An ensemble approach to species distribution modeling reconciles systematic differences in estimates of habitat utilization and range area

– ODMAP Protocol –

Jeremy Harris, Margaret Siple, James Thorson, Edward Laman, Jodi Pirtle

2023-08-10

## Overview

#### Authorship

Contact : [jeremy.harris@dfw.wa.gov](mailto:jeremy.harris@dfw.wa.gov)

Study link: <https://repository.library.noaa.gov/view/noaa/49987>

#### Model objective

Model objective: Mapping and interpolation

Target output: abundance, suitable habitat, range area

#### Focal Taxon

Focal Taxon: Rex Sole (adult)

#### Location

Location: Gulf of Alaska

#### Scale of Analysis

Spatial extent: -167, -130, 52, 60 (xmin, xmax, ymin, ymax)

Spatial resolution: 1

Temporal extent: 1993-2019

Temporal resolution: year

Boundary: political

#### Biodiversity data

Observation type: standardised monitoring data

Response data type: counts

#### Predictors

Predictor types: climatic, topographic

#### Hypotheses

Hypotheses: The Essential Fish Habitat process included environmental covariates that were expected to affect fish density based on previous studies: bottom depth and temperature, water movement, location, seafloor terrain (slope, aspect, curvature, and bathymetric position index [BPI]), substrate rockiness, and biogenic structures (the presence of sponges, crabs, and pennatulaceans) were all included in SDMs as covariates for groundfish density.

#### Assumptions

Model assumptions: Functions are additive and the response variable (in this case, density or presence/absence) belongs to one of the specified distributions (Poisson or negative binomial in the case of density, binomial or cloglog link in the case of presence-absence and hurdle models).

#### Algorithms

Modelling techniques: maxent, gam

Model complexity: The GAMs used in this study included all the potentially relevant covariates and stepwise term elimination was used to remove covariate predictors based on minimizing model-dependent generalized cross-validation (GCV) or unbiased risk estimator (UBRE) scores. Between Poisson and negative binomial GAMs, we used RMSE-based skill testing to identify the best-performing model (lowest RMSE) and keep it in the ensemble. Basis degrees of freedom used in the smoothing function for each habitat covariate were constrained following the methods of Weinberg and Kotwicki (2008). However, attempting to extrapolate model predictions into areas with few data points required additional consideration. In particular, the default smoother when fitting GAMs, a "thin-plate spline," sometimes produced exaggerated predictions in areas of sparse data (Wood 2003). To counter this behavior in one-dimensional smooth terms, we used a smoothing penalty based on the first derivative (as opposed to the default second derivative), which tended to push the effect curve towards zero where data were unavailable. For two-dimensional smooth terms, the same method was applied, but "Duchon" splines were used instead of thin-plate or cubic splines (Duchon 1977), which did a better job of penalizing the smooth function in areas with sparse data. Finally, if a GAM based on thin-plate splines failed, a second version using cubic splines in the one-dimensional smooth terms was attempted. If both versions failed to converge or produced unreasonable results, that particular GAM was excluded from the final ensemble.

Model averaging: The set of best SDMs from each category was then weighted by the inverse of its cross-validated RMSE and constituent SDM weights were normalized to sum to one. Predictions from the ensemble were made by multiplying each constituent prediction by its weight and summing the weighted predictions across SDMs. The result of this exercise was a final ensemble for each species/ life stage that predicts habitat-related abundance.

#### Workflow

Model workflow: We combined the best-fit constituent SDMs into single species life stage-specific ensemble predictions of habitat-related abundance. In practice, this means we first identified the best-performing MaxEnt, paGAM, hGAM, and GAM SDMs. In the GAMs this involved backward stepwise term elimination. For the standard GAM, the Poisson and negative binomial error distributions were modeled separately, and skill testing using the RMSE was employed to select the distribution that best characterized the data. The set of best SDMs from each category was then weighted by the inverse of its cross-validated RMSE and constituent SDM weights were normalized to sum to one. Predictions from the ensemble were made by multiplying each constituent prediction by its weight and summing the weighted predictions across SDMs. The result of this exercise was a final ensemble for each species' life stage that predicts habitat-related abundance.

#### Software

Software: R v3.6.1, mgcv: R package version 1.8-29, maxnet ver. 1.4

Code availability: <https://github.com/alaska-groundfish-efh/EFHSDM>

Data availability: <https://www.fisheries.noaa.gov/foss/f?p=215:28:11295295191265>:::::

## Data

#### Biodiversity data

Taxon names: Rex sole, species name Glyptocephalus zachirus, family Pleuronectidae

Taxonomic reference system: Standard taxonomic system used for Groundfish Assessment Program surveys and studies - this system is available upon request.

<Ecological level>

Data sources: RACE bottom trawl survey data are available online at the DOI above. Models were last updated in 2021.

Sampling design: Stratified random survey design

Sample size: 9714 hauls across all survey years with positive catches. Hauls with positive catches range from 153 (in 1993) to 837 (in in 2005). These numbers should not be compared across years because the survey effort has varied over time. This is just a raw count of the hauls with positive catches of rex sole.

Clipping: The full survey range was used to fit SDMs.

Scaling: NA

Cleaning: Before each SDM was fit, catch per unit effort (CPUE where effort is the area swept) data from the survey were filtered to the species of interest and separate into age classes (adult, subadult, juvenile) suing literature data on size at age. For rex sole, adults were considered to be individuals with a standard fork length >352 mm.

Absence data: These were treated the same as presence data; the SRS survey was carried out across the study area. Hauls with zero catch of Rex Sole were used in this analysis.

Background data: Before each SDM was fit, catch per unit effort (CPUE where effort is the area swept) data from the survey were filtered to the species of interest and separate into age classes (adult, subadult, juvenile) suing literature data on size at age. For rex sole, adults were considered to be individuals with a standard fork length >352 mm.

Errors and biases: There are few potential biases in the data for this species. For some of the other species in the study that are more closely associated with rocky habitats, it is possible that the catchability for the trawl survey may be lower than for species more closely associated with soft bottom habitat.

#### Data partitioning

Training data: For cross-validation, SDM was fitted to a randomly selected "bag" partition containing 90% of the observed abundance at trawl stations (i), predicting abundance at the remaining "out-of-bag" partition containing the other 10% of trawl stations, and comparing the predicted (y) and observed (x) values for the testing subset. The k-fold cross-validation was repeated 10 times until every point in the data set had been tested.

Validation data: See "Selection of training data" section for cross-validation method.

Test data: NA

#### Predictor variables

Predictor variables: Fourteen different predictor variables were used. Covariates are described in the supplementary materials. Interpolation methods are included for each covariate where relevant.

Data sources: There are several sources for the covariate data; we direct the readers to Table 2 in Pirtle et al. (2023) for detailed descriptions of each.

Spatial extent: -160, -130, 52, 60 (xmin, xmax, ymin, ymax)

Spatial resolution: 1 km2

Coordinate reference system: Albers Equal Area Conic (standard parallels 55N and 65N, and center longitude 154W).

Temporal extent: 1993-2019

Temporal resolution: Annual

Data processing: "Independent covariates were chosen based on their potential to influence the distribution and abundance of groundfish. Covariates with Variance Inflation Factors greater than or equal to 5.0 were eliminated (Sigler et al. 2015). These independent covariates include dynamic or static habitat attributes observed on the bottom trawl surveys as well as modeled variables (e.g., ROMS projections). Independent habitat covariates from the time series (1993-2019) were interpolated on regular spatial grids of 1 km2 for subadult and adult life stages using natural neighbor interpolation, inverse distance weighting, ordinary kriging with an exponential semi-variogram, or empirical Bayesian kriging with a semi-variogram estimated using restricted maximum likelihood (REML). For the bottom trawl survey data, covariate raster values were extracted as averaged values along the towpaths at the bottom trawl stations. When predicting species distribution and abundance, the complete raster of each retained covariate was used as input into the final SDMs. In the case of observed, dynamic predictor variables such as bottom temperature measured on bottom trawl surveys, the rasterized multi-year averages of each grid cell were used to represent average conditions over time.  
" "Independent covariates were chosen based on their potential to influence the distribution and abundance of groundfish. Covariates with Variance Inflation Factors greater than or equal to 5.0 were eliminated (Sigler et al. 2015). These independent covariates include dynamic or static habitat attributes observed on the bottom trawl surveys as well as modeled variables (e.g., ROMS projections). Independent habitat covariates from the time series (1993-2019) were interpolated on regular spatial grids of 1 km2 for subadult and adult life stages using natural neighbor interpolation, inverse distance weighting, ordinary kriging with an exponential semi-variogram, or empirical Bayesian kriging with a semi-variogram estimated using restricted maximum likelihood (REML). For the bottom trawl survey data, covariate raster values were extracted as averaged values along the towpaths at the bottom trawl stations. When predicting species distribution and abundance, the complete raster of each retained covariate was used as input into the final SDMs. In the case of observed, dynamic predictor variables such as bottom temperature measured on bottom trawl surveys, the rasterized multi-year averages of each grid cell were used to represent average conditions over time.  
"

Errors and biases: None known for the independent variables (covariate data).

Dimension reduction: See above for details about covariate selection and dimension reduction.

#### Transfer data

Data sources: Not Applicable. There is no intention at present to project this model outside of the range covered by the biannual surveys, or to predict forward in time. There is some minor extrapolation along the edges of the survey grid that may have novel covariates values, and this is handled by discarding models that produce implausible estimates of abundance. See methods for further details.

<Spatial extent>

Spatial resolution: NA

Temporal extent: NA

Temporal resolution: NA

Models and scenarios: NA

Data processing: NA

Quantification of Novelty: NA

## Model

#### Variable pre-selection

Variable pre-selection: Covariates were chosen based on their potential to influence the distribution and abundance of North Pacific groundfish and crab life stages in the ecosystems where these species have been sampled

#### Multicollinearity

Multicollinearity: To eliminate strongly collinear terms, variance inflation factors (VIF) were calculated using the methods of Zuur et al. (2009), and covariates with VIFs greater than or equal to 5.0 were eliminated (Sigler et al. 2015). These independent covariates (or predictor variables) represent dynamic or static habitat attributes observed on the bottom trawl surveys as well as modeled variables describing the marine environment in the study area (e.g., bottom current from Northeast Pacific 5 km resolution Regional Ocean Modeling System (NEP5 ROMS; Danielson et al. 2011).

#### Model settings

maxent: featureSet (Linear; Quadratic; Product; Hinge), regularizationMultiplierSet (.5-3.0), regularizationRule (Best regularization determined by 10-fold CV), offsetSet (Area swept)

gam: family (Various), smoothTerms (See supplemental table S1), offset (log(Area Swept))

Model settings (extrapolation): Not Applicable; in this study we do not project predictions into new areas or years.

#### Model estimates

Coefficients: NA

Parameter uncertainty: There was no resampling to quantify parameter uncertainty for this model. The Coefficient of Variation was mapped separately from the predicted density, so that there was some way to visualize areas with more variability.

Variable importance: Deviance explained based on a jack-knife procedure is available for all covariates. See Pirtle et al. (2023)

#### Model selection - model averaging - ensembles

Model selection: Backward stepwise term elimination.

Model averaging: The set of best SDMs from each category was then weighted by the inverse of its cross-validated RMSE and constituent SDM weights were normalized to sum to one.

Model ensembles: MaxEnt and GAMs were combined in ensemble models for each species and life stage, including adult rex sole.

#### Analysis and Correction of non-independence

Spatial autocorrelation: To reduce the effects of spatial autocorrelation on the results, we chose to combine latitude and longitude into a smoothed bivariate geographic location term included as an independent predictor in SDM formulations. Rooper et al. (2020) demonstrated that this approach can reduce spatial autocorrelation in the modeled results.

Temporal autocorrelation: Temporal correlation was not an issue because all years of data were combined to produce maps of abundance.

Nested data: NA

#### Threshold selection

Threshold selection: For EFH, encounter probability was used to categorize different areas (referred to as the "probability method"). However, see the main paper for a discussion of threshold effects and a comparison of two alternate methods.

## Assessment

#### Performance statistics

Performance on training data: RMSE; AUC; Spearman's Rho; % Deviance Explained

Performance on validation data: RMSE; AUC; Spearman's Rho; % Deviance Explained

Performance on test data: NA

#### Plausibility check

Response shapes: Covariate response plots are available for the top 9 most important covarites in the ensemble. See Pirtle et al. (2023).

Expert judgement: Predicted density maps and predicted presence-absence maps were generated after model fitting and reviewed by stock experts (in this case, stock assessment authors).

## Prediction

#### Prediction output

Prediction unit: Numerical abundance and encounter probability

Post-processing: NA

#### Uncertainty quantification

Algorithmic uncertainty: We estimate the uncertainty of each model based on the variance in predictions across the 10 cross-validation folds. For the ensemble, we modify the methods Burnham and Anderson (2002) to use RMSE weights to compute an expected variance. See Pirtle et al. 2023 for more details.

Input data uncertainty: NA

Parameter uncertainty: NA

Scenario uncertainty: NA

Novel environments: NA