing the models. Slope and rugosity were two habitat variables derived from the 100 m by 100 m bathymetry raster. Slope for each raster grid cell was computed as the maximum difference between the depth at a cell and its surrounding cells. The average summer water temperature at each site was estimated from data collected during Aleutian Islands bottom trawl surveys from 1996-2010. Bottom temperatures are collected during each bottom trawl tow using the SBE-39 attached to the headrope of the net. Mean bottom temperatures for each haul were interpolated to the 100 m by 100 m grid for the entire Aleutian Islands region. These data were interpolated using ordinary kriging (Venables & Ripley 2002) with a spherical semi-variogram model. This resulted in a single temperature raster layer that reflects the average temperature conditions in surveys from 1996-2011 (Fig. 2). When evaluated using leave-one-out cross-validation, the kriging model was a statistically significant fit to the observations (n = 2814, mean squared error = 0.19, $R^2 = 0.38$), capturing the spatial trend in the temperature data. The temperature data used in our models were primarily designed to reflect long-term averages that could be compared spatially to the distribution of corals and sponges. Mean bottom temperature underneath each bottom trawl tow path was used as a habitat variable in the modeling. The 100 m by 100 m raster layers of average temperature were used for prediction.

Three measures of water movement and its potential interaction with the seafloor were used as habitat variables in modeling and prediction. The first variable was the maximum tidal speed on a 10 km^2 grid. Tidal speeds were estimated for 368 consecutive days (January 1st, 2009 to January 3rd, 2010) using a tidal inversion program parameterized for the eastern Bering Sea (Egbert & Erofeeva 2002). This tidal prediction model was used to produce a time series of one year of tidal currents for spring and neap cycles at grid node. The mean values were then interpolated to a 100 m by 100 m grid using inverse distance weighting. The mean of the time series of predicted tidal current was then extracted for the position of each bottom trawl survey haul. This mean value was used as a habitat variable in the modeling.

The second water movement variable was the predicted bottom water layer current speed from ROM's model runs from 1970-2004. This long-term current speed and direction were available as points on a 10 km by 10 km grid. The ROM's model was based on a three-dimensional grid with 60 depth tiers for each grid cell. For example, a point at 60 m water depth would have 60 depth bins at 1 m intervals, while a point at 120 m depth would have 60 depth bins at 2 m depth intervals, etc.). The current speed and direction for the deepest depth bin at each point (closest to the seafloor) was used in this analysis. This regularly spaced data was interpolated to a 100 m by 100 m cell size raster covering the entire Aleutian Islands using inverse distance weighting. Then the values from this raster at each of the bottom trawl survey haul locations were extracted and the mean value computed for the path of each bottom trawl survey tow. The raster was also used for prediction.

The final water current variable used in the modeling was the aspect of the seafloor relative to the mean current direction. The aspect of the seafloor (angle the seafloor faces) in degrees relative to north (0°) was computed using the raster package in R software. This data was produced on a 100 m by 100 m raster grid, the same as the bathymetry data. The current direction used was the mean current direction from the long-term model output from the ROMS model (Danielson et al. 2011). The absolute value of the difference between the current direction and the aspect of the seafloor at the position of each bottom trawl haul was used as a habitat variable in the modeling. This value ranged from 0° (where the currents were flowing in the same direction the seafloor was facing) to 180° (where the mean current was flowing directly opposite the aspect of the seafloor). The raster grid of the aspect variable (on the 100 m by 100 m grid) was used in the prediction.

To reflect average ocean productivity (g C m^{-2} day^{-1}) at each of the bottom trawl survey sites, we used MODIS ocean color data for five spring-summer months (May-September) that encompass the spring and summer phytoplankton blooms over eight years (2003-2011) for the eastern Bering Sea region. These data were downloaded from the Oregon State University Ocean Productivity website. These data were averaged by cell and by month and then averaged again by cell and by year (to account for differences in the number of samples within each cell). The averages were then interpolated to 100 m by 100 m raster grids using inverse distance weighting. The mean value in this grid underlying each bottom trawl survey tow was extracted from this raster. The raster was used for prediction.

Eastings (longitude) and northings (latitude) were very strongly correlated ($R^2 = 0.59$) because of the geographical shape of the eastern Bering Sea slope and as such were included as a bivariate term (location) in the model. The remaining habitat variables used in the models were in a univariate form.

Model Fitting. Two-stage models were fit to the bottom trawl survey catches for Sablefish. In the first stage, presence or absence was predicted using all bottom trawl survey hauls (n=1357) on the outer shelf and slope. In the second stage log(CPUE) was predicted using only bottom trawl hauls that captured sablefish (n = 483). Generalized additive models (Hastie & Tibshirini 1990) using the mgcv package in R (Wood 2006) were used to predict the two dependent variables with the suite of untransformed habitat variables included, so that the full model was

$$y = s(location) + s(depth) + s(temperature) + s(slope) + s(mean\ tidal\ current) + s(mean\ current\ speed) + s(ocean\ color) + s(aspect) + s(phi) + s(sort) + \varepsilon$$

where y was the dependent variable presence or absence of sablefish in bottom trawl hauls and s indicates a thin plate regression spline smoothing function (Wood 2006). In each case the basis degrees of freedom used in the smoothing function was limited to ≤ 4 for univariate variables and ≤ 30 for the bivariate term (location). For presence or absence models a binomial distribution was used for the fitting. The Gaussian distribution with log-transformed CPUE data and a constant of half of the smallest positive value proved to best approximate normality for the CPUE data for sablefish.

A factorial analysis was used to reduce the number of variables in each model. Initially a full model containing the entire variable suite was fit to the data. Then the least significant variable was removed from the model, provided the GCV score for CPUE models or the UBRE score for binomial models was lower with the elimination of the variable, and then the reduced model was re-fit to the data. Stepwise variable removal was continued until a final best-fitting model was reached, where the removal of additional variables did not result in a lower value for GCV or UBRE.

To test the performance of the best-fitting models, the predictions were compared to the observations. For presence and absence models the area under the curve (AUC) was computed to judge model performance. The AUC calculates the probability that a randomly chosen presence observation would have a higher probability of presence than a randomly chosen absence observation using rank data. We used the scale of Hosmer & Lemeshow (2005), where AUC value > 0.5 is estimated to be better than chance, a value > 0.7 is estimated to be acceptable, and values > 0.8 and 0.9 are excellent and outstanding, respectively. Confidence intervals for the AUC (95%) were calculated according to the methodology of DeLong et al (1988). For abundance and diversity models the performance was directly tested by correlating the predictions with the observations.

Sablefish Model Results

The best fitting presence-absence model for sablefish included the location variable (bivariate term of latitude*longitude), bottom depth, seafloor slope, bottom temperature, tidal current, ocean color, and the two measures of sediment character; sorting and phi (Figure 1). The model explained 68% of the deviation in the data set. The model fit very well, with an AUC of 0.97, which is considered "outstanding". A threshhold probability for determining presence or absence was set at 0.5 (i.e. for trawl hauls with a probability of presence of sablefish >0.5, presence was designated). Using this threshhold resulted in about a 9% error rate for predicting presence or absence.

The best fitting model of abundance (trawl survey log(CPUE)) included; location, depth, slope, bottom temperature, ocean color and phi (Figure 2). This model also fit the data well, explaining 43% of the deviation in the data set (Figure 3). When a threshhold value of 0.08 was used to determine presence or absence (the threshhold value that balanced the error rate between predicted presence or absence), a map of abundance showed that occurred at medium depths (500-800 m) along the slope (Figure 4). Almost no sablefish were predicted to occur on the eastern Bering Sea shelf. Interestingly, most of the areas of high abundance of sablefish were predicted to occur in canyons (Bering, Pribilof and Zhemchug).

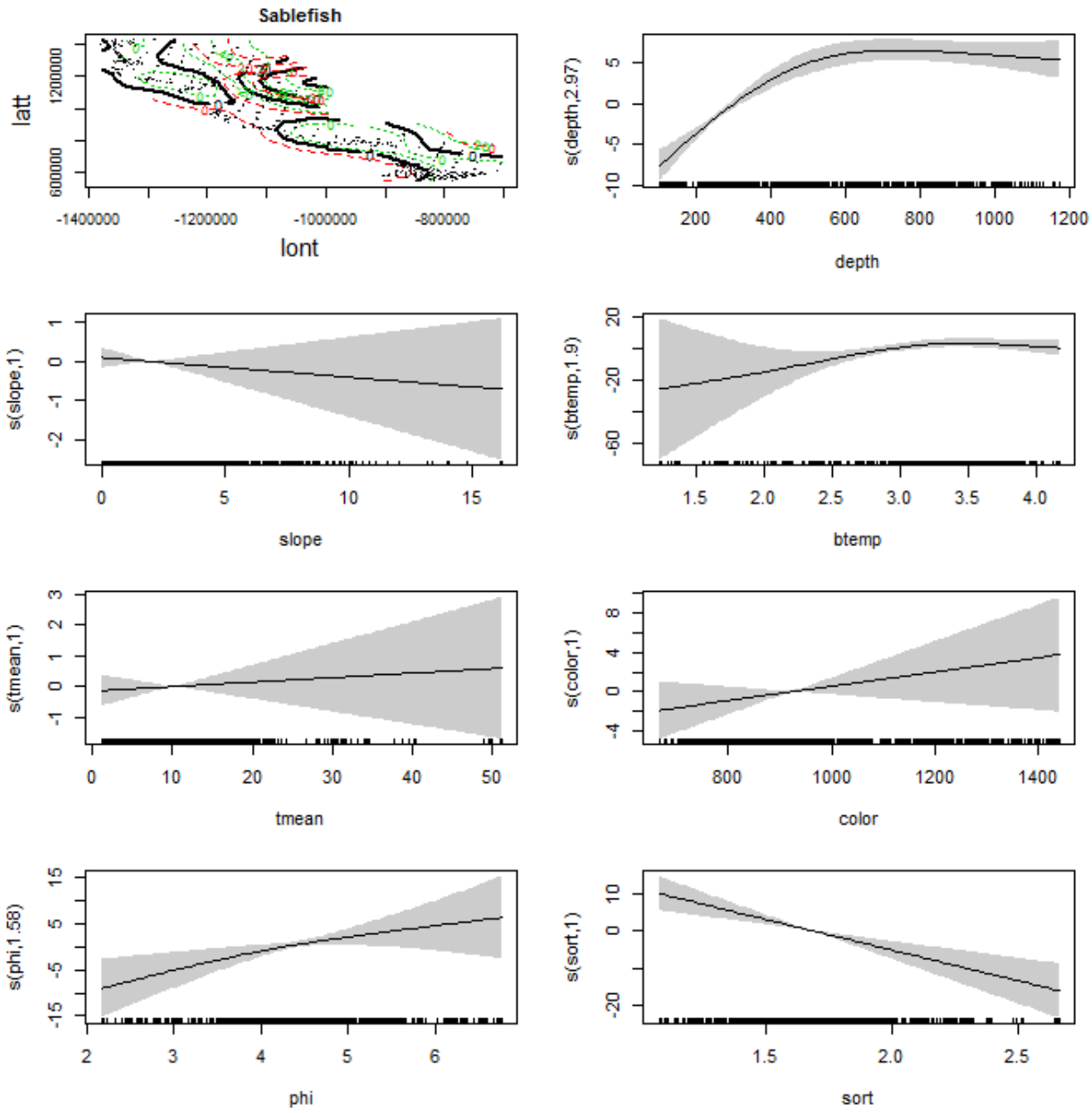

Figure 1. Response of sablefish probability of presence with significant variables in the best fitting model of sablefish presence or absence.

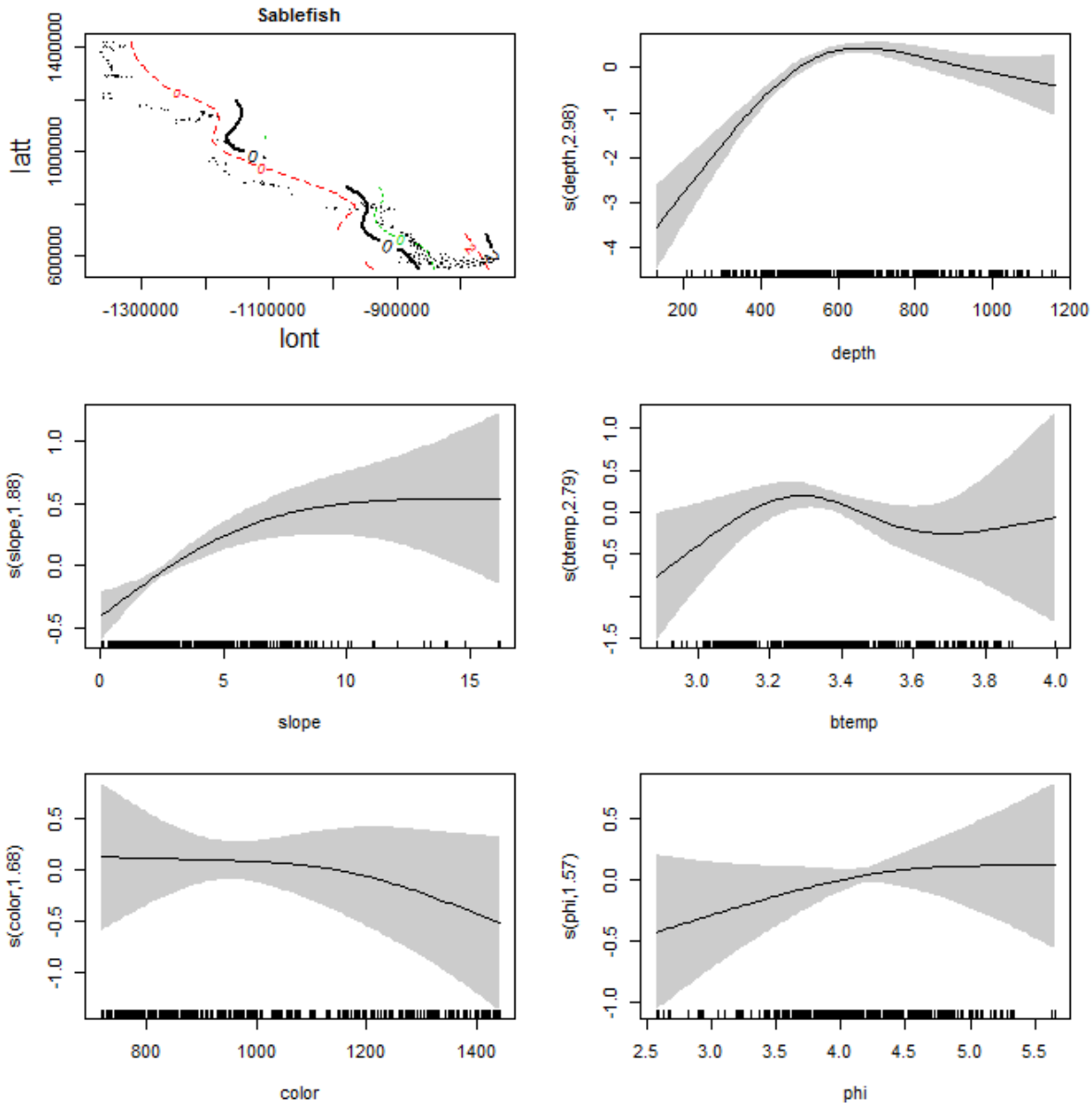

Figure 2. Response of sablefish log-transformed abundance with significant variables in the best fitting model of sablefish abundance. Data were only bottom trawl hauls with sablefish present.

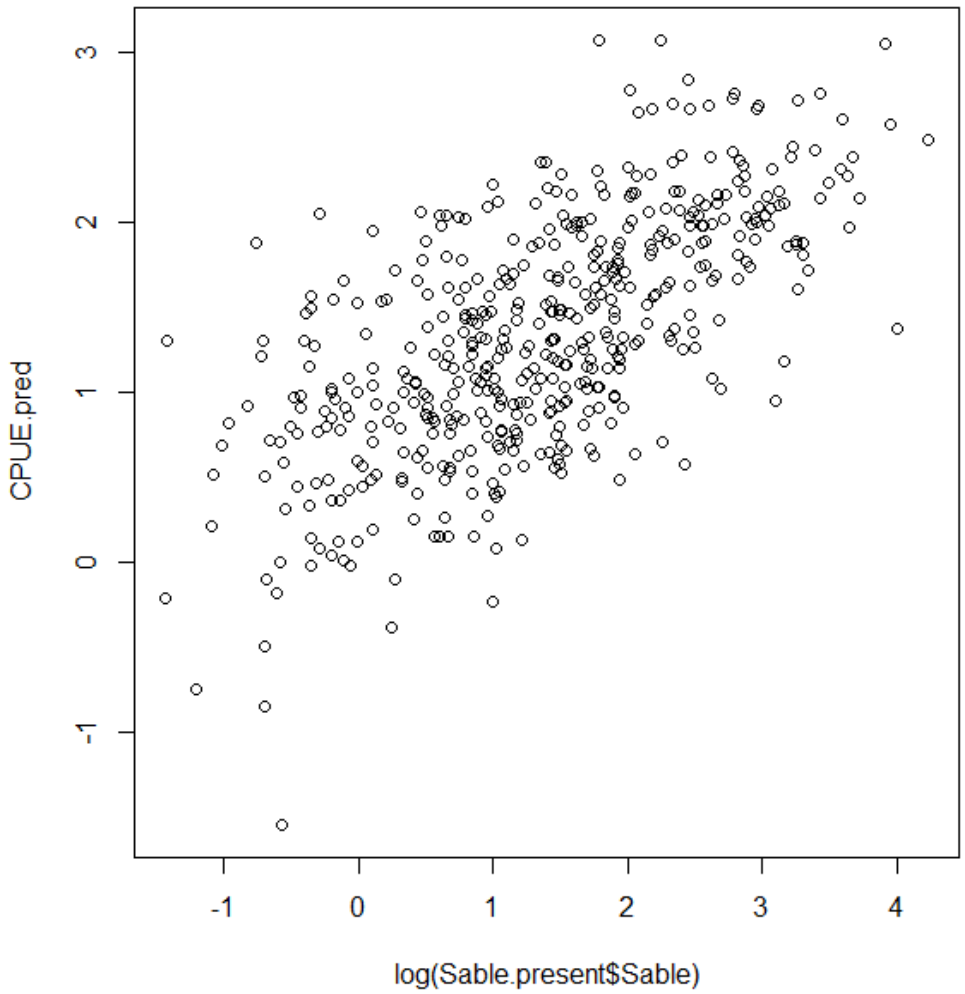

Figure 3. Predicted (y-axis) and observed (x-axis) log-transformed sablefish catches from bottom trawl survey tows where sablefish were present. Predictions were made using the best fitting GAM model.

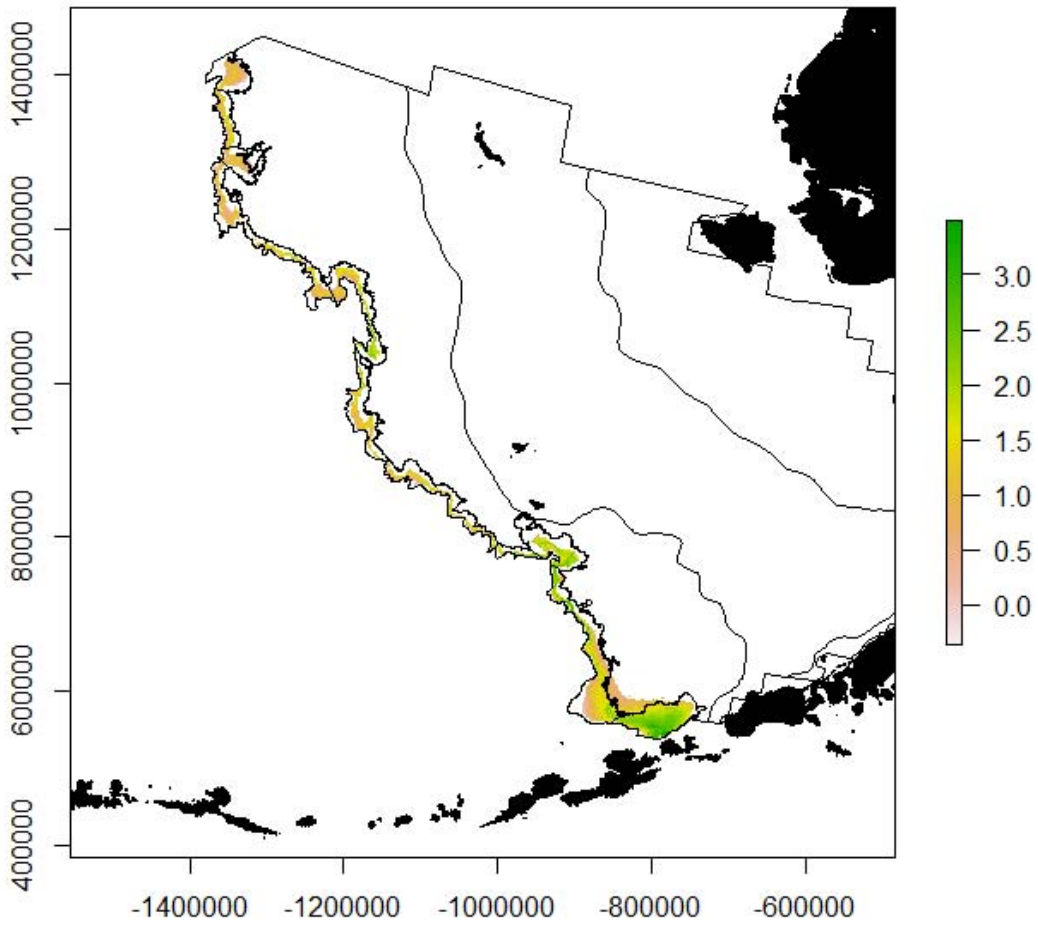

Figure 4. Predicted sablefish abundance (log-transformed) from the hurdle model using a threshhold value of 0.5, meaning that at grid cells where the probability of presence is predicted to be >0.5, sablefish abundance is predicted using the two-stage model.