Vol. 500: 159–174, 2014
doi: 10.3354/meps10659

MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser

Published March 17

Multivariate random forest models of estuarine-associated fish and invertebrate communities

Katharine Miller[1,*], Falk Huettmann[2], Brenda Norcross[3], Mitch Lorenz[1]

[1]Auke Bay Laboratories, National Marine Fisheries Service, Juneau, Alaska, USA
[2]Institute of Arctic Biology, University of Alaska, Fairbanks, Alaska, USA
[3]School of Fishery and Ocean Science, University of Alaska, Fairbanks, Alaska, USA

ABSTRACT: Models that evaluate species–habitat relationships at the community level have been gaining attention with increasing interest in ecosystem management. Developing models that can incorporate both a large number of predictor variables and a multivariate response (a vector of individual species occurrences or abundances) is challenging. One promising new approach is multivariate random forests (MRF), a method that combines multivariate regression trees with bootstrap resampling and predictor subsampling from traditional random forests. Random forest models have been shown to be highly accurate and powerful in their predictive ability in a wide variety of applications. They can effectively model nonlinear and interacting variables. Our research evaluated change in estuarine assemblage composition along habitat gradients in Southeast Alaska using landscape-scale habitat variables and MRF. For 541 estuaries, we identified 24 predictor variables describing the geomorphic and habitat environment on land and in the estuary. MRF models were constructed in R software for combined fish and invertebrate assemblages. Cluster analysis of model proximities revealed strong spatial variation in community composition in relation to differences in tidal range, precipitation, percent of eelgrass, and amount of intertidal habitat. This research presents a new science-based management template that can be used to inform and assess species management and protection strategies, as well as to guide future research on species distributions.

KEY WORDS: Estuaries · Multivariate models · Random forest

INTRODUCTION

The shift in focus from single-species management to ecosystem methods has increased interest in community or assemblage-based species distribution models. A common approach to multi-species distribution modeling is to model each species independently with respect to environmental variation and evaluate the overlaps in species distributions. An assumption of these single-species models is that species respond to environmental differences in an individualistic manner, which may not be the case where biotic interactions and dispersal pathways limit the species' realized niche. The use of metrics like species richness or diversity have the severe limitation that they do not capture the identity of the species in sampled areas and therefore can provide no information on how communities are structured. A species' distribution may be influenced by the distribution of other taxa within the study area (Bonthoux et al. 2013) and species may have similar environmental tolerances. By modeling each species separately, single-species models are limited in their ability to detect shared patterns of environmental response across taxa or identify interactions that influence species distributions (Ferrier & Guisan 2006).

Species–habitat relationships are affected by biotic and abiotic processes that occur on a variety of spa-

*Corresponding author: katharine.miller@noaa.gov

tial scales. In estuaries, species tend to have wide environmental tolerances that make them adaptable to many different environments and to high environmental variability, and their response to environmental factors can be nonlinear or discontinuous (Mueter & Norcross 1999, Gutiérrez-Estrada et al. 2008). Furthermore, the effects of most environmental factors do not occur in isolation from effects of other factors, which makes it difficult for researchers to attribute simple causality in explaining variation in assemblage composition. In Alaska, a number of studies have identified species–habitat relationships for individual species in estuaries or the nearshore (Abookire et al. 2001, Stoner et al. 2007), but the key patterns and processes that influence structure in these estuarine communities are still undefined.

One reason so much attention has focused on single-species models is that there are relatively few methods that can fit a large number of environmental predictor variables to a vector of dependent variables (species presence or abundance) resulting in a model that can predict community composition in unsampled areas. A promising new approach in this regard is multivariate random forests (MRF) (Segal & Xiao 2011), a method that combines multivariate regression trees (De'ath 2002) with bootstrap resampling and predictor subsampling from traditional random forests (Breiman 2001). Random forest models have been shown to be highly accurate in their predictive ability in a wide variety of applications (Magness et al. 2010). They can effectively model nonlinear and interacting predictor variables and can identify the predictor variables with the strongest influence on community composition patterns. By modeling the response of a community of species to environmental variables, models such as MRF incorporate information on species co-occurrence that can be used to evaluate the influence of species interactions on community composition. Results from these models can be used to extrapolate beyond the sampled assemblages to predict community composition in unsampled areas.

A challenge for community modeling is the large number of species that occur in only a few samples or occur in low numbers. In some cases, these species can account for up to half the species in the dataset. Species may be rare in samples because they are found in only a few of the habitats sampled, or they may be present at a broad number of sites and either occur in low numbers or are unable to be captured consistently with the sampling gear. Similarly, life-history or behavioral traits, such as schooling, may result in spatial clumping of species. In most single-species models, species with low occurrence in the data are often excluded because they do not exhibit good statistical properties. Modeling methods that use similarity matrices also tend to exclude rare species. Commonly used similarity metrics, such as the Bray-Curtis measure, are strongly affected by species abundance (Clarke et al. 2006) and are biased toward dominant species. The accuracy of these indices diminishes sharply as the number of rare species increases (Wisz et al. 2013, Schoch et al. 2014).

Excluding species with low occurrence is not without risk. These species often constitute the largest component of species richness in a community (Magurran & Henderson 2003), and recent evidence suggests that these species may play unique ecological roles (Mouillot et al. 2013). Multiresponse models, those that use a vector of species occurrence or abundance as the response variable, are able to incorporate data on both abundant and rare species and improve understanding of rare species distributions (Ferrier & Guisan 2006, Bonthoux et al. 2013). Multivariate regression trees have been used successfully to evaluate community structure for communities with rare species (Engle et al. 2007), suggesting that MRF models also will perform well for these species.

Our research evaluated change in estuarine assemblage composition along habitat gradients in Southeast Alaska. This region has approximately 22 500 km of shoreline divided among 1100 islands in an area known as the Alexander Archipelago. The area's large size and remoteness make it difficult to comprehensively sample for habitat characteristics and species. Therefore, we evaluated the influence of landscape-scale variables on changes in community composition in estuaries. Models predicting community composition over large spatial scales are common in terrestrial ecology (Oppel & Huettmann 2010), but their application to marine environments has been limited. Most marine landscape models have focused on specific environments or habitat types, such as coral reefs (Wedding & Friedlander 2008), mangroves (Jelbart et al. 2006), or seagrasses (Whitlow & Grabowski 2012). Few studies have investigated the relationship between landscape structure and composition of estuarine communities. In Australia, a comparison of fish assemblages among tropical estuaries found that estuary-level variables, such as tidal range, intertidal area, and distance to closest estuary, explained more variation in fish assemblages than site-specific physical variables such as salinity, substrate, and turbidity (Sheaves & Johnston 2009). These results support other research

indicating that environmental variables at intermediate scales may explain spatial patterns in species assemblages better than either site-specific or large-scale variables (Wiens 2011). Digital datasets of environmental variables are becoming increasingly available and are much less expensive and easier to obtain than site-specific environmental parameters, especially in remote and challenging environments like Southeast Alaska. If these variables can be used to develop models for detecting changes in fish and invertebrate community composition, they could become important tools in marine conservation and research.

MATERIALS AND METHODS

Study region

This research was conducted in the Alexander Archipelago, a collection of approximately 1000 mountainous islands in Southeast Alaska, USA. The study area (Fig. 1) extends from Lance Point in Lynn Canal (58° 44′ N, 135° 13′ W) to Cape Chacon in Dixon Entrance at the Canadian border (54° 41′ N, 132° 01′ W). 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. 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. Flows decrease in winter as a result of freezing, and increase again in the late spring and summer from melting of snow and ice. The flow of freshwater affects not only nearshore estuarine circulation, but is 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.

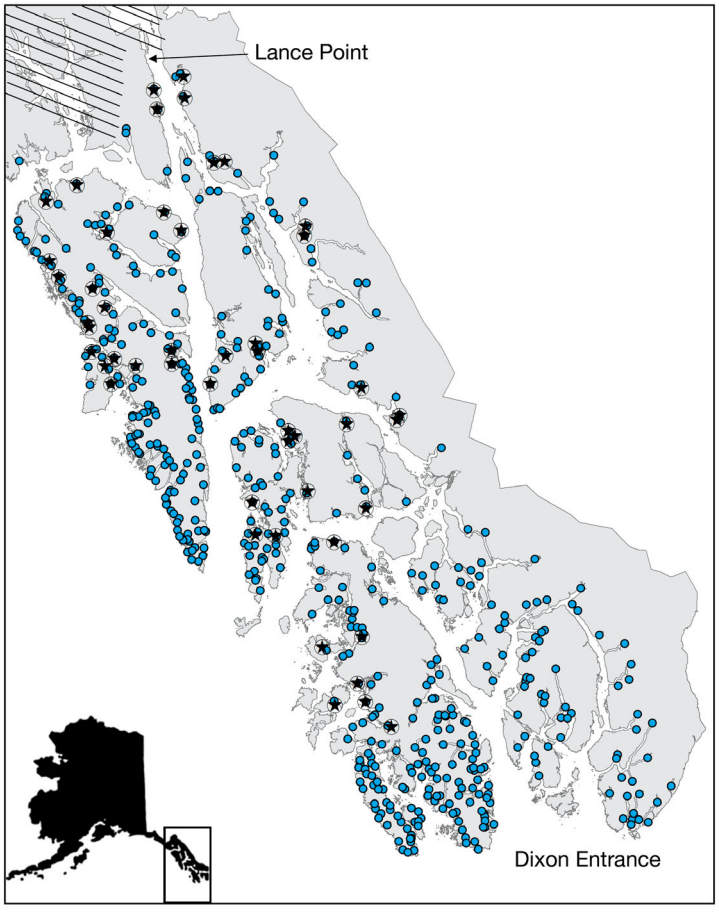

Fig. 1. Study area in Southeast Alaska showing sampling locations (★), and estuaries to which model results were predicted (●). Hash marks indicate exclusion of Glacier Bay and North Lynn Canal

The estuaries in the study area differ in their hydrologic and geomorphologic characteristics. In many Southeast Alaska estuaries, tidal energy is often much higher than energy from freshwater inflow. 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. 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 varies greatly across estuaries in the study area. 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.

Previous research identified approximately 12 000 estuaries in Southeast Alaska using the intersection of fresh and marine waters as the defining criteria (Schoch et al. 2014); however, this definition does not

take into account the degree to which an estuary is enclosed and somewhat isolated from other coastal waters. The degree of enclosure has important implications for estuarine circulation as well as the physical and chemical properties of the estuarine waters. For this research, we developed the following definition for an estuary: a coastal indentation with a restricted connection to saltwater and an aquatic environment affected by the physical and chemical characteristics of both fluvial drainage and marine systems. Using this definition, we delineated manually 541 estuarine polygons, including 49 polygons for estuaries for which we had biological data, between the high tide line and the 30 m depth contour (Fig. 1) in ArcGIS 10™.

Biological sampling

Forty-nine estuaries were sampled for fish and mobile invertebrates between April and September from 1998 and 2005 (Fig. 1). Sampling was conducted during daylight hours using an otter trawl (3 × 1 m, with 6 mm square mesh in the cod end) deployed with a bridle scope of approximately 5:1. The trawl was towed at a speed of approximately 3 knots along a depth contour between 5 and 10 m. The depth of individual tows varied within this range depending on bottom structure of the estuary. One tow in each direction was made along the same transect at high and low slack water, equaling 4 replicates at each station. The latitude and longitude of the beginning and ending points were recorded along with the average depth of the tow. Captured fish were identified to species and their total length was measured to the nearest millimeter in the field. Invertebrates were identified to the highest taxonomic separation in the field and counted.

Life stage, measured as the length at 50% maturity, was obtained from the Alaska Fishery Science Center's (AFSC) Life History database (http://access.afsc.noaa.gov/reem/LHWeb) for commercially harvested fish and some forage fish. This information was used to classify fish as adults or juveniles. For fish not in the AFSC database, other published sources (Froese & Pauley 2012 for the species or a close relative) were used to obtain length-at-maturity information. Several species occurred in the data entirely as either juveniles or adults and could be analyzed according to life stage. For species with a mix of juvenile and adult life stages, the amount of available data was generally insufficient to analyze each life stage separately, so the data were pooled and modeled together. Catch per unit effort (CPUE) was calculated as the number of species caught divided by tow length and then standardized to number of fish per 100 m. Most estuaries were sampled only a single time during the study period. For estuaries sampled more than once, CPUE was calculated as the total combined catch in a month divided by the average trawl length and standardized to number of fish per 100 m. CPUE for both fish and invertebrates was transformed to relative abundance using the Hellinger transformation by dividing the number of individuals of each species in each estuary by the total abundance of all species present in the estuary and taking the square root of the ratio. This transformation is widely used in multivariate analysis (Legendre & Gallagher 2001), including multivariate regression tree approaches (Wehrly et al. 2012) for species data containing many zeros.

Environmental data

For each of the 541 estuarine polygons in this research, we identified 24 predictor variables describing the geomorphic and habitat environment of the estuaries. Data were compiled from GIS layers from the Southeast Alaska GIS library (http://seakgis.alaska.edu), the National Oceanic and Atmospheric Administration (NOAA, http://tidesandcurrents.noaa.gov/stations.html?type=Bench+Mark+Data+Sheets), and the Alaska ShoreZone database (http://alaskafisheries.noaa.gov/shorezone/) (Table 1). All variables except the ones for intertidal vegetation were standardized to a mean of zero and standard deviation of one. The vegetation variables were recorded as percent of the vegetation type within each estuary polygon. The great diurnal tide range and mean tide range for each estuary were compiled from NOAA tide data. Estuaries without tide stations were attributed the tidal ranges from the nearest estuary with tidal data. Variables that were used to describe the structure of the estuary included the open water area, intertidal area, length of the intertidal perimeter, width at the estuary mouth, and bathymetric slope and depth. Open water area was the surface area of open water at low tide and was the difference between the estuary area and the intertidal area. Minimum depth, and maximum and average bathymetric slope, in each estuary polygon were calculated using the ArcGIS Spatial Analyst™ extension. Bathymetric slope is a proxy for the amount of habitat available at varying depths. The width of the estuary mouth was measured for each estuary along a line

between the landmasses on each side of the estuary entrance, or at the 30 m depth contour.

We included information on the intertidal environment, which provides important foraging habitat for subtidal species, such as juvenile Dungeness crab *Metacarcinus magister* (Holsman et al. 2003), as well as habitat for eelgrass (*Zostera marina*) communities. The intertidal area and perimeter length were obtained from the US Department of Agriculture (USDA) Tongass National Forest High and Low Tidelines dataset (http://seakgis.alaska.edu) for the intertidal areas within each estuary polygon. We calculated the intertidal ratio as the ratio of the intertidal perimeter to the intertidal area. This variable is an index of the shape and complexity of the intertidal environment. Area:perimeter ratios are widely used in landscape analyses to study species distributions and densities with respect to habitat size and edge effects (Martins et al. 2010), and more recently are being applied in studies of the marine environment (Wedding et al. 2011).

We also included variables describing the size and slope of the catchment surrounding the estuaries. Catchment size was derived from 12-digit hydrologic units depicting catchment boundaries. We measured catchment slope within a 5 km buffer around the estuary and used a digital elevation model and ArcGIS Spatial Statistics™ to calculate maximum and average slopes within each buffer.

Freshwater inflow into Southeast Alaska estuaries is difficult to calculate. Much of the study area is remote and undeveloped and there is a paucity of stream flow data even for large rivers. To capture the influence of freshwater on estuarine communities, we compiled minimum monthly precipitation over the study period from the Scenarios Network for Alaska and Arctic Planning (SNAP, http://www.snap.uaf.edu/data.php) climate model for Alaska into 5 seasonal variables: spring (February to April), summer (May to July), autumn (August to October), winter (November to January), and annual. We calculated fluvial flow per unit of estuary by multiplying the catchment with the average annual rainfall and a runoff coefficient (RV) and dividing by the open water area of the estuary to standardize flow per unit of estuary (Digby et al. 1998):

$$\text{Fluvial flow} = \text{Catchment} \times \text{Average annual precipitation} \times \text{RV} / \text{Open water area} \quad (1)$$

RV is based on the impervious fraction (I) of the drainage area (ADEC 2004):

$$\text{RV} = 0.05 + 0.9(I) \quad (2)$$

We calculated the variable I as the non-vegetated, non-ice portions of the catchment from the 2001 National Land Cover Dataset for Alaska (http://gisdata.usgs.gov/TDDS/DownloadFile.php?TYPE=nlcdak&FNAME=AK_NLCD_2001_land_cover_3-13-08.zip). At the scale of this analysis, the variable I was sufficiently small that the runoff coefficient was essentially a constant (0.059) across all catchments.

Physical characteristics have been used in a number of studies to classify estuaries and coastal areas (Engle et al. 2007, Schoch et al. 2014) and to predict species assemblages (Ellis et al. 2006, van der Wal et al. 2008, Schmiing et al. 2013). In Southeast Alaska, nearly the entire shoreline has been classi-

Table 1. Predictor variables used in multivariate random forest (MRF) model. No.: no. of parameters per variable; na: variables without a spatial scale

Variable	No. of parameters	Spatial scale of data source	Source
Intertidal area (m^2)	1	1:63360	USFS Tongass GIS, SEAK GIS Library
Intertidal perimeter (m)	1	1:63360	USFS Tongass GIS, SEAK GIS Library
Intertidal ratio (m)	1	1:63360	Derived
Open water (m^2)	1	1:63360	USFS Tongass GIS, SEAK GIS Library-derived
Catchment (m^2)	1	1:63360	USGS Hydrologic Unit Maps , SEAK GIS Library
Catchment slope (°)	1	300 m	USGS Digital Elevation Model, SEAK GIS Library
Tidal range (feet)	2	na	NOAA
Width (m)	1	1:63360	Measured
Estuary slope (°)	2	5 m	NMFS AKR Bathymetry-derived
Depth (m)	1	5 m	NMFS AKR Bathymetry-derived
Minimum seasonal precipitation (mm)[a]	4	2 km	SNAP Climate Model
Minimum annual precipitation (mm)[a]	1	2 km	SNAP Climate Model
Fluvial flow (mm)[a]	1	na	Derived
Continuous/patchy subtidal vegetation (%)	6	na	ShoreZone-derived

[a]Over a timescale of 1998–2005

Table 2. ShoreZone coastal class variables (Harney et al. 2008) used in the multivariate random forest (MRF) model. No.: the numeric ShoreZone designation for the variable; na: not applicable

Substrate	Sediment	Width	Slope	Coastal class	No.
Rock	na	Narrow (<30 m)	Steep (>20°)	Rock cliff	3
			Inclined (5–20°)	Rock ramp, narrow	4
Rock and sediment	Gravel	Wide (>30 m)	Inclined (5–20°)	Ramp with gravel beach, wide	6
			Flat (<5°)	Platform with gravel beach, wide	7
		Narrow (<30 m)	Steep (>20°)	Cliff with gravel beach	8
			Inclined (5–20°)	Ramp with gravel beach	9
			Flat (<5°)	Platform with gravel beach	10
	Sand and gravel	Wide (>30 m)	Inclined (5–20°)	Ramp with gravel and sand beach, wide	11
			Flat (<5°)	Platform with gravel and sand beach, wide	12
		Narrow (<30 m)	Steep (>20°)	Cliff with gravel and sand beach	13
			Inclined (5–20°)	Ramp with gravel and sand beach	14
			Flat (<5°)	Platform with gravel and sand beach	15
Sediment	Gravel	Wide (>30 m)	Flat (<5°)	Gravel flat, wide	21
		Narrow (<30 m)	Inclined (5–20°)	Gravel beach, narrow	22
	Sand and gravel	Wide (>30 m)	Flat (<5°)	Sand and gravel flat or fan	24
		Narrow (<30 m)	Inclined (5–20°)	Sand and gravel beach, narrow	25
			Flat (<5°)	Sand and gravel flat or fan	26
	Sand/mud	Wide (>30 m)	Flat (<5°)	Sand flat	28
			Flat (<5°)	Mudflat	29
	Organics	na	na	Estuaries	31
Anthropogenic	Human-made	na	na	Human-made, permeable	32
				Human-made, impermeable	33
Channel	Current	na	na	Channel	34
Glacier	Ice	na	na	Glacier	35

fied using the ShoreZone mapping and classification protocol (Harney et al. 2008). This method uses oblique, low-altitude video and still images to classify the shoreline according to natural breaks in geomorphic, sedimentary, and biological features. Shoreline segments are classified according to substrate type, sediment type, across-shore width, and slope. Assemblages of sessile coastal biota present within a shoreline segment are given a categorical descriptor of either continuous (>50% cover within the unit) or patchy (<50% coverage). Shorelines are further defined by their habitat class, which is an index that combines geology, wave exposure, and biota into a single variable. For this analysis, we used shoreline classes and habitat classes that were represented in 5% or more of the sampled estuaries, for a total of 23 shorelines classes and 13 habitat classes (Table 2). Variables were calculated as the percentage of each class with respect to the total perimeter of the estuary polygon. We also included percent of continuous or patchy canopy kelp (*Alaria* spp.), eelgrass (*Zostera marina*), and soft brown kelps (*Saccharina latissima*) within the intertidal and shallow subtidal zones of the estuaries.

Data analysis

MRF models were constructed in R 2.13.2 (Segal & Xiao 2011) for combined fish and invertebrate assemblages to investigate spatial variation in species' relative abundance in relation to the environmental predictor variables. MRFs are a modification of multivariate regression trees (MRTs), which have been widely used to model change in marine communities (Claudet et al. 2006, Ruppert et al. 2010, Kortsch et al. 2012). The MRF model improves the stability and performance of MRTs by building an ensemble of several hundred trees using boostrapped subsamples of the original data and aggregating the results (Segal & Xiao 2011). In our MRF model, the response is the relative abundance of each species at each sampled estuary. The prediction error for each tree is calculated using the data omitted from the bootstrap sample for that tree, and the prediction error for the forest is the average prediction error of the individual trees. Variable importance is calculated in the same manner as for traditional random forests, by randomly permuting the values of the variables, running them through the model, and evaluating the change in the mean squared error (MSE). Variables having

the greatest effect on MSE have more influence on model accuracy (Breiman 2001).

We applied the MRF algorithm to construct 100, 200, 300, 500, and 1000 trees. After 300 trees, performance was not substantially enhanced by the adding trees, so we selected 300 as the maximum number of trees for our model. For the number of variables to use for constructing each tree, we used the default value of 1/3 of the predictor variables.

In the proximity matrix, random forest models also provide a measure of site similarity based on both the physical and biological variables at an estuary. This matrix is constructed by comparing the location of estuaries in the terminal nodes of each tree in the forest and giving higher proximity values to estuaries in the same node. Proximity values for each tree are summed and normalized by dividing by the number of trees in the forest. Subtracting 1 from the proximity values of the matrix converts the data to squared Euclidean distances (Segal & Xiao 2011). We used multidimensional scaling and partition around the medoid (PAM) clustering (Kaufman & Rousseeuw 1990) on the distance matrix to classify the 49 sampled estuaries. The optimal number of clusters (k) was determined by selecting the k with the maximum definition of silhouette width, which is a measure of the difference between intra-cluster similarity and similarity with the next closest cluster. Silhouette widths close to 1 indicate perfectly assigned clusters (Rousseeuw 1987). To predict the class membership of the unsampled estuaries, we used the clusters from the MRF PAM clustering as the response variable in a traditional random forest model. The resulting model provided the splitting rules for assigning unsampled estuaries to each cluster and provided a misclassification rate that we used to evaluate model performance. All modeling and analysis was done in R (R Development Core Team 2008).

RESULTS

The modeling dataset contained 22 species of fish from 12 families, and 14 species of invertebrates from 11 families. Snake prickleback *Lumpenus sagitta*, yellowfin sole *Limanda aspera*, and starry flounder *Platichthys stellatus* were the most numerous fish in the data, comprising 15, 14, and 12% of the total catch over all samples and years, respectively. Four fish species were captured at 50% or more of the estuaries: Pacific staghorn sculpin *Leptocottus armatus*, crescent gunnel *Pholis laeta*, starry flounder, and rock sole *Lepidopsetta* sp. Shell shrimp *Crangon alaskensis* and spot shrimp *Pandalus platyceros* were the most abundant invertebrate species, comprising 24 and 19% of the total catch, respectively, but 90% of the spot shrimp catch occurred at a single estuary. Three species and one family of invertebrates were captured at over 50% of the estuaries: sunflower sea star *Pycnopodia helianthoides*, shell shrimp, gammarid amphipods (Gammaridiae), and helmet crab *Telmessus cheiragonus* (Table 3).

Several fish species occurred more frequently or entirely as juveniles in the data. Species occurring only as juveniles included Pacific herring *Clupea pallasi*, Pacific cod *Gadus macrocephalus*, lingcod *Ophiodon elongatus*, kelp greenling *Hexagrammos decagrammus*, butter sole *Isopsetta isolepis*, and great sculpin *Myoxocephalus polyacanthocephalus*. Species whose abundance was predominantly composed of juveniles were yellowfin sole, rock sole, and Pacific sand lance *Ammodytes hexapterus*. Species with mixes of juveniles and adults were starry flounder, Pacific staghorn sculpin, and shiner perch *Cymatogaster aggregata*. Species for which life stage could not be determined from the literature were the snake prickleback, tube-snout *Aulorhynchus flavidus*, and sturgeon poacher *Podothecus accipenserinus*.

PAM clustering of estuaries using the MRF proximities identified 3 clusters (Fig. 2a,b), with silhouette widths of 0.31, 0.33, and 0.42. The absence of negative silhouette widths indicates that all estuaries were correctly assigned to clusters. The first 2 components explained approximately 55% of the point variability in the data. Using these clusters as dependent variables, the traditional random forest model had a classification error rate of 4%, with one estuary in Cluster 1 (19 estuaries), zero estuaries in Cluster 2 (20 estuaries), and one estuary in Cluster 3 (10 estuaries) being misclassified. Tidal range, minimum precipitation, and percent of continuous eelgrass were the most influential variables in the model (Fig. 3). Variables describing the amount of intertidal habitat, open water area, and characteristics of the catchment surrounding the estuary also were influential (Fig. 4). The data for all variables except vegetation had been standardized to a mean of zero. This means a zero value indicated an average value for the variable, with values above average plotted to the right of center and values below average plotted to the left. Vegetation variables were measured in percent of estuary perimeter and all had positive values.

Estuaries in Cluster 1 had intermediate tidal range values, and small open water and intertidal areas,

Table 3. Catch per unit effort (CPUE; number of individuals per 100 m) of fish and invertebrates from trawl samples

Fish

Family Species	CPUE
Clupeidae (herrings)	
Clupea pallasi (Pacific herring)	87
Gadidae (cods)	
Gadus macrocephalus (Pacific cod)	21
Aulorhynchidae (tubesnouts)	
Aulorhynchus flavidus (tubesnout)	146
Gasterosteidae (sticklebacks)	
Gasterosteus aculeatus (three-spined stickleback)	59
Sygnathidae (pipefishes)	
Syngnathus leptorhynchus (Bay pipefish)	192
Hexagrammidae (greenlings)	
Ophiodon elongates (lingcod)	9
Hexagrammos stelleri (whitespotted greenling)	7
Hexagrammos decagrammus (kelp greenling)	104
Hexagrammos lagocephalus (rock greenling)	47
Cottidae (sculpins)	
Icelinus borealis (northern sculpin)	7
Leptocottus armatus (Pacific staghorn sculpin)	122
Enophrys bison (buffalo sculpin)	87
Myoxocephalus polyacanthocephalus (great sculpin)	24
Agonidae (poachers)	
Podothecus accipenserinus (surgeon poacher)	31
Stichaeidae (pricklebacks)	
Lumpenus sagitta (snake prickleback)	397
Pholidae (gunnels)	
Pholis laeta (crescent gunnel)	241
Ammodytidae (sand lances)	
Ammodytes hexapterus (Pacific sand lance)	99
Pleuronectidae (righteye flounders)	
Hippoglossus stenolepis (flathead sole)	13
Platichthys stellatus (starry flounder)	302
Lepidopsetta sp. (rock sole)	218
Isopsetta isolepis (butter sole)	11
Limanda aspera (yellowfin sole)	376

Invertebrates

Family Species	CPUE
Pychnopodiidae	
Pycnopodia helianthoides (sunflower sea star)	251
Tethyididae	
Melibe leonina (hooded nudibranch)	396
Crangonidae	
Crangon alaskensis (shell shrimp)	1996
Gammarids	68
Hippolytidae (broken back shrimp)	
Heptacarpus stylus (stiletto shrimp)	51
Hippolyte sp.	821
Cancridae	
Metacarcinus magister (Dungeness crab)	331
Cancer productus (red rock crab)	87
Cancer sp.	152
Cheiragonidae	
Telmessus cheiragonus (helmet crab)	384
Majidae (spider crabs)	112
Oregoniidae	
Oregonia gracilis (graceful decorator crab)	179
Epialtinae	
Pugettia producta (northern kelp crab)	339
Pandalidae	
Pandalus danae (dock shrimp)	87
Pandalus platyceros (spot shrimp)	1536
Pandalus sp.	85

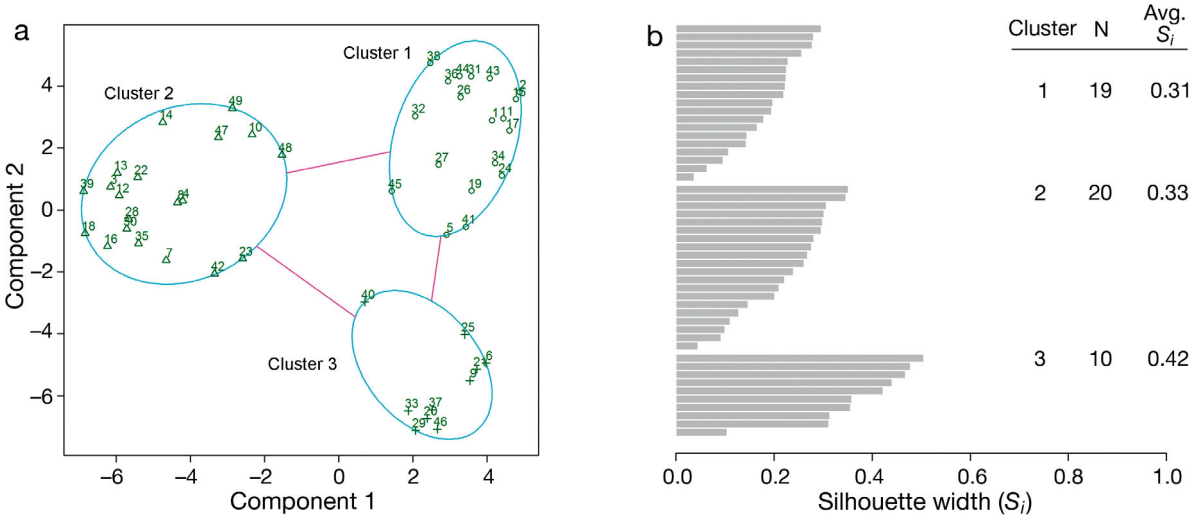

Fig. 2. (a) Multidimensional scaling plot and (b) silhouette widths of multivariate random forest (MRF) proximity values. Ellipses delineate partition around the medoid (PAM) cluster membership, and numbers are sample estuaries. The 2 components explain 54.85% of the variation in the data. Lines indicate the distance between clusters. Silhouette plot shows strength of group membership

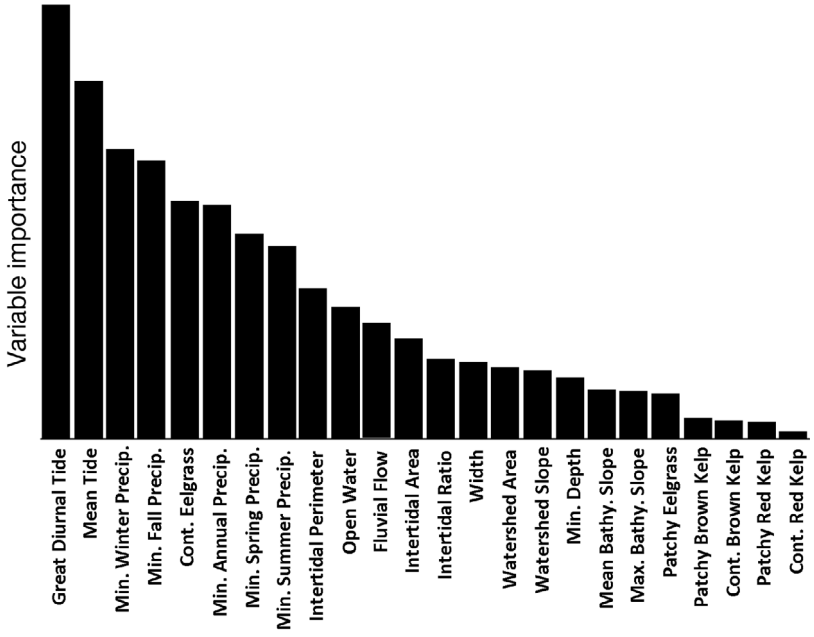

Fig. 3. Relative importance of the predictor variables scaled by the decrease in the mean squared error of the model when the variable is permuted. Higher bars equate to higher variable importance. Cont.: continuous, Bathy.: bathymetry-derived

but high estuary slopes and depths and high minimum precipitation throughout the year (Fig. 4). These estuaries had relatively small catchments and open water areas, but high fluvial flow and higher than average precipitation. Cluster 1 estuaries had the most even mix of fish species (Fig. 5), with all species represented except the white spotted greenling *Hexagrammos stelleri*. Dock shrimp *Pandalus danae* and graceful decorator crab *Oregonia gracilis* also were absent from this cluster (Fig. 6). Lingcod, Pacific cod, Pacific sand lance, rock greenling *H. lagocephalus*, northern sculpin *Icelinus borealis*, and buffalo sculpin *Enophrys bison* were found in this cluster but occurred at low relative abundances or were absent from other clusters. Average continuous and patchy kelp coverage in these estuaries was 60%, compared with 40% for estuaries in Clusters 2 and 3.

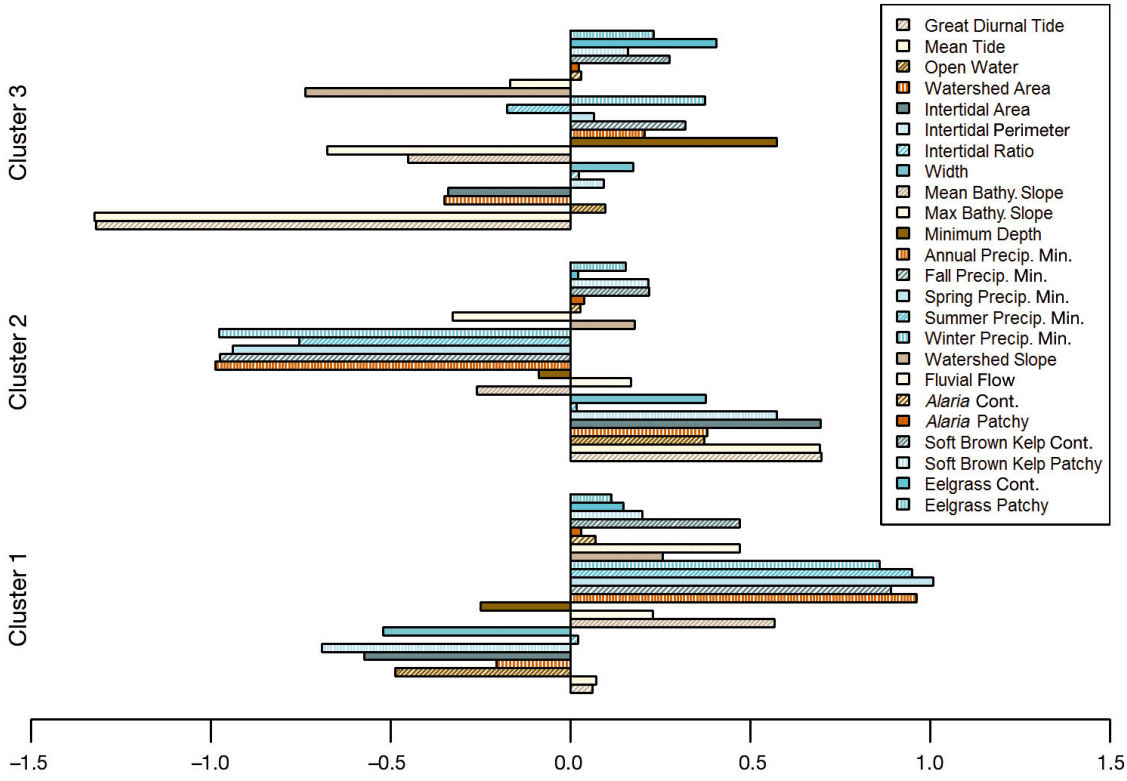

Fig. 4. Average value of predictor variables for sampled estuaries by cluster. Data for all variables except vegetation were standardized to a mean of zero, so a zero value indicates an average value for the variable, with values above average plotted to the right of center and values below average plotted to the left. Vegetation variables are measured in percent of estuary perimeter and all have positive values. Cont.: continuous, Bathy.: bathymetry-derived

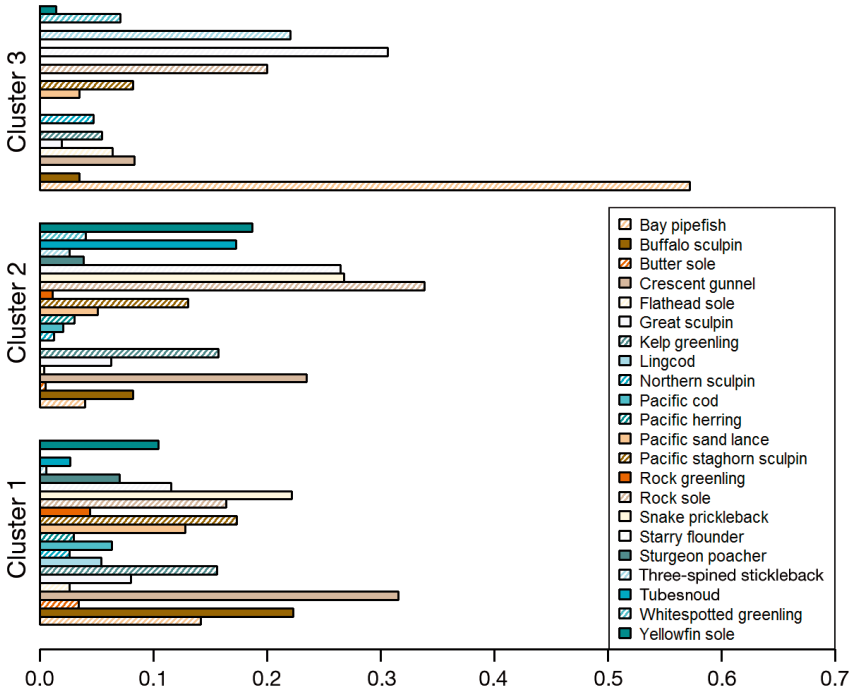

Fig. 5. Average relative abundance of fish species for sampled estuaries by cluster for sampled estuaries

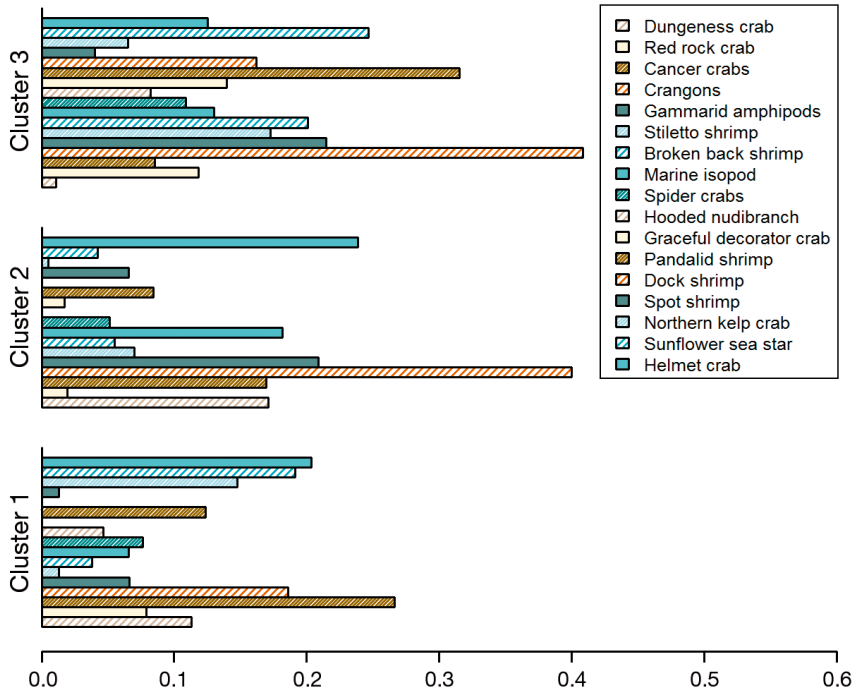

Fig. 6. Average relative abundance of invertebrates by cluster for sampled estuaries

Estuaries in Cluster 2 were characterized by low tidal ranges, large open water and intertidal values, and low minimum precipitation and fluvial flow (Fig. 4). Precipitation patterns in these estuaries were the reverse of those of estuaries in Cluster 1. Estuaries in this cluster had the highest relative abundance of flatfishes (Fig. 5) and Dungeness crab *Metacarcinus magister* (Fig. 6). Cluster 2 estuaries also had high relative abundances of tube-snouts. Dock shrimp and graceful decorator crabs were absent from this cluster, and abundances of other shrimp were generally lower than in the other 2 clusters. Lingcod were the only fish absent from this cluster.

Average precipitation in Cluster 3 estuaries was lower than for estuaries in Cluster 1 but was higher than average for all seasons except summer (Fig. 4). Cluster 3 estuaries had the lowest average tidal ranges of all estuaries in the data and the steepest slopes. Intertidal area in Cluster 3 was lower than average for the dataset, but both intertidal perimeter and intertidal ratio were higher than average. These estuaries had the lowest number of fish taxa (Fig. 5). Butter sole, lingcod, Pacific cod, Pacific herring, rock greenling, snake prickleback, sturgeon poacher, and tube-snouts were all absent from this cluster. In contrast, estuaries in Cluster 3 had the higher occurrence of invertebrate species, particularly shrimp (Fig. 6). In this cluster, estuaries averaged between 39 and 93% patchy or continuous eelgrass coverage. This cluster was characterized by high relative abundances of bay pipefish *Syngnathus leptorhynchus* and three-spined sticklebacks *Gasterosteus aculeatus*. Average relative abundance of Dungeness crab was lower in this cluster than in the other two. Cluster 3 estuaries had high abundances of rock sole and starry flounder, similar to estuaries in Cluster 2, but low relative abundances of yellowfin sole.

The majority of the ShoreZone coastal class and habitat class variables had very low or negative variable importance, indicating low direct predictive value for the MRF model. Environmental variables are naturally intercorrelated. Variable predictor scores

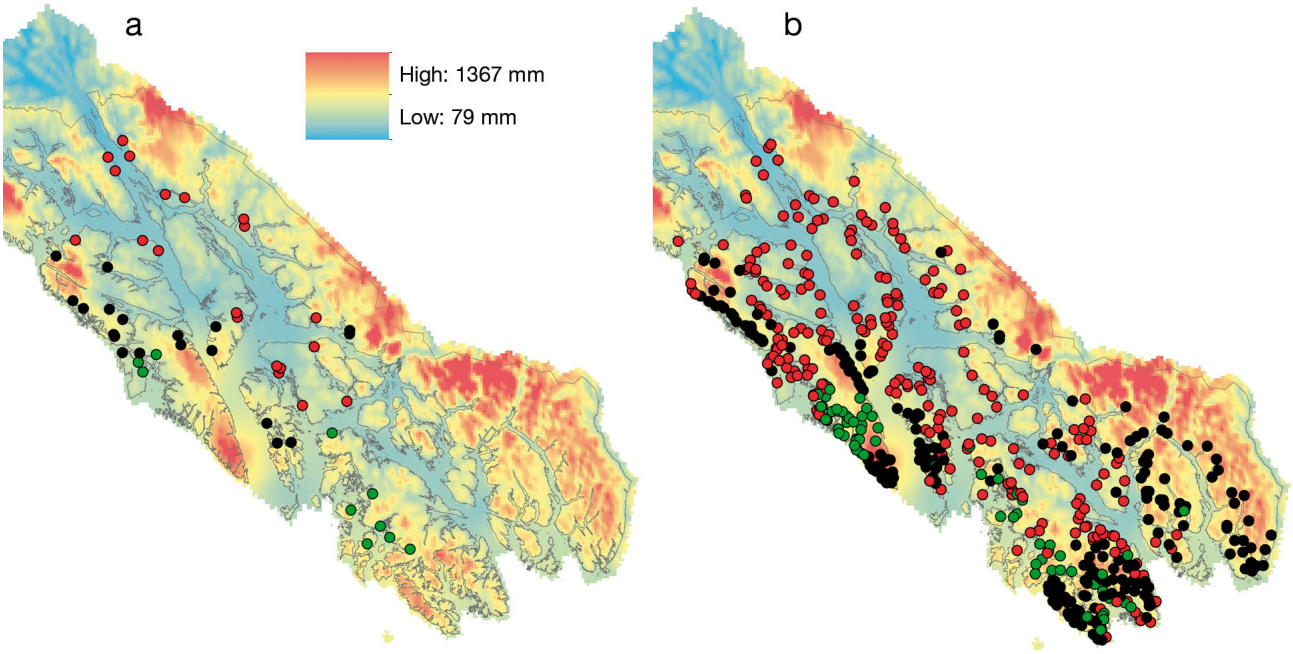

Fig. 7. (a) Spatial distribution of 49 sample estuaries by cluster from multivariate random forest (MRF) model and partition around the medoid (PAM) analysis, and (b) predicted cluster membership of 492 delineated estuaries. Estuaries colored by cluster membership and overlaid on winter minimum precipitation. Cluster 1 = ●, Cluster 2 = ●, and Cluster 3 = ●

in random forest models take into account interactions among variables (Lunetta et al. 2004), and variables that do not have high individual prediction scores may influence and increase or decrease the prediction scores of other variables. One of the strengths of random forest methods is that they can detect variable dependencies even when these relationships are nonlinear or discontinuous (Cutler et al. 2007). However, step-wise removal of variables with low or negative importance can result in deleting interacting variables that are otherwise important to the scores of variables that are retained, resulting in unstable models. A number of methods have been proposed to select the most relevant variables in random forest models (Sandri & Zuccolotto 2006, Genuer et al. 2010), but there is no agreement on how or whether variable removal should be done. Rather than implementing a step-wise variable removal method that might introduce bias into the model from interacting variables, we chose to remove all ShoreZone coastal and habitat class variables from the analysis and run the models on the remaining 24 environmental variables.

The spatial distribution of clusters reflected the influence of the highest performing variables: tidal range and winter precipitation (Fig. 7a). Estuaries in Cluster 1 were located on both the outer coast and along the deeper channels where precipitation from the interaction of ocean storms and coastal mountains is high. Cluster 2 estuaries were located among the inland waters where tidal range is lowest and the rain shadow of the coastal mountains results in lower precipitation than the other clusters. Estuaries in Cluster 3 were located on the outer coast adjacent to the open ocean. These estuaries tended to have the lowest tidal ranges due to their proximity to the shelf. These general patterns of cluster distribution held when cluster membership was predicted for the 492 other estuaries in the study area, but more overlap occurred as a result of the other variables in the model (Fig. 7b).

DISCUSSION

The MRF models identified spatial patterns in Southeast Alaska fish and invertebrate communities in relation to precipitation, tidal range, percent of eelgrass, and amount of intertidal habitat. Classification error rate for the model was very low, suggesting that this approach is an effective method for predicting community composition in the study area. One of the strongest patterns in the environmental data is the difference in precipitation and tidal exchange between the clusters, but it can be difficult to tease apart the influence of individual variables in models where the importance of any variable may be influenced by other interacting variables in the data

(Knudby et al. 2010). Both precipitation and tidal exchange vary substantially across the study area at small spatial scales. Precipitation patterns are strongly influenced by catchment elevation, with ocean storms releasing moisture as air is adiabatically cooled by the high mountainous landscape (Weingartner et al. 2009). This process results in areas of high precipitation near the coast, and in rain shadows along the interior passes. Tidal height increases as water moves inland away from the continental shelf and into the interior of the archipelago. Tidal currents are strongly influenced by channel width, bathymetric structure, and depth, resulting in substantial variations in tidal current strengths at the scale of several kilometers. The interaction of these variables results in a spatial distribution of estuary clusters that generally aligns with patterns of precipitation (Fig. 7b), but with deviations based on estuary-specific differences in bathymetry, estuary size, and amount of intertidal habitat. The complex interactions between ecosystem components would likely not have been identified using a linear model.

Precipitation and fluvial flow variables were included in the analysis to capture differences in salinity and buoyancy-driven circulation between estuaries, but the relationship between community composition and salinity is complex. Several studies on juvenile groundfish in Alaska found only small or insignificant correlations between salinity and abundance (Norcross et al. 1997, Abookire et al. 2001). In Kachemak Bay, Alaska, Pacific herring and sand lance are substantially more abundant in the interior of the estuary in less-saline surface waters (Abookire et al. 2000). However, in the Skagit River estuary in Puget Sound, Washington State, there is no correlation between Pacific herring annual abundance and river discharge (Sandri & Zuccolotto 2006).

Precipitation may also be a proxy for other oceanographic processes. Higher precipitation can freshen the surface water layer and result in stratification, which stabilizes the water column and can lead to enhanced primary productivity. In the presence of sufficient nutrients, stratification strong enough to prevent mixing to depths greater than the euphotic zone will result in a plankton bloom. Higher freshwater discharge is also associated with the development of tidal fronts: 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. Constrictions, such as a narrowing of the estuary mouth, act as hydraulic controls that can enhance formation of these fronts (Largier 1993). Nutrients are drawn into the stratified surface layer of the front by diapycnal mixing at the frontal boundary, and over a period of time can enhance phytoplankton production (Largier 1993, Johnson & Costello 2002). At the same time, convergent flows along the frontal boundary advect and concentrate plankton (Franks & Chen 1996), which attract grazers and higher-trophic-level predators (Kingsford & Suthers 1994). These fronts may also act as barriers to larval transport by deflecting flow along the frontal boundary and limiting transport across the boundary (Eggleston et al. 1998). This may help retain and distribute planktonic larvae within the estuary.

Freshwater discharge into Cluster 1 was higher year-round than for the other 2 clusters. These estuaries had higher than average fluvial flow, similar to estuaries in Cluster 3, but Cluster 1 estuaries also had narrow average mouth widths, steep slopes, and small open water areas: all factors that enhance stratification and water column stability. The conditions in Cluster 1 may favor enhanced productivity that could explain the more even mix of fish species and the higher relative abundance of species such as sand lance and Pacific herring, which tend to favor higher-productivity water (Arimitsu et al. 2004).

The model results confirmed several expected species–habitat relationships. For example, northern kelp crab *Pugettia producta* is an herbivore which feeds on kelp and utilizes kelp pigments to maintain its shell color similar to its surrounding habitat (Lunetta et al. 2004). These crabs were abundant in Cluster 1, which had the highest percentage of average and continuous patchy kelp coverage. Other species whose high relative abundance also may be explained by the presence of vegetation were crescent gunnels and northern sculpin, which are regularly captured in kelp and eelgrass habitats in Southeast Alaska (Johnson et al. 2003), and juvenile lingcod and juvenile Pacific cod, which prefer structured habitats that include kelp and eelgrass beds (Petrie & Ryer 2006, Laurel et al. 2007). Cluster 3 estuaries had the highest average abundance of eelgrass, and these estuaries were characterized by high relative abundances of bay pipefish *Syngnathus leptorhynchus* and three-spined sticklebacks *Gasterosteus aculeatus*, both eelgrass-associated species (Johnson et al. 2003).

Cluster results explained a little over half the point variability in the data, indicating that variables not included in the model are influencing species distributions and community composition. One such variable might be substrate type. Several studies have demonstrated strong associations between fish (Abookire et al. 2001) and invertebrates (Hovel &

Wahle 2010) and substrate type. Although we did not find inclusion of subtidal geology variables from the ShoreZone dataset to be informative in our models, this may be a result of the way the variables were derived. These variables were calculated as the percent of each substrate occurring within the estuary, a method that is not sensitive to the patch size of the habitat and sediment types. Therefore, an estuary with 30% continuous coverage of sand, for example, is equivalent to an estuary with 3 small and well-spaced patches of sand of 10% each. While inclusion of the absolute area of each substrate type within the estuary might improve the performance of these variables in the model, it would be preferable to have some measure of both the size and separation of habitats within each estuary. Size and shape of habitat patches are important factors affecting species abundance, diversity, and habitat use (Morin 2011).

The spatial arrangement of habitat in estuaries may also be important in influencing community composition. ShoreZone variables were extracted for the entire estuary polygon and may not correspond directly to the habitat in the area sampled. Trawl sampling is constrained to occur in areas with minimal rocks or hard structures, and these may represent only a portion of the substrates in the estuary. Including substrate variables directly under sampling transects along with information on adjacent habitats may enhance the performance of ShoreZone substrate variables in the model. Our future research will attempt to include this type of finer-scale substrate information in the models.

Biotic factors, such as competition, dispersal limitation, and predation are known to constrain species distributions (Boulangeat et al. 2012, Wisz et al. 2013), and some of the variance that is not captured in our model could be the result of biotic interactions. Unfortunately, incorporating relevant biotic factors into species distributions models is complicated by lack of data on species interactions at the scale of the analysis, and confounding effects of abiotic and biotic variables on individual species (Elith & Leathwick 2009). By using the species assemblage as the response variable, the multivariate random forest model used in this research implicitly incorporates patterns of species co-occurrence that could be reflections of species interactions, but the model cannot identify specific relationships with biotic factors or whether such factors are having a direct or indirect effect on the model. Methods for incorporating biotic interactions into species distribution models include adding competing or predator/prey species as explanatory variables (Kissling et al. 2012), including estimates of habitat productivity (Wisz et al. 2013), and including data on dispersal (Boulangeat et al. 2012). Incorporating these data into multispecies models is especially challenging due to the complexity of species interactions within a community. For marine species, incorporation of dispersal data is also hampered by lack of oceanographic data on a scale relevant to the analysis. Our ongoing work in this area involves evaluating the strength of biotic interactions through the development of niche models that use behavioral, functional, or phylogenetic traits. We will be evaluating ways to incorporate insights on the importance of biotic factors to estuarine communities gained from these null models into our community analyses.

Several Alaskan estuarine species have strong seasonal patterns. An example is the tube-snout, which is found in a variety of sandy and rocky habitats with adjacent eelgrass or kelp. Tube-snouts occur in the nearshore to depths of 30 m. In nearshore seine-net fish sampling in Southeast Alaska from 1998 to 2000, tube-snouts were captured in low numbers, but consistently, throughout Southeast Alaska (Johnson et al. 2003). In the present analysis, tube-snouts were entirely absent from Cluster 3, although species that frequently co-occur with them, such as the bay pipefish, were abundant in that cluster, and Cluster 3 has the highest percentage of eelgrass. This disparity may be explained by the season in which the sampling occurred in our analysis. In this study, field data were collected over a period of 7 yr, with most estuaries sampled only once during that period. As a result, not all estuaries were sampled in the same month or season. Seasonality and interannual variation in the data may have introduced uncertainty into some aspects of the analysis. Although tube-snouts are year-round residents of estuaries and nearshore areas, sampling in Prince William Sound, Alaska, in 2006 and 2007 (Johnson et al. 2010) captured substantially more tube-snouts in September (331) than in April (21) or July (between 67 and 95). Estuaries in Cluster 3 were sampled primarily in April, with some samples occurring in May and June. In contrast, estuaries in Cluster 2, which has the highest relative abundance of tube-snouts, were sampled relatively uniformly between April and September. Unfortunately, the September Prince William Sound surveys were conducted in a single year, giving no information on interannual differences, and no other information on seasonality in tube-snouts is available. Similarly, seasonal movement of species from nearshore areas to offshore areas in Alaska has been documented in other studies (Abookire & Norcross 1998,

Stone & O'Clair 2001), but, as with the present research, sampling was not consistent between months and years. Interannual differences in temperature and oceanographic conditions affect timing of migratory behavior and make it difficult to compare species composition at estuaries sampled in different months and years. Most sampling in Southeast Alaska occurs from March to August and sometimes September, but there has been no research comparing species abundances for all of these months across more than 1 yr at the same location. Until such research can be conducted, it is unclear whether changes in the seasonal abundance of species have an impact on the results of this analysis, and predicted community composition should be assumed to apply only to the months for which sampling occurred.

CONCLUSIONS

Understanding the mechanisms that influence community composition is fundamental to ecology and a precursor for ecosystem management. Multispecies management strategies require knowledge not only of abiotic factors affecting species distributions (fundamental niche), but also of the relationship between species within a community. Spatially explicit models of estuarine communities can be useful in developing management and conservation strategies for fish and invertebrate species, as well as providing insight into estuarine ecosystem processes. To this end, methods that can quantitatively model multiple species will provide additional insight into species' functional roles and interactions. In the past, many multivariate analyses of species with respect to habitat have been constrained to model a univariate response: either individual species or an index such as species diversity or richness. The advance of machine learning algorithms, such as MRF, has provided powerful methods for modeling species and environment data together and producing information on spatial patterns of communities. This is the first implementation of an MRF model to marine fish and invertebrate communities, and the first research to evaluate the relationship between landscape structure and estuarine community composition in Southeast Alaska. At the regional scale, estuaries clustered from the model show strong association with spatial patterns of precipitation and tidal height. At a more local scale, the amount of intertidal habitat and availability of kelp and eelgrass habitats influenced the relative of abundance of individual species within the communities. Evaluating both large- and fine-scale patterns in community composition can inform species management and protection strategies, as well as guide future research on species co-occurrence. As with other remote and relatively inaccessible areas, the majority of the Alaska coast has not been systematically sampled. Using the model results to predict community composition in unsampled areas can enhance understanding of nearshore marine processes and identify areas to be targeted for additional research or protection.

LITERATURE CITED

Abookire AA, Norcross BL (1998) Depth and substrate as determinants of distribution of juvenile flathead sole (*Hippoglossoides elassodon*) and rock sole (*Pleuronectes bilineatus*), in Kachemak Bay, Alaska. J Sea Res 39: 113–123

Abookire AA, Piatt JF, Robards MD (2000) Nearshore fish distributions in an Alaskan estuary in relation to stratification, temperature and salinity. Estuar Coast Shelf Sci 51:45–59

Abookire AA, Piatt JF, Norcross BL (2001) Juvenile groundfish habitat in Kachemak Bay, Alaska during late summer. Fish Bull 8:45–56

ADEC (Alaska Department of Environmental Conservation) (2004) Total maximum daily load (TMDL) for fecal coliform in the waters of Little Survival Creek in Anchorage, Alaska. ADEC, Anchorage, AK. http://dec.alaska.gov/water/tmdl/approvedtmdls.htm

Arimitsu ML, Piatt JF, Romano MD, Douglas DD (2004) Distribution of forage fishes in relation to oceanography of Glacier Bay National Park. In: Piatt JF, Gende SM (eds) US Geol Surv Sci Invest Rep 1007-5047. Proc Fourth Glacier Bay Sci Symp, October 26–28, 2004. US Geological Survey, Juneau, AK, p 102–106

Bonthoux S, Baselga A, Balent G (2013) Assessing community-level and single-species models predictions of species distributions and assemblage composition after 25 years of land cover change. PLoS ONE 8:e54179

Boulangeat I, Gravel D, Thuiller W (2012) Accounting for dispersal and biotic interactions to disentangle drivers of species distributions and their abundances. Ecol Lett 15: 584–593

Breiman L (2001) Random forests. Mach Learn 45:5–32

Clarke KR, Chapman MG, Somerfield PJ, Needham HR (2006) Dispersion-based weighting of species counts in assemblage analyses. Mar Ecol Prog Ser 320:11–27

Claudet J, Pelletier D, Jouvenel JY, Bachet F, Galzin R (2006) Assessing the effects of marine protected area (MPA) on a reef fish assemblage in a northwestern Mediterranean marine reserve: identifying community-based indicators. Biol Conserv 130:349–369

Cutler DR, Edwards TC Jr, Bears KH, Cutler A, Hess K, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. Ecology 88:2783–2792

De'ath G (2002) Multivariate regression trees: a new technique for modeling species-environment relationships. Ecology 83:1105–1117

Digby MJ, Saenger P, Whelan MB, 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/

Eggleston DB, Armstrong DA, Elis WE, Patton WS (1998) Estuarine fronts as conduits for larval transport: hydrodynamics and spatial distribution of Dungeness crab postlarvae. Mar Ecol Prog Ser 164:73–82

Elith J, Leathwick J (2009) Species distribution models: ecological explanation and prediction across space and time. Annu Rev Ecol Syst 40:677–697

Ellis J, Ysebaert T, Hume T, Norkko A and others (2006) Predicting macrofaunal species distributions in estuarine gradients using logistic regression and classification systems. Mar Ecol Prog Ser 316:69–83

Engle VD, Kurtz JC, Smith LM, Chancy C, Bourgeois P (2007) A classification of U.S. estuaries based on physical and hydrologic attributes. Environ Monit Assess 129: 397–412

Ferrier S, Guisan A (2006) Spatial modelling of biodiversity at the community level. J Appl Ecol 43:393–404

Franks PJS, Chen C (1996) Plankton production in tidal fronts: a model of Georges Bank in summer. J Mar Res 54:631–651

Froese R, Pauley D (2012) FishBase. www.fishbase.org (version 10/2012)

Genuer R, Poggi JM, Tuleau-Malot C (2010) Variable selection using random forests. Pattern Recognit Lett 31: 2225–2236

Gutiérrez-Estrada JC, Vasconcelos R, Costa MJ (2008) Estimating fish community diversity from environmental features in the Tagus estuary (Portugal): multiple linear regression and artificial neural network approaches. J Appl Ichthyol 24:150–162

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

Holsman K, Armstrong DA, Beauchamp DA, Reusink JL (2003) The necessity for intertidal foraging by estuarine populations of subadult Dungeness crab, *Cancer magister*: evidence from a bioenergetics model. Estuaries 26: 1155–1173

Hovel KA, Wahle RA (2010) Effects of habitat patchiness on American lobster movement across a gradient of predation risk and shelter competition. Ecology 91:1993–2002

Jelbart JE, Ross PM, Connolly RM (2006) Edge effects and patch size in seagrass landscapes: an experimental test using fish. Mar Ecol Prog Ser 319:93–102

Johnson MP, Costello MJ (2002) Local and external components of the summertime plankton community in Lough Hyne, Ireland a stratified marine inlet. J Plankton Res 24: 1305–1315

Johnson SW, Murphy ML, Csepp DJ, Harris P, Thedinga J (2003) A survey of fish assemblages in eelgrass and kelp habitats of Southeast Alaska. NMFS-AFSC-139. US Department of Commerce, Seattle, WA

Johnson SW, Thedinga J, Neff AD, Harris P, Lindeberg MR, Maselko JM, Rice SD (2010) Fish assemblages in nearshore habitats of Prince William Sound, Alaska. Northwest Sci 84:266–280

Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. John Wiley & Sons, New York, NY

Kingsford MJ, Suthers IM (1994) Dynamic estuarine plumes and fronts: importance to small fish and plankton in coastal waters of NSW, Australia. Cont Shelf Res 14: 655–672

Kissling WD, Dormann CF, Groeneveld J, Hickler T and others (2012) Towards novel approaches to modelling biotic interactions in multispecies assemblages at large spatial extents. J Biogeogr 39:2163–2178

Knudby A, Brenning A, LeDrew E (2010) New approaches to modelling fish-habitat relationships. Ecol Model 221: 503–511

Kortsch S, Primicerio R, Beuchel F, Renaud PE, Rodrigues J, Lønne OJ, Gulliksen B (2012) Climate-driven regime shifts in Arctic marine benthos. Proc Natl Acad Sci USA 109:14052–14057

Largier JL (1993) Estuarine fronts: How important are they? Estuaries 16:1–11

Laurel BJ, Stoner AW, Ryer C, Abookire AA (2007) Comparative habitat associations in juvenile Pacific cod and other gadids using seines, baited cameras, and laboratory techniques. J Exp Mar Biol Ecol 351:42–55

Legendre P, Gallagher ED (2001) Ecologically meaningful transformations for ordination of species data. Oecologia 129:271–280

Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P (2004) Screening large-scale association study data: exploiting interactions using random forests. BMC Genet 5:32–45

Magness DR, Huettmann F, Morton JM (2010) Using random forests to provide predicted species distribution maps as a metric for ecological inventory and monitoring programs. In: Smolinski TG, Milanova MG, Hassanien AE (eds) Applications of computational intelligence in biology: current trends and open problems. Studies in computational intelligence, Vol 122. Springer-Verlag, Berlin, p 209–229

Magurran AE, Henderson PA (2003) Explaining the excess of rare species in natural species abundance distributions. Nature 422:714–716

Martins GM, Thompson RC, Neto AI, Hawkins SJ, Jenkins SR (2010) Exploitation of intertidal grazers as a driver of community divergence. J Appl Ecol 47:1282–1289

Morin PJ (2011) Community ecology. John Wiley & Sons, Chichester

Mouillot D, Bellwood DR, Baraloto C, Chave J and others (2013) Rare species support vulnerable functions in high-diversity ecosystems. PLoS Biol 11:e1001569

Mueter FJ, Norcross BL (1999) Linking community structure of small demersal fishes around Kodiak Island, Alaska, to environmental variables. Mar Ecol Prog Ser 190:37–51

Neal EG, Walter MT, Coffeen C (2002) Linking the Pacific decadal oscillation to seasonal stream discharge patterns in Southeast Alaska. J Hydrol (Amst) 263:188–197

Norcross BL, Mueter FJ, Holladay BA (1997) Habitat models for juvenile pleuronectids around Kodiak Island, Alaska. Fish Bull 95:504–520

Oppel S, Huettmann F (2010) Using a random forest model and public data to predict the distribution of prey for marine wildlife management. In: Cushman S, Huettmann F (eds) Spatial complexity, informatics and wildlife conservation. Springer, Tokyo, p 151–165

Petrie ME, Ryer C (2006) Laboratory and field evidence for structural habitat affinity of young-of-the-year lingcod. Trans Am Fish Soc 135:1622–1630

R Development Core Team (2008) R: a language and environment for statistical computing. R Foundation for

Statistical Computing, Vienna

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Comput Appl Math 20:53–65

Ruppert JLW, Fortin MJ, Rose GA, Devillers R (2010) Environmental mediation of Atlantic cod on fish community composition: an application of multivariate regression tree analysis to exploited marine ecosystems. Mar Ecol Prog Ser 411:189–201

Sandri M, Zuccolotto P (2006) Variable selection using random forests. In: Zani S, Cerioli A, Riani M, Vechi M (eds) Data analysis, classification and the forward search. Proceedings of the Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society, Parma, June 6–8, 2005. Springer, Heidelberg, p 263–270

Schmiing M, Afonso P, Tempera F, Santos RS (2013) Predictive habitat modelling of reef fishes with contrasting trophic ecologies. Mar Ecol Prog Ser 474:201–216

Schoch GC, Albert D, Shanley C (2014) An estuarine habitat classification for a complex fjordal island archipelago. Estuaries Coasts 37:160–176

Segal M, Xiao Y (2011) Multivariate random forests. Data Min Knowl Discov 1:80–87

Sheaves M, Johnston R (2009) Ecological drivers of spatial variability among fish fauna of 21 tropical Australian estuaries. Mar Ecol Prog Ser 385:245–260

Stone RP, O'Clair CE (2001) Seasonal movements and distribution of Dungeness crabs *Cancer magister* in a glacial southeastern Alaska estuary. Mar Ecol Prog Ser 214: 167–176

Stoner AW, Spencer ML, Ryer C (2007) Flatfish-habitat associations in Alaska nursery grounds: use of continuous video records for multi-scale spatial analysis. J Sea Res 57:137–150

van der Wal D, Herman PMJ, Forster RM, Ysebaert T, Rossi F, Knaeps E, Plancke YMG, Ides SJ (2008) Distribution and dynamics of intertidal macrobenthos predicted from remote sensing: response to microphytobenthos and environment. Mar Ecol Prog Ser 367:57–72

Wedding LM, Friedlander AM (2008) Determining the influence of seascape structure on coral reef fishes in Hawaii using a geospatial approach. Mar Geod 31:246–266

Wedding LM, Lepczyk CA, Pittman SJ, Friedlander AM, Jorgensen S (2011) Quantifying seascape structure: extending terrestrial spatial pattern metrics to the marine realm. Mar Ecol Prog Ser 427:219–232

Wehrly KE, Breck JE, Wang L, Szabo-Kraft L (2012) A landbased classification of fish assemblages in sampled and unsampled lakes. Trans Am Fish Soc 141:414–425

Weingartner T, Eisner L, Eckert GL, Danielson S (2009) Southeast Alaska: oceanographic habitats and linkages. J Biogeogr 36:387–400

Whitlow WL, Grabowski JH (2012) Examining how landscapes influence benthic community assemblages in seagrass and mudflat habitats in southern Maine. J Exp Mar Biol Ecol 411:1–6

Wiens JJ (2011) The niche, biogeography and species interactions. Philos Trans R Soc B 366:2336–2350

Wisz MS, Pottier J, Kissling WD, Pellissier L and others (2013) The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. Biol Rev Camb Philos Soc 88:15–30

Editorial responsibility: Christine Paetzold, Oldendorf/Luhe, Germany

Submitted: January 11, 2013; Accepted: November 20, 2013
Proofs received from author(s): February 21, 2014