Received Date : 23-Mar-2015
Revised Date : 31-Jul-2015
Accepted Date : 31-Jul-2015
Article type : Article

Efficient spatial models for predicting the occurrence of subarctic estuarine-associated fishes: implications for management

K. B. Miller
Auke Bay Laboratories, Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA, Juneau, AK, USA
F. Huettmann
EWHALE lab, Institute of Arctic Biology, Biology & Wildlife Department, University of Alaska Fairbanks (UAF), Fairbanks, AL, USA
B. L. Norcross
School of Fisheries and Ocean Sciences, University of Alaska Fairbanks, Fairbanks, AK, USA

Running Title: Spatial Predictive Models of Estuarine Fish

Correspondence: Katharine B. Miller, Auke Bay Laboratories, Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA, 17109 Pt Lena Loop Rd, Juneau, AK 99801, USA (Email: katharine.miller@noaa.gov)

Abstract In many of the nearshore areas where development is most likely to occur, essential fish habitat data are incomplete and there is little information on species occurrence that can be used to inform management decisions. This research investigated the use of multivariate remotely sensed geomorphic and landscape data to develop accurate predictive models of subarctic, estuarine-associated fishes. The Random Forest algorithm was used to predict the

occurrence was 26 fish species captured in 49 estuaries in Southeast Alaska. Model prediction accuracy ranged from 100% to 42% for species presence and 87% to 15% for species absence. Model goodness-of-fit and accuracy was assessed by comparing the number of species occurrences predicted by the model against the observed presences and absences of species in an independent dataset. Sixty percent of the models were able to predict species presence with an accuracy of 70% or better. The models were used to predict species occurrence for 521 unsampled Southeast Alaskan estuaries to provide a regional map of predicted species distributions.

KEYWORDS:

Introduction

Understanding species spatial distributions in relation to environmental and habitat features is essential for effective management of marine ecosystems and fish habitat. Quantitative species distribution models can assist in identification of habitat for protection and spatial planning (Austin 2002; Maxwell *et al.* 2009; Valavanis *et al.* 2008), can increase understanding of ecosystem structure and function, can guide research on productivity of habitats for specific species, and can help predict changes in species occurrence as a result of climate change or invasive species (Williams *et al.* 2009;Mueter & Litzow 2008;Stojlgren *et al.* 2010). Predictive species distribution models have been widely used in terrestrial and freshwater ecological studies (Magness *et al.* 2010; Ohse *et al.* 2009; Lawler *et al.* 2011), but just in the last decade have they begun to be applied to marine ecosystems (Oppel & Huettmann 2010; Wei *et al.* 2010; Robinson *et al.* 2011).

In the United States, the Magnuson-Stevens Fishery Conservation and Management Act (2006) requires that activities undertaken by federal agencies, or with federal funding, be evaluated by the National Oceanic and Atmospheric Administration (NOAA) if those actions may adversely affect essential fish habitat (EFH). However, for many of the nearshore locations where human activities are most likely to be proposed, knowledge about EFH is incomplete or EFH has not been defined due to the limited number of site-specific surveys of fish occurrence or habitat use. This is especially true in places like Southeast Alaska where much of the 36 000 km

of shoreline is accessible only by boat or float plane yet has the potential to be adversely affected by logging, mining, shipping and other activities that require access to marine waters. Models that effectively predict the distribution of species from limited sample data can help improve the management of development activities and inform recommendations for mitigation of potential adverse impacts to fish habitat. This allows a precautionary approach to be achieved quantitatively.

Using limited sample data to predict species occurrence in unsampled locations requires similar environmental and habitat data for both sampled and unsampled areas (Cushman & Huettmann 2010). With the advent of remotely-sensed datasets, a large amount of potential habitat data has become available for use in predictive modeling. In terrestrial ecosystems, landscape-scale data such as elevation, vegetative cover and precipitation patterns are commonly used to predict species occurrence and abundance (Magness *et al.* 2010). Freshwater fisheries research also has demonstrated that landscape and stream geomorphic variables (e.g. channel complexity, stream size) can accurately predict the occurrence of freshwater fish (Brenden *et al.* 2007;Cèrèghino *et al.* 2005;Elmendorf & Moore 2008;Wilkins & Snyder 2011). In the marine environment, studies assessing how land and seascape structure and complexity influence species occurrence and abundance have generally been confined to work on coral reefs (Wedding & Friedlander 2008; Pittman *et al.* 2007). By contrast, this research investigates the use of multivariate remotely sensed geomorphic and landscape data to develop accurate predictive models of subarctic, estuarine-associated fish.

Development of accurate models predicting species occurrence or abundance is hampered by the complexity of species-habitat relationships and the processes that affect those relationships on a variety of spatial scales (Anderson *et al.* 2009; Elith & Leathwick 2009; Oppel & Huettmann 2010; Hardy *et al.* 2011; Huettmann & Diamond 2006). Particularly in estuaries, species tend to have wide environmental tolerances; their response to environmental factors can be inherently multivariate, nonlinear or discontinuous (Gutiérrez-Estrada *et al.* 2008;Mueter & Norcross 1999). Furthermore, the effects of most environmental factors do not occur in isolation from the effects of other factors thus requiring non-parametric methods that are capable of modeling complex interactions between predictors in a robust fashion. In spatial analysis of ecological data, it is also not uncommon for both the predictor and its response to be autocorrelated and clustered: habitats closer to one another will be more similar than habitats

farther apart, just as the probability of species occurrence at adjacent locations is likely to be higher than between more remote locations (Segurado *et al.* 2006). Nonlinearity and spatial autocorrelation present difficulties for traditional modeling methods (Zar 2009).

Regression methods have been commonly used to evaluate marine species-habitat relationships. However, methods such as Generalized Linear Models (GLM) are not well suited to model nonlinear relationships with high performance output (Elith *et al.* 2006). It is also often difficult for survey data to meet normality (parametric) assumptions of linear models. Generalized Additive Models (GAMs) provide a more flexible approach that can approximate non-linearity between the predictor variables and the response. GAMs have become very popular for modeling fish occurrence and abundance in relation to habitat variables (Mourato *et al.* 2014; Schmiing *et al.* 2013; Drexler & Ainsworth 2013) and they have been used to predict species abundance in unsampled areas (Drexler & Ainsworth 2013). However, for both GLMs and GAMs, spatial autocorrelation violates assumptions of independence in the data and potentially inflates the significance of the model results. Use of step-wise procedures to reduce model complexity is known to introduce additional bias because of the tendency for spatially correlated variables to be selected over non-correlated ones (Segurado *et al.* 2006; Whittingham *et al.* 2007). Both GAMs and GLMs are also quite limited in the number of predictor variables that can be used in the model requiring either subjective preselection of variables or the use of mathematical variable reduction techniques (e.g. principal components analysis (PCA)) to identify variables with the strongest relationship to the response (Guisan *et al.* 2002; Marra & Wood 2011). However, modern climate and landscape studies in ecosystems easily can include 20 predictors just to express the basic relationships (Cushman & Huettmann 2010; Drew *et al.* 2011); more predictors are usually needed for a good generalisation.

With the advent of powerful computers, machine learning algorithms have been developed that can handle large datasets composed of nonlinear and autocorrelated data (Cushman & Huettmann 2010; Drew *et al.* 2011). The random forest algorithm (Breiman 2001) is a machine learning algorithm that has highly accurate predictive ability (Prasad *et al.* 2006; Magness *et al.* 2010; Stojlgren *et al.* 2010) in a wide variety of applications. Random Forests (RFs) are collections of regression trees constructed by randomly drawing with replacement from the entire dataset to create separate training and testing data, and then randomly drawing with replacement from the predictor variables of the training dataset to construct individual decision

trees. Usually, several hundred regression trees are grown using this dual randomization approach. The best split for each predictor variable is determined by averaging the results across all trees. Correlation between individual trees is reduced by randomizing both the samples and the predictor variables, which decreases the error estimate of the entire ensemble. Predictor variables are considered individually when constructing the decision trees, and the data are unconstrained by the assumptions of the underlying distribution. By constructing individual trees on a subsample of the data, RFs can handle situations where the number of predictors is much larger than the number of samples (p >> n) (Strobl *et al.* 2009; Oppel & Huettmann 2010). This process also breaks relationships between variables and makes it possible for the RF to model the effect of nonlinear and interacting variables on species occurrence and identify the variables with the strongest influence on species distributions (see details and chapters in Cushman & Huettmann 2010; Drew *et al.* 2011; Breiman 2001).

In this research, predictive RF models were constructed for 26 species of fish captured in 49 estuaries in remote areas of Southeast Alaska from 1998 to 2005. To allow for a valid generalization and inference, the accuracy of the model predictions was evaluated against a set of independently collected data from an additional 71 estuaries during the same timeframe to quantitatively determine how well these models perform for research and management purposes.

Study site

The study area (Fig. 1) was the Alexander Archipelago: a collection of approximately 1000 mountainous islands in Southeast Alaska from Dixon Entrance at the Canadian border (54º 47’35”, 130º 38’06”W) to Lance Point in Lynn Canal (58º 44’ 141”, 135º 13’ 996”). The coastline is generally steep and the islands are separated by deep channels and fjords. The entire archipelago is a temperate rainforest: precipitation varies locally and regionally with a general gradient of lower precipitation in the northwest and higher precipitation in the southeast. Precipitation is strongly influenced by the coastal geology and topography (Weingartner *et al.* 2009). Average annual precipitation in the region is in excess of 1000 mm yr^{-1} (Neal *et al.* 2002) with much of the precipitation being released directly into the marine waters via numerous small streams and wetlands. Stream flow is highly seasonal and influenced both by precipitation and by snow and ice melt. The highest stream flows tend to occur in autumn when precipitation rates are high (Mundy *et al.* 2010). Flows decrease in winter as a result of freezing, and increase again

in the late spring and summer with the melting of snow and ice. The flow of fresh water affects not only the nearshore estuarine circulation, but is also the driver for larger-scale oceanographic circulation within Southeast Alaska's interior channels and on the continental shelf (Weingartner *et al.* 2009). Stream and river temperatures are influenced both by air temperatures and by runoff from glaciers, snowmelt, and precipitation.

The estuaries in the study area differ in their hydrological and geomorphological characteristics. In many Southeast Alaska estuaries, tidal energy is often much higher than energy from freshwater inflow (Weingartner *et al.* 2009). Southeast Alaska has mixed semi-diurnal tides with tidal height increasing as the tide moves from the continental shelf into the interior of the archipelago (Inazu *et al.* 2009). The difference in height between mean higher high water and mean lower low water is between 1.18 to 5.13 m in the study area(Noaa 2012). Tidal velocities are strongly influenced by bathymetry and channel morphology, and these, in turn, affect estuarine circulation, nutrient fluxes, and sediment dynamics (Weingartner *et al.* 2009). Coastal geology also varies greatly across estuaries in the study area (Harney *et al.* 2008). Most estuaries have a mixture of soft and hard substrate shorelines, but the amount of each type of substrate varies depending on both oceanographic and terrestrial processes.

For this research, 541 estuaries between the high tide line and the 30 -m depth contour were delineated in ArcGIS 10 ™ for model projection (Fig. 2). Glacier Bay was excluded from the research because circulation within the bay is constrained by the shallow sill at the entrance to this fjord system. As a result, the processes structuring fish and invertebrate communities within that bay are different than those in open estuaries elsewhere in Southeast Alaska (Matthews 1981).

Methods

Species data

Fish were sampled in 49 Southeast Alaska estuaries between 1998 and 2005 using both trawl and seine gear (Fig. 2, red dots). Sampling was conducted during daylight hours between February and September at high and low slack water. Fish were captured using an otter trawl (3 m x 1 m, with 6 mm square mesh in the cod end) deployed with a bridle scope of approximately 20 m. The trawl was towed at a speed of approximately 3 knots along a depth contour between 5 and 10 m. The exact depth of individual tows varied within this range depending on benthic structure of the

estuary. One tow in each direction was made along the same transect at high and low slack water (4 tows total). The latitude and longitude of the beginning and ending points of the trawl were recorded along with the average depth of tow. Fish also were sampled with a 37-m long variable mesh beach seine that tapered from 5 m wide at the center to 1 m at the ends. Outer panels were each 10 m with 32-mm stretch mesh, intermediate panels were each 4 m with 6-mm square mesh, and the bunt was 9 m with 2.3-mm square mesh. The net was set as a round haul by fixing one end on the beach, backing the skiff while deploying the net, and bringing the other end to shore approximately18 m from the first end. The latitude and longitude for each sample were recorded using a handheld Garmin GPSmap 76S™ with an accuracy of ± 3 m. Captured fish were identified to species and measured in the field to the nearest millimeter total length. Seventy percent of the estuaries were sampled only once.

The length at 50% maturity for commercially harvested fish and some forage fish was obtained from the Alaska Fishery Science Center's (AFSC) Life History database and was used to classify fish as adults or juveniles (AFSC 2011). For fish species not in the AFSC database, a variety of published sources was used to obtain length at 50% maturity information. For species that occurred in the data as a mix of juveniles and adults, life stages were pooled and modeled together. Most sites were sampled only once. The presence or absence of fish at each site was recorded by the month in which the sampling occurred.

To verify the predictive models, the Alaska Nearshore Fish Atlas (Johnson *et al.* 2005) was used as an independent dataset. This dataset is a compilation of seine sampling at 71 sites in Southeast Alaska from 1998 to 2004 (Fig. 2, yellow dots). To ensure independence of the data, estuaries from the Fish Atlas were only employed if they did not occur in the sample dataset used to develop the predictive models. Because of differences in gear and sampling method between the Fish Atlas data and the sampled data used in this study, only presence/absence data were used to develop predictive models. This allows for a robust comparison, because a presence record is a confirmed species occurrence, and it that presents a snapshot in time of presence and absence.

Environmental variables

Estuaries were standardized for size. For each estuary in the sampled, independent and unsampled datasets 107 predictor variables representative of major estuary components

(watershed structure, estuary structure, habitat, and hydrodynamics) were identified. This is the most complete dataset of estuarine predictor data available for Alaska. For each of these variables, remotely sensed landscape data were used as surrogates for environmental data measured in situ. The GIS data layers were acquired from the Southeast Alaska GIS library (Seak 2013), NOAA (Noaa 2013), and the Alaska ShoreZone database (Shorezone 2012) (Table 1). Variables that described the structure of the estuary included estuary length, width at mouth, area, perimeter, intertidal area, depth, bathymetric slope, and open water area. Estuary length was measured as the distance from the high tide line to the estuary mouth. The open water area was the amount of open water at low tide and was the difference between the estuary area and the intertidal area. Each estuary was assigned a classifier for the type of water body into which the estuary drains (bay, inlet, channel, or open ocean). This was the only categorical variable in the predictor variable dataset. For those estuaries draining into a bay or inlet, the distance from the estuary mouth to open water, either a major channel or the open ocean, was measured and included as a variable.

Variables describing the watershed surrounding the estuaries included the size and slope of the watershed, the type and amount of land cover (e.g. vegetation, bare land, development, glaciers), and the degree of land cover fragmentation. Watershed size was derived from 12-digit hydrologic units (USGS 1995). A 5-km buffer was placed around each estuary in the GIS and the watershed slope was measured within this buffer using a digital elevation model. The buffer size was chosen because it corresponded to the size of the smallest watershed. The 2001 National Land Cover Dataset (NLCD) for Alaska (2001) was used to extract percent area of vegetation, bare land, development, and glacial ice in the 1-km buffer using ArcGIS™. FRAGSTATS™ (Mcgarigal *et al.* 2002) was used to calculate five measures of land cover patchiness: total area, number of patches, patch density, largest patch index and landscape shape index from the clipped grids.

To capture the influence of fresh water on estuarine communities, minimum monthly precipitation over the study period was compiled from the Parameter-elevation Relationships on Independent Slopes (PRSM) climate model for Alaska (Snap 2011). Fluvial flow was calculated after the method of Digby *et al*. (1998) by multiplying the catchment area with the average annual rainfall and a runoff coefficient divided by the open water area of the estuary. At the scale of this analysis, the runoff coefficient was essentially a constant across all watersheds. The total

length of streams within the 1-km buffer around the estuary and in the intertidal area was measured using the USFS streams data layer for the Tongass National Forest(USDA 2002). Surface salinity data for Southeast Alaska are not available at the spatial scale of this research, so precipitation and fluvial flow variables were included in the analysis to capture differences in salinity and buoyancy-driven circulation between estuaries.

The great diurnal tide range (difference between mean higher high water and mean lower low water) and mean tide range for each estuary was compiled from NOAA tide data (Noaa 2012). Estuaries without measured tidal data were attributed the tidal ranges from the nearest estuary with tidal data. Estuary depth and slope, and the depth and location of bars or sills were included as predictor variables to capture the influence of bathymetry on tidal energy and flow. Bathymetric contours were evaluated to identify bars/sills, which were defined as constrictions where minimum depths were half or less of the average depth of the estuary. The distance between the bar/sill and the estuary mouth was measured for each estuary in which they occurred.

Geomorphological and biological characteristics of the intertidal portion of each estuary were obtained from the Alaska ShoreZone dataset (http://fakr.noaa.gov/shorezone/default.htm). ShoreZone is a mapping and classification system that uses oblique, low altitude aerial video and still images to classify segments of the shoreline according to natural breaks in geomorphic, sedimentary, and biological features (Harney *et al.* 2008). Shorezone coastal class, which is an index of substrate type, sediment type, across-shore width and slope, was included as the proportion of the estuary in each coastal class. Variables for the percentage of continuous and patchy subtidal red algae (e.g. *Neorhodomela* sp.), *Alaria*, soft brown kelps (i.e. *Saccharina latissima*), dark brown kelps (i.e. stalked *Laminaria* sp.), and eelgrass (*Zostera marina*) were also included. The shorezone habitat class combines the physical and biological information for a shoreline unit into a single variable that describes the intertidal biota together with the geomorphology (Harney *et al.* 2008). The proportion of each Shorezone habitat class was also evaluated in the model.

Model development

Random Forest models were built for the occurrence of 26 fish species from 15 families using the open source randomForest package for R (Liaw & Wiener 2002). In the Random Forest

algorithm, the number of variables selected for prediction (mtry) and, to a lesser extent, the number of trees grown in a Forest (ntree) can be used to tune the models to obtain better results. For each fish species, forests with 100, 500, 1000 and 1500 trees were grown using from 1 to 30 variables (mtry) per tree and the model with the best predictive accuracy was selected. The receiver operating characteristic (ROC) was used to evaluate model performance for each species (Brown & Davis 2006). The ROC curves were created using the pROC package in R for each model using the relative index of occurrence predicted by the models.

Model goodness of fit was assessed by comparing species occurrence predicted by the model against the observed occurrence in the independent data. The point on the ROC curve where the slope of the curve is equal to 1, or the highest sum of specificity and sensitivity (Jimenez-Valverde & Lobo 2007) was used as the probability threshold above which species were considered to be present. This was applied to the relative index of species occurrences (RIO) from the models to obtain the percentage of correctly classified presences and absences, and the model results were sorted by the percentage of presences accurately predicted. Each model was also used to predict species occurrence in the 541 unsampled estuaries. A map for the predicted occurrence of each species was saved as a shapefile that could be used alone or in association with other species data to evaluate co-occurrence patterns.

Predictor variable importance in Random Forest is calculated by permuting the predictor variables individually in the testing data and measuring the decrease in prediction accuracy for models computed with the permuted data compared to models computed with the original data. If model accuracy decreases with the permuted value, this indicates that the variable has a strong association to the response (Liaw & Wiener 2002). To explore relationships between species occurrence and the predictor variables, partial dependence plots were constructed for the 30 most influential variables.

Results

Of the 26 species captured, chum salmon, *Oncorhynchus keta* (Walbaum), was the most numerous fish species, comprising 12% of the total catch across all sites and years, followed closely by pink salmon, *Oncorhynchus gorbuscha* (Walbaum) with 11% of the catch. Three species were captured at over 60% of the sites: Pacific staghorn sculpin, *Leptocottus armatus*

(Girard), rock sole, *Lepidopsetta* sp. (Gill), and crescent gunnels, *Pholis laeta* sp. (Cope). Several fish species occurred more frequently or entirely as juveniles in both the sample and independent data. Species occurring only as juveniles included all four species of salmon, Pacific herring, *Clupea palasii* (Valenciennes), Pacific cod, *Gadus microcephalus* (Tilesius), lingcod, *Ophiodon elongatus* (Girard), kelp greenling (*Hexagrammos decagrammus* (Pallas), silverspot sculpin, *Blepsia scirrhosis* (Pallas), butter sole, *Isopsetta isolepis* (Lockington), and great sculpin, *Myoxocephalus polyacanthocephalus* (Pallas). Species whose abundance was predominantly composed of juveniles were yellowfin sole, *Limanda aspera* (Pallas), rock sole, and Pacific sand lance, *Ammodytes hextaperus* (Pallas). Species with mixes of juveniles and adults were starry flounder, *Platichthys stellatus* (Pallas), Pacific staghorn sculpin, and shiner perch, *Cymatogaster aggregata*, Gibbons. Species for which life stage could not be determined from the literature were the snake prickleback, *Lumpenus sagitta* (Wilimovsky), tube-snout, *Aulorhynchus flavidus* (Gill), and sturgeon poacher, *Podothecus accipenserinus* (Tilesius).

The AUCs of the Random Forest models for the individual species ranged between 0.94 and 0.63. Prediction accuracy ranged from 100 to 42% for species presence and 87 to 15% for species absence (Table 2). Strong models were defined as those with presence prediction accuracy of 80% or higher, moderate models as those with prediction accuracies between 70% and 80%, and poor models as those with prediction accuracies below 70%. The predictive accuracy of the models was generally lower for species absences than for species presence. Sixty percent of the Random Forest models were able to predict species presence in the independent data with an accuracy of 70% or better, but only32% of the models could predict both presence and absence with moderate to good accuracy. Model strength was not related to how common or rare a species was.

Predictor variable importance differed among models. Unlike modeling methods that fit a few pre-selected variables to species occurrence based on an assumed inference, the random forest algorithm identifies a unique set of predictors of highest importance to each species from all the predictors in the offered data set. Certain classes of variables occurred in nearly all models. These included variables for estuary depth and slope, intertidal area, precipitation, Shorezone vegetation and Shorezone substrate and habitat classes. For example, substrate type was an important habitat feature for all flatfish models, but substrate preferences differed by species. Yellowfin sole had higher predicted occurrence in areas with wide (> 30 m) intertidal

areas composed of sand and mud. Starry flounder and rock sole showed a preference for narrow intertidal areas composed of sand and gravel. Butter sole occurrence was associated with wide (> 30 m) gravel beaches. Most of the important predictors agreed with known species habitat preferences. For example, occurrence probability for the bay pipefish and the kelp greenling increased at higher percentages of patchy and continuous eelgrass which is a common juvenile habitat for these species (Johnson *et al.* 2005;Hosack *et al.* 2006). Stream area was important to anadromous pink and coho salmon occurrence. Variables describing estuary size and connectivity were important in fewer than 5% of the models while watershed variables (watershed size, slope and percent and type of cover) had the lowest occurrence and importance across models. Model strength was higher for seasonal residents than year around residents, with 76% of the seasonal residents having moderate to strong models; however, sampling month was only an important predictor for five of the eight strong-moderate model species: shiner perch, lingcod, pink salmon, chum salmon and Pacific cod. Figure 3 shows variable importance plots for the six most important variables for each of these species. The x-axis in these plots is the increase in the model mean squared error (decrease in performance) when the variable is permuted. The y-axis is the code for the predictor variable listed in decreasing order of importance. For shiner perch and lingcod, month was the most important variable among other influential variables, but for pink salmon, chum salmon, and Pacific cod month was the most influential variable by a wide margin resulting in a 14% to 25% change in the model mean squared error when the variable was permuted.

The univariate relationship between an important predictor variable and species occurrence for each model can be visualized by examining the partial dependence plots for the variable. In the case of the variable “month” these plots reflect seasonal estuarine use (Fig. 4). Shiner perch, which occurred primarily as adults, had higher predicted abundance in mid-summer when it migrates into estuaries to spawn. Juvenile lingcod and Pacific cod also show higher predicted abundance in late summer. In contrast, juvenile pink and chum salmon are abundant in estuaries in early spring as they transition from freshwater to saltwater and are rarely found in the estuaries in mid-summer.

Discussion

Species distribution models that accurately predict species occurrence in relation to habitat characteristics provide an important tool for habitat protection and ecosystem management; they can provide essential habitat information for fisheries decisions. However, the ability to develop accurate models based on little field sampling in complex areas is often constrained by a lack of *in-situ* habitat data and the modeling approach used. Although remotely-sensed environmental data are commonly used to develop distribution models for terrestrial species, the quality, accessibility and application of these data to marine species has been limited. These results demonstrate the great potential of using landscape scale data with random forest models to predict the occurrence of subarctic estuarine fish species. This machine learning method is able to extract a valid predictive signal from the data. With these relatively simple approaches, two-thirds of the models had a predicted accuracy above 70% for species presence when assessed against independent data.

Predicted species occurrences from the models can be mapped using GIS to evaluate spatial patterns of occurrence throughout the region. These maps may also be overlaid on each other for insight on community composition and habitat partitioning making it possible for habitat managers to evaluate the importance of specific areas at the individual species and community level. As an example of this, the predicted distribution of shiner perch (orange) and lingcod are shown in Figure 5. Shiner perch is among the most abundant species in coastal areas of Southeast Alaska (Johnson *et al.* 2005) occurring in large numbers in association with subtidal vegetation, particularly eelgrass (Murphy *et al.* 2000;Johnson & Thedinga 2006). Lingcod also seeks structurally complex habitat that may include sublittoral kelp and eelgrass, but may also be associated with hard substrates (rocks or shell) or anthropogenic structures (Petrie & Ryer 2006). The overlap in the predicted distribution for these species occurs primarily along the outer coast and outer portions of bays along the major channels (Fig. 5). Maps such as this provide insights and information that is not currently available in any other way and would be difficult and costly to obtain through sampling in such a large and complex area. In addition to being used to inform habitat management decisions, these maps can also suggest areas for targeted sampling or in-situ studies as a condition of a development permit.

Distributions of species can also be evaluated in relation to specific predictor variables; however, the relationship between species distribution patterns and individual predictor variables can be difficult to interpret in multi-variable models where variables are known to interact.

Random forest models are not frequency statistics and thus do not assess the statistical significance of variables to the response in the same manner as parametric regression methods. The strength of RF models lies in the ability to identify a suite of variables with strong influence on the predictive power of each species model from a large number of potential predictors and infer from the obtained prediction (Breiman 2001).

The variables in the RF model may serve as surrogates for a number of processes that cannot be easily measured. This is one of the advantages of the RF modeling approach: it allows the development of accurate predictive models using all relevant and available data. As a result, the models are not constrained by predetermined assumptions of what the underlying model is, how variables interact, which variable is more important, or what each variable's function is in the model. As an example, minimum precipitation was among the most important variables for many of the strong and moderate models, with individual species exhibiting positive or negative associations.

Precipitation and fluvial flow variables can be surrogates for measures of salinity and buoyancy-driven circulation. The presence of habitat features that are sensitive to salinity, such as eelgrass, may influence the relationship between precipitation and species occurrence in the models. Precipitation also may be a proxy for other oceanographic processes, such as stratification which can enhance primary productivity and food availability. High freshwater discharge is associated with the development of tidal fronts which are areas of mixing that occur at the interface between stratified water and well mixed saline water as a result of tidal inflow into the estuary. These fronts may act as barriers to larval transport helping to retain and distribute planktonic larvae within the estuary (Genuer *et al.* 2010; Svetnik *et al.* 2002).

Variables may also represent spatial patterns that would be difficult to capture in other ways. Starry flounder and butter sole models showed opposite trends with respect to precipitation (Fig. 6). Starry flounder was less likely to occur in areas with high precipitation and butter sole were more likely to occur in those locations. Starry flounder is a eurohaline species that spawns in low salinity areas of the upper estuary (Tomiyama & Omori 2008;Wada *et al.* 2007). It is found throughout Southeast Alaska and often captured in small trawls and seine gear. By contrast, butter sole is relatively infrequently caught in nearshore surveys. It moves offshore to spawn and moves into estuaries after its first year of life (Richardson *et al.* 1979). Given these life history characteristics, the relationship between these two species and precipitation would be

expected to be the opposite of what the models show if precipitation was acting as a surrogate for salinity. In this case, however, precipitation is reflecting climactic patterns that are driven by coastal geography and topography. Precipitation is higher along the coastal mountains where storm systems from the Gulf of Alaska first make landfall. These areas are also subject to offshore winds that advect water from the shelf into coastal bays and estuaries (Weingartner *et al.* 2009). The random forest models predict higher butter sole occurrence along the coast adjacent to spawning locations on the shelf where larvae can be advected into adjacent estuarine rearing areas. For starry flounder, the models predict a broad distribution within interior estuaries.

Capturing coastal morphology into a variable that could be used to model fish occurrence would be extremely difficult. Precipitation can act as a surrogate for these data, but knowledge of the species ecology and life history is necessary for correct interpretation. Additional hypothesis testing may be needed to understand the importance and influence of model variables. In this research, the focus was on the predictive accuracy of the models. Associations between predictors and response identified in the models provide insight into the ecological relationships that can be further evaluated with additional data or research.

Species distribution models predict the realized distribution of a species, not necessarily the potential distribution (Cushman & Huettmann 2010; Magness *et al.* 2010; Lawler *et al.* 2011). Most of the models in this study were able to predict species presence with a higher rate of accuracy than prediction of species absence. In many cases, the models predict that species should be present in a location when the data indicate that they are not (false positive). Species may not occur in suitable habitats for a variety of reasons, including biotic interactions such as competition or predation, or dispersal limitations. These factors were not accounted for in the random forest models. Predicting that a species is absent when it is actually present (false negative predictions) may also occur. These may be the result of imperfect species detection from sampling inefficiencies, gear selectivity, gear efficiency, or species catch (Hattab *et al.* 2013). In this analysis, false negatives could be the result of different capture efficiencies for the different gear used in the sample and independent datasets. For cryptic or hard to capture species, it is likely that "absences" from the data are not absences from the habitat but a rather a failure to capture the species during sampling. False positives could result in management or

protection measures being taken in an area where the species does not occur while false negatives might result in failure to protect a species where it does occur.

Determining that a species is absent from a habitat is much more difficult than determining that a species is present. Acquiring true absence data requires a substantial sampling effort including repetitive sampling which, over large spatial scales in complex habitats, is generally cost prohibitive. For this research, during the 10 years for which annual data were collected most of the estuaries could only be sampled a single time. The issue of false positives could be improved by incorporating additional information on species interactions and co-occurrence, which would provide insight into the importance of biotic interactions on species distributions. Including predator or prey species as predictor variables (Leathwick & Austin 2001) and incorporating dispersal vectors (Boulangeat *et al.* 2012) are two approaches that have been used in species distribution models to account for biotic interactions. To include biotic interactions it would be necessary to have detailed data on predator-prey dynamics within the estuary and these are largely unknown for many species important to the ecosystem but not of commercial interest, such as sculpin (Spies *et al.* 2011). The models could also be refined by incorporating *in-situ* sampling performed as a condition of permitting in locations where species were predicted to occur but are not recorded as present in the data.

The models developed in this research can provide habitat managers with refined data to use in assessing potential impacts to EFH especially in nearshore waters where EFH may not be well defined or defined at all. There are no EFH designations for commercially fished species in the interior waters of Southeast Alaska (NPFMC 2009), yet a third of the commercially managed species occur in estuaries here as either juveniles or adults (Lorenz 2005). An example is Pacific cod for which EFH has only been designated along the shelf and therefore does not overlap predicted occurrence of Pacific cod in estuaries in Southeast Alaska (Fig. 7). The majority of non-fishing impacts to EFH will occur in areas adjacent to the coast where EFH has not been defined. Further, state managed fisheries for groundfish, including Pacific cod and lingcod and several flatfish species, also occur within the interior waters which are not designated EFH. These models can provide managers with a science-based justification for protecting nearshore habitat for these species.

Acknowledgements

We would like to thank Mitch Lorenz, Scott Johnson, and John Thedinga, Research Fisheries Scientists, NOAA for the use of their fish sampling data.

Disclaimer

The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the National Marine Fisheries Service. Reference to trade names does not imply an endorsement by the National Marine Fisheries Service, NOAA.

References

Magnuson-Stevens Fishery Conservation and Management Act of 2006. *16 U.S.C. 1801 § 3 (10)*. United States.

AFSC (2011) Alaska Fisheries Science Center, Resource Ecology and Fisheries Management Division Life History Database. *NOAA Fisheries* *http://access.afsc.noaa.gov/reem/LHWeb/Index.php*. Seattle, WA.

Anderson T. J., Syms C., Roberts D. A. & Howard D. F. (2009) Multi-scale fish-habitat associations and the use of habitat surrogates to predict the organization and abundance of deep-water fish assemblages. *Journal of Experimental Marine Biology and Ecology* **379,** 34-42.

Austin M. P. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling* **157,** 101-118.

Boulangeat I., Gravel D. & Thuiller W. (2012) Accounting for dispersal and biotic interactions to disentangle drivers of species distributions and their abundances. *Ecology Letters* **15,** 584-593.

Breiman L. (2001) Random Forests. *Machine Learning* **45,** 5-32.

Brenden T. O., Wang L., Clark Jr. R. D. & Seelbach P. W. (2007) Comparison between model-predicted and field-measured stream habitat features for evaluating fish assemblage-habitat relationships. *Transactions of the American Fisheries Society* **136,** 580-592.

Brown C. D. & Davis H. T. (2006) Receiver operating characteristic curves and related decision measures: A tutorial. *Chemometrics and Intelligent Laboratory Systems* **80,** 24-38.

Cèrèghino R., Santoul F., Compin A., Figuerola J. & Mastrorillo S. (2005) Co-occurrence patterns of some small-bodied freshwater fishes in southwestern France: implications for fish conservation and environmental management. *Ambio* **34,** 440-444.

Cushman S. & Huettmann F. 2010. *Spatial Complexity, Informatics and Wildlife Conservation,* Tokyo, Japan, Springer.

Digby M. J., Saenger P., Whelan M. B., Mcconchie D., Eyre B., Holmes N. & Bucher D. (1998) A physical classification of Australian estuaries. *Centre for Coastal Management, Southern Cross University,Urban Water Research Association of Australia, (http://au.riversinfo.org/library/nrhp/estuary_clasifn/).* Lismore, NSW.

Drew C. A., Wiersma Y. & Huettmann F. 2011. *Predictive Modeling in Landscape Ecology,* New York, NY, Springer.

Drexler M. & Ainsworth C. H. (2013) Generalized additive models used to predict species abundance in the Gulf of Mexico: an ecosystem modeling tool. *PLoS One* **8,** e64458.

Elith J., Graham C. H., Anderson R. P., Dudik M., Ferrier S., Guisan A. *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **29,** 129-151.

Elith J. & Leathwick J. (2009) Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology and Systematics* **40,** 677-697.

Elmendorf S. C. & Moore K. A. (2008) The use of community-composition data to predict the fecundity and abundance of species. *Conservation Biology* **22,** 1523-1532.

Genuer R., Poggi J.-M. & Tuleau-Malot C. (2010) Variable selection using random forests. *Pattern Recognition Letters* **31,** 2225-2236.

Guisan A., Edwards Jr T. C. & Hastie T. (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling* **157,** 89-100.

Gutiérrez-Estrada J. C., Vasconcelos R. & Costa M. J. (2008) Estimating fish community diversity from environmental features in the Tagus estuary (Portugal): Multiple linear regression and artificial neural network approaches. *Journal of Applied Ichthyology* **24,** 150-162.

Hardy S., Lindgren M., Hanumantharao K. & Huettmann F. (2011) Predicting the distribution and ecological niche of unexploited snow crab (*Chionoecetes opilio*) populations in

Alaska waters: A first open-access ensemble model. *Integrative and Comparative Biology* **51,** 608-622.

Harney J., Morris M. & Harper J. (2008) Shorezone coastal habitat mapping protocol for the Gulf of Alaska. *Coastal & Ocean Resources, Inc. (http://www.fakr.noaa.gov/shorezone)*.

Hattab T., Ben Rais Lasram F., Albouy C., Sammari C., Romdhane M. S., Cury P., Leprieur F. & Le Loc'h F. (2013) The use of a predictive habitat model and a fuzzy logic approach for marine management and planning. *PLoS One* **8,** e76430.

Hosack G. R., Dumbauld B. R., Ruesink J. L. & Armstrong D. A. (2006) Habitat associations of estuarine species: comparisons of intertidal mudflat, seagrass (*Zostera marina*), and oyster (*Crassostrea gigas*) habitats. *Estuaries and Coasts* **29,** 1150-1160.

Huettmann F. & Diamond A. W. (2006) Large-scale effects on the spatial distribution of seabirds in the Northwest Atlantic. *Landscape Ecology* **21,** 1089-1108.

Inazu D., Sato T., Miura S., Ohta Y., Nakamura K., Fujimoto H., Larsen C. F. & Higuchi T. (2009) Accurate ocean tide modeling in Southeast Alaska and large tidal dissipation around Glacier Bay. *Journal of Oceanography* **65,** 335-347.

Jimenez-Valverde A. & Lobo J. M. (2007) Threshold criteria for conversion of probability of species presence to either-or presence-absence. *Acta Oecologica* **31,** 361-369.

Johnson S. W., Neff A. D. & Thedinga J. (2005) An atlas on the distribution and habitat of common fishes in shallow nearshore waters of southeastern Alaska. *U.S. Department of Commerce NOAA Technical Memo NMFS-AFSC-157***,** 89.

Johnson S. W. & Thedinga J. (2006) Fish use and size of eelgrass meadows in Southeastern Alaska: a baseline for long-term assessment of biotic change. *Northwest Science* **79,** 141-155.

Lawler J. J., Wiersma Y. & Huettmann F. 2011. Designing predictive models for increased utility: Using species distribution models for conservation planning, forecasting and risk assessment. *In:* C. A. Drew, Y. Wiersma & F. Huettmann (eds.) *Predictive Modeling in Landscape Ecology.* New York, NY: Springer**,** pp 271-290.

Leathwick J. & Austin M. P. (2001) Competitive interactions between tree species in New Zealand's old-growth indigenous forests. *Ecology* **82,** 2560-2573.

Liaw A. & Wiener M. (2002) Classification and regression by randomForest. *R News* **2/3,** 18-22.

Lorenz J. M. (2005) Appendix 1: Fish Habitat Assessment and Classification of Alaskan Estuaries. *Stock Assessment and Fishery Evaluation Report for the Groundfish Resources of the Gulf of Alaska* Anchorage, AK: NPFMC,,North Pacific Fishery Management Council.

Magness D. R., Huettmann F. & Morton J. M. 2010. Using random forests to provide predicted species distribution maps as a metric for ecological inventory and monitoring programs. *In:* T. G. Smolinski, M. G. Milanova & A.-E. Hassanien (eds.) *Applications of Computational Intelligence in Biology.* New York: Springer**,** pp. 209-229.

Marra G. & Wood S. N. (2011) Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis* **55,** 2372-2387.

Matthews J. B. (1981) The seasonal circulation of the Glacier Bay, Alaska fjord system. *Estuarine, Coastal and Shelf Science* **12,** 679-700.

Maxwell D. L., Stelzenmuller V., Eastwood P. D. & Rogers S. I. (2009) Modelling the spatial distribution of plaice (*Pleuronectes platessa*), sole (*Solea solea*), and thornback ray (*Raja clavataI*) in UK waters for marine management and planning. *Journal of Sea Research* **61,** 258-267.

Mcgarigal K., Cushman S. & Ene E. (2002) FRAGSTATS: Spatial pattern analysis program for categorical maps.Available at the following website: http://www.umas.edu/landco/research/fragstats/fragstats.html. Amherst: University of Massachusetts.

Mourato B. L., Hazin F., Bigelow K., Musyl M., Carvalho F. & Hazin H. (2014) Spatio-temporal trends of sailfish,Istiophorus platypteruscatch rates in relation to spawning ground and environmental factors in the equatorial and southwestern Atlantic Ocean. *Fisheries Oceanography* **23,** 32-44.

Mueter F. J. & Litzow M. A. (2008) Sea ice retreat alters the biogeography of the Bering Sea continental shelf. *Ecological Applications* **18,** 309-320.

Mueter F. J. & Norcross B. L. (1999) Linking community structure of small demersal fishes around Kodiak Island, Alaska to environmental variables. *Marine Ecology Progress Series* **190,** 37-51.

Mundy P. R., Allen D. M., Boldt J., Bond N. A., Dressel S., Farley E. V. *et al.* (2010) Status and trends of the Alaska Current region. *In:* M. S. M. & M. J. Dagg (eds.) *Marine Ecosystems of the North Pacific Ocean, 2003-2008.* PICES Special Publication 4**,** pp 393-387.

Murphy M. L., Johnson S. W. & Csepp D. J. (2000) A comparison of fish assemblages in eelgrass and adjacent subtidal habitats near Craig, Alaska. *Alaska Fishery Bulletin* **7,** 11-21.

Neal E. G., Walter M. T. & Coffeen C. (2002) Linking the Pacific decadal oscillation to seasonal stream discharge patterns in Southeast Alaska. *Journal of Hydrology* **263,** 188-197.

Noaa. 2012. *NOAA Tides and Currents* [Online]. National Oceanic and Atmospheric Administration,Center for Operational Oceanographic Products, http://tidesandcurrents.noaa.gov. Available: http://tidesandcurrents.noaa.gov.

Noaa (2013) NOAA Tides and Currents (http://tidesandcurrents.noaa.gov/). Accessed on 2/2/2013.

NPFMC, North Pacific Fishery Management Council (2009) Amendment 91, Bering Sea/Aleutian Islands Groundfish Fisheries Management Plan. Anchorage, AK North Pacific Fishery Management Council.

Ohse B., Huettmann F., Ickert-Bond S. & Juday G. (2009) Modeling the distribution of white spruce (*Picea glauca*) for Alaska with high accuracy: an open access role-model for predicting tree species in the last remaining wilderness areas. *Polar Biology* **32,** 1717-1724.

Oppel S. & Huettmann F. 2010. Using a random forest model and public data to predict the distribution of prey for marine wildlife management. *In:* S. Cushman & F. Huettmann (eds.) *Spatial Complexity, Informatics and Wildlife Conservation.* Tokyo, Japan: Springer**,** pp. 151-164.

Petrie M. E. & Ryer C. (2006) Laboratory and field evidence for structural habitat affinity of young-of-the-year lingcod. *Transactions of the American Fisheries Society* **135,** 1622-1630.

Pittman S. J., Christensen J. D., Caldow C., Menza C. & Monaco M. E. (2007) Predictive mapping of fish species richness across shallow-water seascapes in the Caribbean. *Ecological Modelling* **204,** 9-21.

Prasad A. M., Iverson L. R. & Liaw A. (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* **9,** 181-199.

Richardson S. L., Dunn J. R. & Naplin N. A. (1979) Eggs and larvae of butter sole, *Isopsetta isolepi*, (Pleurinectidae), off Oregon and Washington. *Fisheries Bulletin* **78,** 401-419.

Robinson L. M., Elith J., Hobday A. J., Pearson R. G., Kendall M. A., Possingham H. P. & Richardson A. J. (2011) Pushing the limits in marine species distribution modeling: lessons from the land present challenges. *Global Ecology and Biogeography* **20,** 789-802.

Schmiing M., Afonso P., Tempera F. & Santos R. S. (2013) Predictive habitat modelling of reef fishes with contrasting trophic ecologies. *Marine Ecology Progress Series* **474,** 201-216.

Seak (2013) Southeast Alaska GIS Library (http://seakgis.alaska.edu/) Accessed on 2/2/2013.

Segurado P., Araújo M. B. & Kunin W. E. (2006) Consequences of spatial autocorrelation for niche-based models. *Journal of Applied Ecology* **43,** 433-444.

Shorezone (2012) NOAA Fisheries ALaska Shorezone (http://alaskafisheries.noaa.gov/shorezone/) Accessed on 2/2/2013.

Snap. 2011. University of Alaska, (http://www.snap.uaf.edu/data.php). Available: www.snap.uaf.edu.

Spies I., Ormseth O. A., Martin M. & Tenbrink T. T. (2011) Asssessment of the sculpin complex in the Gulf of Alaska. *Stock Assessment and Fishery Evaluation Report.* Anchorage, AK: NPFMC,,North Pacific Fishery Management Council.

Stojlgren T. J., Ma P., Kumar S., Rocca M., Morisette J. T., Jarnevish C. S. & Benson N. (2010) Ensemble habitat mapping of invasive plant species. *Risk Analysis* **30,** 224-235.

Strobl C., Malley J. & Tutz G. (2009) An introduction to recursive partitioning. *University of Munich Technical Report Number 55*.

Svetnik V., Liaw A. & Tong C. Variable selection in random forest with application to quantitative structure-activity relationship. *In:* N. Intrator & F. Masulli, eds. 7th Course on Ensemble Methods for Learning Machines, September 22-28, 2002, 2002 Salterno, Italy Springer-Verlag.

Tomiyama T. & Omori M. (2008) Habitat selection of stone and starry flounders in an estuary in relation to feeding and survival. *Estuarine, Coastal and Shelf Science* **79,** 475-482.

USDA (2002) High and low tidelines. *USDA Forest Service, Tongass National Forest, Southeast Alaska GIS Library, (http://seakgis.alaska.edu).*

USGS (1995) Alaska hydrologic units. *U.S. Geological Survey (http://nhd.usgs.gov/data.html) accessed: 2009*.

Valavanis V. D., Pierce G. J., Zuur A. F., Palialexis A., Saveliev A., Katara I. & Wang J. (2008) Modelling of essential fish habitat based on remote sensing, spatial analysis and GIS. *Hydrobiologia* **612,** 5-20.

Wada T., Aritaki M., Yamashita Y. & Tanaka M. (2007) Comparison of low-salinity adaptability and morphological development during the early life history of five pleuronectid flatfishes, and implications for migration and recruitment to their nurseries. *Journal of Sea Research* **58,** 241-254.

Wedding L. M. & Friedlander A. M. (2008) Determining the influence of seascape structure on coral reef fishes in Hawaii using a geospatial approach. *Marine Geodesy* **31,** 246-266.

Wei C.-L., Rowe G. T., Escobar-Briones E., Boetius A., Softwedel T., Caley M. J. *et al.* (2010) Global patterns and predictions of seafloor biomass using random forests. *PlosOne* **5,** e15323.

Weingartner T., Eisner L., Eckert G. L. & Danielson S. (2009) Southeast Alaska: oceanographic habitats and linkages. *Journal of Biogeography* **36,** 387-400.

Whittingham M. J., Krebs J. R., Swetnam R. D., Vickery J. A., Wilson J. D. & Freckleton R. P. (2007) Should conservation strategies consider spatial generality? Farmland birds show regional not national patterns of habitat association. *Ecology Letters* **10,** 25-35.

Wilkins B. C. & Snyder N. P. (2011) Geomorphic comparison of two Atlantic coastal rivers: toward an understanding of physical controls on Atlantic salmon habitat. *River Research and Applications* **27,** 135-156.

Williams J. N., Seo C., Thorned J., Nelson J. K., Erwin S., O'brien J. M. & Schwartz M. W. (2009) Using species distribution models to predict new occurrences for rare plants. *Diversity and Distributions* **15,** 565-576.

Zar J. H. 2009. *Biostatistical Analysis,* Saddle River, NJ, Pearson Prentice-Hall.

FIGURES:

Figure 1: Map of the study area

Figure 2: Estuaries in southeast Alaska. Circles are sampling sites used in model development, squares are sample sites for independent data, and x's are the estuaries to which model results were predicted

Figure 3: Variable importance plots showing the six most influential variables for species from the top five fish species models. X-axis is the change in model performance (mean-squared error) when the variable is permuted and the y-axis is the variable code.

Figure 4: Partial dependence plots for the top five fish species models for the variable "month." X-axis is the numerical month and y-axis is the average trend of the variable in the model.

Figure 5: Predicted occurrence of Shiner perch (gray circles) and lingcod (black circles) showing overlaps where the species are expected to occur in the same estuary.

Figure 6: Predicted occurrence of starry flounder (black) and butter sole (green) with respect to average annual precipitation during the study period with warmer (red) colors representing higher precipitation.

Figure 7: Map showing currently designated EFH for Pacific cod (gray) in relation to predicted species occurrence in estuaries (black)

Table 1. Predictor variables used in predictive models. NA = data without a spatial scale

Type	No.	Unit	Time Scale	Spatial Scale of data source	Source
Estuary Area	1	m^2	NA	1:63,360	USFSTongass GIS - derived
Estuary Perimeter	1	m		1:63,360	USFSTongass GIS - derived
Intertidal Area	1	m^2		1:63,360	USFSTongass GIS
Intertidal Perimeter	1	m		1:63,360	USFSTongass GIS
Open water	1	m^2		1:63,360	USFSTongass GIS - derived
Watershed Area	1	m^2		1:63,360	USGS Hydrologic Unit Maps
Streams in watershed					
Estuarine streams	3	m		1:63,360	USFSTongass GIS
Total streams					
Percent estuarine					
Tidal range	2	Feet		NA	NOAA
Type of waterbody	2	Category		NA	Derived
Distance to waterbody	1	m		1:63,360	Derived
Width	1	m		1:63,360	measured
Length	1	m		1:63,360	measured
Depth of bar/sill	1	m		5 m	NMFSAKR Bathymetry -
Width of bar/sill	1	m		1:63,360	measured

Estuary slope					
Mean slope	3	Degrees		5 m	NMFSAKR Bathymetry -
Maximum slope					derived
Range of slope					
Depth					
Mean depth	3	M		5 m	NMFSAKR Bathymetry -
Maximum depth					derived
Range of depth					
Land cover patchiness					
Total area	5	Varies		30m	2001 National Land Cover
Number of patches					Dataset
Patch density					
Largest patch index					
Landscape shape index					
Annual precipitation	1	mm	1998-2005	2 km	PRISM Climate Model
Monthly precipitation	12	mm	1998-2005	2 km	PRISM Climate Model
Fluvial flow	1	Flow m^2	1998-2005	NA	Derived
Land cover					
Ice					
Developed					
Barren					
Deciduous	10	Percent		30 m	2001 National Land Cover
Evergreen					Dataset
Mixed vegetation					
Dwarf					
Scrub-shrub					
Woody wetlands					
Emergent herbaceous					
Slope of watershed	2	Degrees		300 m	USGS Digital Elevation Model
Habitat class	17	Percent		Coastal unit	Alaska ShoreZone dataset
Geology class	25	Percent		Coastal unit	Alaska ShoreZone dataset
Inter/subtidal vegetation (continuous/patchy)					
Red algae	8	Percent		Coastal unit	Alaska ShoreZone dataset
Aleria					
Soft brown kelp					
Dark brown kelp					
Eelgrass					

Table 2. Occurrence model results for fishes sorted by area under the curve (AUC). For life stage, capital letters indicate higher occurrence in the data than lower case letters. A = adults, J = juveniles, A/J = equal numbers adults and juveniles, A(j) = more adults than juveniles, J(a) = more juveniles than adults, and M

= species for which life stage could not be determined based on length of fish in samples. An * indicates that the model for that species was validated from OOB rather than independent datas

Fish Species	Life Stage	Present Correct (%)	Absence Correct (%)	AUC	No. of Sites (%)	Estuarine occupancy
Shiner perch *Cymatogaster aggregata*	A/J	100	87	0.94	39	seasonal
Lingcod *Ophiodon elongatus*	J	100	85	0.93	16	seasonal
Pink salmon *Oncorhynchus gorbuscha*	J	88	81	0.85	64	seasonal
Starry flounder *Platichthys stellatus*	A/J	84	68	0.77	34	year around
Dolly varden *Salvelinus malma*	A(j)	84	63	0.78	32	seasonal
Silverspot sculpin* *Blepsias cirrhosus*	J	83	71	0.78	42	unknown
Silverspot sculpin*Chum salmon *Oncorhynchus keta*	J	82	81	0.85	60	Seasonal
Kelp greenling *Hexagrammos decagrammus*	J	81	67	0.80	43	year around
Pacific cod *Gadus macrocephalus*	J	78	74	0.86	43	seasonal
Pacific herring *Clupea pallasi*	J	77	61	0.73	39	seasonal
Butter sole *Isopsetta isolepis*	J	75	74	0.88	8	seasonal
Rock sole *Lepidopsetta sp.*	J(a)	75	73	0.77	47	seasonal
Yellowfin sole *Limanda aspera*	J(a)	72	68	0.77	18	seasonal
Great sculpin**M. polyacanthocephalus*	J	71	65	0.72	8	unknown
Bay Pipefish *Syngnathus leptorhynchus*	A/(j)	71	62	0.70	53	year around
Sockeye salmon *Oncorhynchus nerka*	J	69	66	0.74	18	seasonal
Threespine stickleback *Gasterosteus aculeatus*	A	69	64	0.72	51	year around
Coho salmon *Oncorhynchuskisutch*	J	68	30	0.68	55	seasonal
Sturgeon poacher *Podothecus accipenserinus*	M	68	15	0.70	15	year around
Snake prickleback**Lumpenus sagitta*	M	61	59	0.67	54	year around
Tubesnout *Aulorhynchus flavidus*	M	62	62	0.62	47	year around
Crescent gunnel *Pholis laeta*	A(j)	60	50	0.63	78	year around
Pacific staghorn sculpin *Leptocottus armatus*	A/J	50	20	0.70	66	year around
Pacific sand lance *Ammodytes hexapterus*	J(a)	47	34	0.68	48	seasonal
Buffalo sculpin *Enophrys bison*	M	42	40	0.64	47	unknown

Table 1: Predictor variables used in predictive models. NA = data without a spatial scale.

Type	No.	Unit	Time Scale	Spatial Scale of data source	Source
Estuary Area	1	Sq. Meters	NA	1:63,360	USFSTongass GIS - derived
Estuary Perimeter	1	Meters		1:63,360	USFSTongass GIS - derived
Intertidal Area	1	Sq. Meters		1:63,360	USFSTongass GIS
Intertidal Perimeter	1	Meters		1:63,360	USFSTongass GIS
Open water	1	Sq. Meters		1:63,360	USFSTongass GIS - derived
Watershed Area	1	Sq. Meters		1:63,360	USGS Hydrologic Unit Maps
Streams in watershed					
Estuarine streams	3	Meters		1:63,360	USFSTongass GIS
Total streams					
Percent estuarine					
Tidal range	2	Feet		NA	NOAA
Type of waterbody	2	Category		NA	Derived
Distance to waterbody	1	Meters		1:63,360	Derived
Width	1	Meters		1:63,360	measured
Length	1	Meters		1:63,360	measured
Depth of bar/sill	1	Meters		5 m	NMFSAKR Bathymetry -
Width of bar/sill	1	Meters		1:63,360	measured
Estuary slope					
Mean slope	3	Degrees		5 m	NMFSAKR Bathymetry -
Maximum slope					derived
Range of slope					
Depth					
Mean depth	3	Meters		5 m	NMFSAKR Bathymetry -
Maximum depth					derived
Range of depth					
Land cover patchiness					
Total area	5	Varies		30m	2001 National Land Cover
Number of patches					Dataset
Patch density					
Largest patch index					
Landscape shape index					
Annual precipitation	1	Millimeters	1998-2005	2 km	PRISM Climate Model
Monthly precipitation	12	Millimeters	1998-2005	2 km	PRISM Climate Model
Fluvial flow	1	Flow/sq m	1998-2005	NA	Derived
Land cover					
Ice					
Developed					
Barren					
Deciduous	10	Percent		30 m	2001 National Land Cover
Evergreen					Dataset
Mixed vegetation					
Dwarf					
Scrub-shrub					

Woody wetlands				
Emergent herbaceous				
Slope of watershed	2	Degrees	300 m	USGS Digital Elevation Model
Habitat class	17	Percent	Coastal unit	Alaska ShoreZone dataset
Geology class	25	Percent	Coastal unit	Alaska ShoreZone dataset
Inter/subtidal vegetation (continuous/patchy)				
Red algae	8	Percent	Coastal unit	Alaska ShoreZone dataset
Alaria				
Soft brown kelp				
Dark brown kelp				
Eelgrass				

Table 2: Occurrence model results for fishes sorted by area under the curve (AUC). For life stage, capital letters indicate higher occurrence in the data than lower case letters. A = adults, J = juveniles, A/J = equal numbers adults and juveniles, A(j) = more adults than juveniles, J(a) = more juveniles than adults, and M = species for which life stage could not be determined based on length of fish in samples. An * indicates that the model for that species was validated from OOB rather than independent data.

Fish Species	Life Stage	Present Correct (%)	Absence Correct (%)	AUC	No. of Sites (%)	Estuarine occupancy
Shiner perch *Cymatogaster aggregata*	A/J	100	87	0.94	39	seasonal
Lingcod *Ophiodon elongatus*	J	100	85	0.93	16	seasonal
Pink salmon *Oncorhynchus gorbuscha*	J	88	81	0.85	64	seasonal
Starry flounder *Platichthys stellatus*	A/J	84	68	0.77	34	year around
Dolly varden *Salvelinus malma*	A(j)	84	63	0.78	32	seasonal
Silverspot sculpin* *Blepsias cirrhosus*	J	83	71	0.78	42	unknown
Chum salmon *Oncorhynchus keta*	J	82	81	0.85	60	seasonal
Kelp greenling *Hexagrammos decagrammus*	J	81	67	0.80	43	year around
Pacific cod *Gadus macrocephalus*	J	78	74	0.86	43	seasonal
Pacific herring *Clupea pallasi*	J	77	61	0.73	39	seasonal
Butter sole *Isopsetta isolepis*	J	75	74	0.88	8	seasonal
Rock sole *Lepidopsetta sp.*	J(a)	75	73	0.77	47	seasonal
Yellowfin sole *Limanda aspera*	J(a)	72	68	0.77	18	seasonal

FISH SPECIES	Life Stage	Present Correct (%)	Absence Correct (%)	AUC	No. of Sites (%)	Estuarine occupancy
Great sculpin*						
M. polyacanthocephalus	J	71	65	0.72	8	unknown
Bay Pipefish						
Syngnathus leptorhynchus	A/(j)	71	62	0.70	53	year around
Sockeye salmon						
Oncorhynchus nerka	J	69	66	0.74	18	seasonal
Threespine stickleback						
Gasterosteus aculeatus	A	69	64	0.72	51	year around
Coho salmon						
Oncorhynchuskisutch	J	68	30	0.68	55	seasonal
Sturgeon Poacher						
Podothecus accipenserinus	M	68	15	0.70	15	year around
Snake prickleback*						
Lumpenus sagitta	M	61	59	0.67	54	year around
Tubesnout						
Aulorhynchus flavidus	M	62	62	0.62	47	year around
Crescent gunnel						
Pholis laeta	A(j)	60	50	0.63	78	year around
Pacific staghorn sculpin						
Leptocottus armatus	A/J	50	20	0.70	66	year around
Pacific sand lance						
Ammodytes hexapterus	J(a)	47	34	0.68	48	seasonal
Buffalo sculpin						
Enophrys bison	M	42	40	0.64	47	unknown