# Abstract

Modeling the abundance and distribution of species over time in response to anthropogenic and environmental processes allows managers to efficiently allocate resources to measure and predict future responses. However, the modeled processes may be biased if they fail to account for the spatiotemporal heterogeneity in species distributions. Here we present an exploration of three common models – generalized linear mixed-models (GLMM), general additive mixed-models (GAMM), and random forest models (RF) – for estimating the spatiotemporal distribution of juvenile and spawning Coho salmon (*Oncoryhnchus kitsuch*) densities in coastal tributaries of Oregon, USA. Using AIC to compare the parsimony of model forms, we found that both the GAMM and GLMM models with the lowest AIC included spatiotemporal random deviates. Similarly, we found the RF model with the lowest root mean square error (RMSE) for the out-of-bag samples included spatiotemporal features. Comparing the predictability for out-of-bag samples for juvenile and adult densities zero, one, and two years into the future, the RF model performed the best, followed by the GLMM and then the GAMM; however, the advantage of the RF model decreased markedly for one and two years into the future. Among the features in the RF model, xx and xx provide the most predictive ability for the juvenile and adult densities, respectively.

# Introduction

# Methods

## Data

Data on juvenile Coho salmon densities (#/km) and habitat characteristics were collected from 24 Oregon coastal watersheds from 1998 to 2019 (Figure Map). The habitat data for each stream segment included: stream slope, maximum gradient, stream width, distance to the ocean, spring precipitation, MWMT\_index, and W3Dppt. We included all the habitat covariates where the correlation <0.7 (Hosmer Jr et al., 2013).

## Generalized linear mixed-effect model (GLMM)

Generalized linear mixed models (GLMM) describe the dependent variable arising from any number of statistical distributions (e.g., binomial, Poisson, Gaussian, etc.), and the effects of the independent variables may be fixed and or random. Our data include both zeros and positive values, and the independent effects are a combination of fixed (i.e., environmental) and or random effects (i.e., spatial). To address the challenges of our data, we used sdmTMB package in R (Barnett, Ward, and Anderson 2022) – a high level wrapper for modeling continuous spatiotemporal heterogeneity across the study domain using stochastic partial differential equations (SPDE) to characterize the Gaussian Markov random fields (Lindgren, Rue, and Lindstrom 2011).

We assumed a Tweedie distribution for the observation, where the expected density of juvenile or spawning Coho salmon () is a linear combination of fixed effects describing the relationship between juvenile and spawner densities and the parameters for the spatial precision matrix, and the random effects for deviations in juvenile and spawner abundance across space and time are estimated using the Laplace approximation in Template Model Builder (Kasper et al. 20XX). We made several modifications to the data to improve the parameter estimation: 1) the coordinate data were transformed from latitude and longitude to UTM (Universal Transverse Mercator) coordinates using zone 10 and distances in kilometers (Figure 1), and 2) all of the covariates in the model were Z-scored with a mean of zero and standard deviation of one. Unlike the GAMM or RF models, the GLMM is a parametric model that does not require any “tuning.” The fixed effects portion of the model is similar in syntax to other linear modeling R packages, but the spatial arguments and estimation of the spatial and or spatiotemporal random effects are included using Boolean flags,

R> sdmTMB("Juv.km ~ 1 + fSTRM\_ORDER + StrmSlope + WidthM +

OUT\_DIST + SolMean + MWMT\_Index +

W3Dppt + StrmPow + SprPpt + IP\_COHO",

data = juvData, #juvenile only data

mesh = mesh, #inla mesh passed to sdmTMB

family = tweedie(link = "log"), #tweedie

time = "SpwnYr", #temporal column

spatial = TRUE, #include equilibrium spatial

spatiotemporal = TRUE, #spatiotemporal

anisotropy = TRUE, #equal decorrelation axes

silent=TRUE) #remove messages

We ran 12 model forms for the GLMM: including all or none of the environmental covariates, and different permutations of temporal, spatial, and spatiotemporal random deviates. Parsimony of the different model forms are evaluated using marginal AIC (Akaike 1979), and we used the simulation-based R package DHARMa (Hartig 2022) to examine issues of model misspecification based on the residuals for the form with the lowest AIC.

## Generalized additive model (GAMM)

Generalized additive mixed-models (GAMM) are an extension of generalized linear models that allow for non-linear relationships between the response and predictor variables. The complexity in the non-linear processes between the response and predictor is controlled by the basis dimension *k*, which we set equal to 4 to allow for complexity while avoiding overfitting (Santana et al. 2012). The spatiotemporal variability of the model is defined by a three-dimensional spline; however, the spatiotemporal variance in the GAMM is permitted to follow an unknown smooth function rather than being restricted to a Gaussian Markov process in the GLMM. The GAMMs were constructed using the mgcv package R (Wood 2004) which has a syntax that is similar to the GLMM.

R> gam(Juv.km ~

fSTRM\_ORDER + s(StrmSlope,k=4) + s(WidthM,k=4) + s(OUT\_DIST,k=4) + s(SolMean,k=4) + s(MWMT\_Index,k=4) +

s(W3Dppt,k=4) + s(StrmPow,k=4) + s(SprPpt,k=4) + s(IP\_COHO,k=4) + #Environmental fixed effect with k = 4

s(UTM\_E\_km, UTM\_N\_km, JuvYr), #Spatiotemporal effects

data=juvData, #data

family = "tw") #Tweedie observation model

Similar to the GLMM, each form included all or none of the covariates and different permutations for the temporal, spatial, and spatiotemporal effects for a total of 12 models. GAMM are a semi-parametric model with likelihoods, as such we used AIC to determine the parsimony of the 12 GAMM forms, and then we used the DHARMa package to examine model misspecification for the GAMM with the lowest AIC.

*Random forest model (RF)*

Random forest models use a non-parametric method to classify the importance of different features (i.e., independent variables) for estimating the dependent variable (i.e., juvenile and adult Coho density). These models are particularly convenient if the distribution of the dependent variable is unknown, or something other than the assumed Tweedie distribution in the GLMM or GAMM models. Additionally, RF can be used to find the most relevant features of a model which can improve computational speed and prediction accuracy.

The RF algorithm is as follows: boot-strap the data, construct the decision tree by partitioning the dependent/covariate classifications to provide the lowest sums-of-squares error for each data subset, bag the decision tree classifications by fitting many boot-strapped datasets, and finally, cross-validate the error rates of the decision tree using the use the “out-of-bag” data not included in the bootstrapped data. Random forest are unique among machine learning algorithms, because they use a random set of dependent variables at each branch when “growing” a decision tree.

Random forest models require two user inputs: the number of covariates used at each split in the tree (*m*), and the number of trees to grow such that each covariate get predicted a couple of times (*n*). For our purposes, we completed a gridded search of *m* equal to 3, 5, 7, 9, 11 splits, and *n* equal to 200, 400, 600, 800, and 1000 trees. Methods for determining which dependent variables to include in the RF models are well documented for a suite of R packages (Degenhardt, Seifert, and Szymczak 2019); however, to maintain consistency between the GLMM, GAMM, and RF approaches, we included all of the same dependent variables in the models and with iteration of spatial, temporal, or spatiotemporal effects similar to the GLMM and GAMM models. The syntax for the model is

>R randomForest(Juv.km ~

StrmSlope + WidthM + OUT\_DIST +

SolMean + MWMT\_Index + W3Dppt +

StrmPow + SprPpt + IP\_COHO +

JuvYr + UTM\_E\_km + UTM\_N\_km, #spatiotermporal

mtry = 3, #Number of features/covariates

ntree = 600, #Number of trees

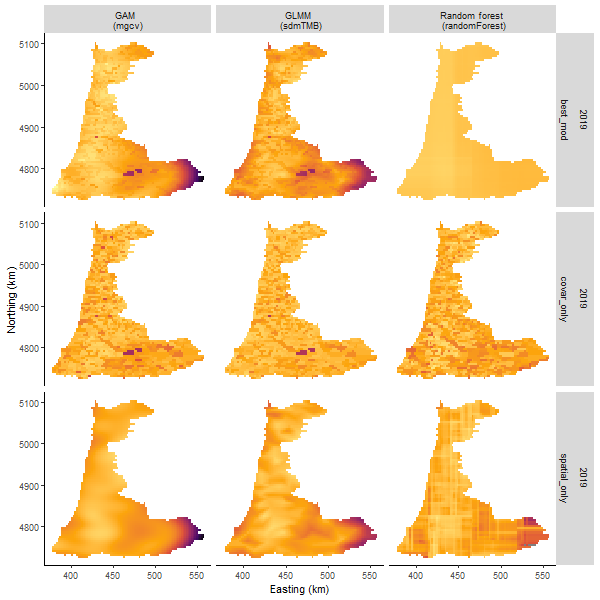
data=juvDat) #data

Because RF models are non-parametric, AIC is not appropriate of comparing different forms; instead, we used root-mean square error (RMSE) for the out-of-bag samples to compare forms.

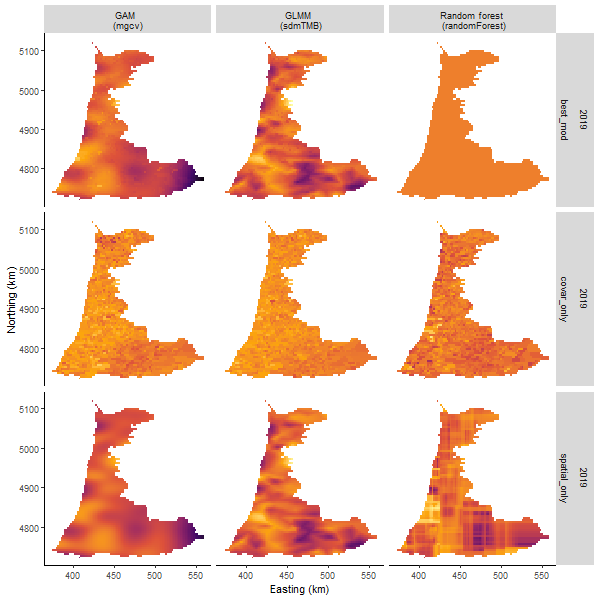
*Comparing the predictive ability of the models.*

To assess the predictive ability of the top models we trained each model by removing the last one or two years of data. Then we predicted the out-of-bag years and calculated the RMSE for the last five years of data (i.e., years with both training and out-of-bag data). The model with the lowest RMSE was assumed to be the model with the most predictive ability.

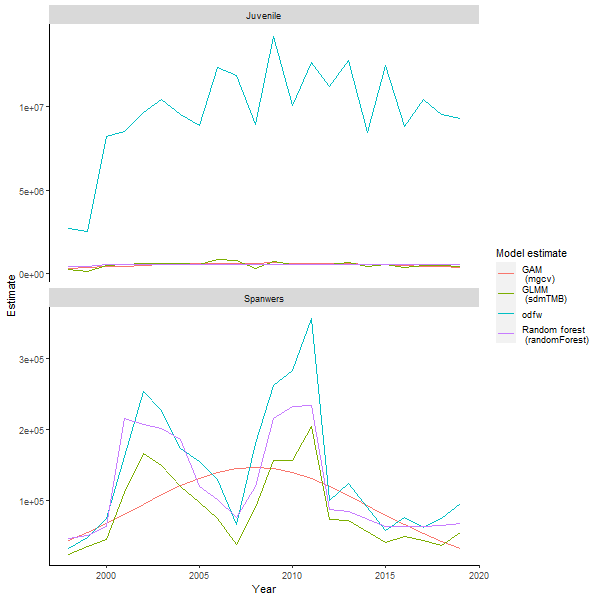
**Results**



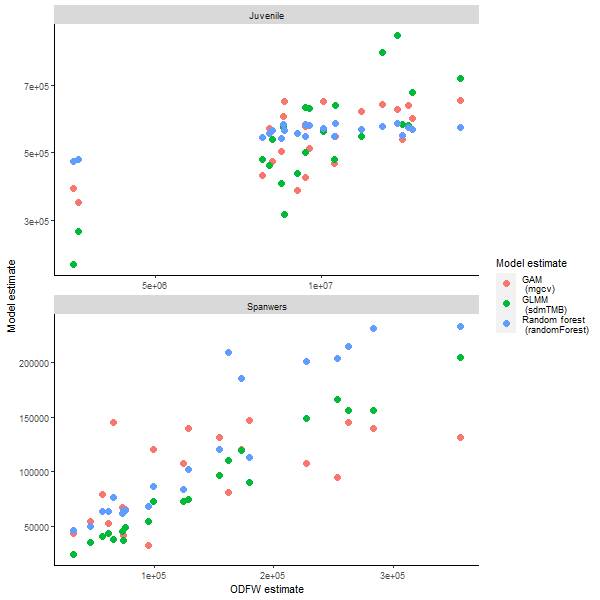
Predicted distribution of juvenile spawners during 2019 using the best fit, covariate-only, and the spatial-only forms of the GAMM, GLMM, and random forecast models (ggplot\_comp\_spatial\_by\_lifeStage.r).



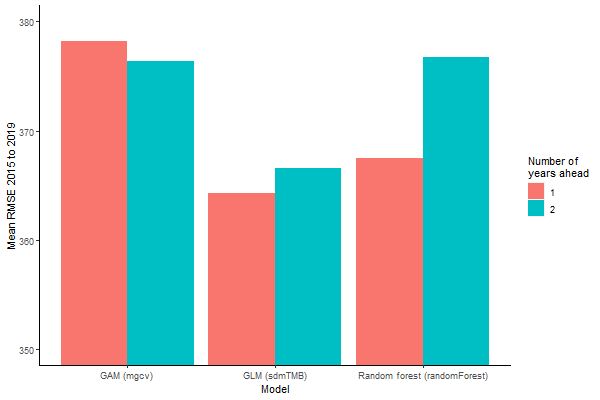
Predicted distribution of adult spawners during 2019 using the best fit, covariate-only, and the spatial-only forms of the GAMM, GLMM, and random forecast models (ggplot\_comp\_spatial\_by\_lifeStage.r).



Estimated indexes of juvenile and spawner abundance for Coho salmon in the Coastal Oregon ESU form 1998 to 2019 (ggplot\_spatial\_index.r).



Comparison between the indexes of juvenile and spawner abundance for Coho salmon in the Coastal Oregon ESU form 1998 to 2019 (ggplot\_spatial\_index2.r).



**Supplemental Material**

For the purposes of reproducibility, we created four wrapper functions in R. The first function, **function\_wrangle\_data.r**, reads in the data and returns a data frame of juvenile and spawner density for the Coastal Oregon Evolutionary Significant Unit. The second function, **function\_model\_search.r**, creates a tagged list of model formulas and arguments for each of the three model types: GAMM (R package mgcv), GLMM (R package sdmTMB), and random forest (R package RandomForest). For each model type (i.e., GAMM, GLMM, and RF), the third function, **function\_model\_exploration.r**, measures the performance of each model type for predicting historical data and projecting salmon densities into the future.