LANDSCAPE MODELING OF THREESPINE STICKLEBACK OCCURRENCE

IN SMALL SOUTHEAST ALASKA LAKES

By

Dave Gregovich

RECOMMENDED: ________________________________

____

____
Advisory Committee Chair

WW Smoker
Director, Fisheries Division

APPROVED: ________________________________
Dean, School of Fisheries and Ocean Sciences

Lawrence K Duffy
Dean of the Graduate School

Dec 8, 2007
Date

LANDSCAPE MODELING OF THREESPINE STICKLEBACK OCCURRENCE

IN SMALL SOUTHEAST ALASKA LAKES

A

THESIS

Presented to the Faculty

of the University of Alaska Fairbanks

in Partial Fulfillment of the Requirements

for the Degree of

MASTER OF SCIENCE

By

Dave Gregovich, B.S.

Fairbanks, Alaska

December 2007

Abstract

Although threespine stickleback (*Gasterosteus aculeatus* L.) are known to inhabit a wide range of habitats, their distribution in lakes across Southeast Alaska is not known. Threespine stickleback are an important prey item for many consumers in freshwater ecosystems. Additionally, isolated populations may be genetically unique and thus important from a conservation perspective. This study focused on identifying landscape factors and models useful in predicting the presence of threespine stickleback in small (0.5-5 ha) lakes of Southeast Alaska. Stickleback occurrence was assessed via snorkeling and minnow trapping in 54 lakes, which were divided into calibration (n=36) and prediction (n=18) data sets. A number of models representing four methodologies—generalized linear models, generalized additive models, classification trees, and artificial neural networks—were built based on the calibration set, cross-validated, and evaluated by prediction to the test set of lakes. Lake elevation, distance from saltwater, and slope of lake outlet stream were the most useful predictors of stickleback occurrence. Results suggest that the likelihood of stickleback presence is highest in low elevation lakes near the coast. Human development and recreational activity also tends to be common in these areas, and so land-use planning should account for the high potential occurrence of threespine stickleback here.

Table of Contents

Page

Page

Page

List of Figures

Page

List of Tables

List of Appendices

Acknowledgements

I would like to thank my advisor Mark Wipfli and agency supervisor (ADF&G) Brian Frenette for their guidance, committee members Joe Margraf and Gordon Haas for their assistance, and statisticians Franz Mueter (Sigma Plus Consulting), and Randy Mullen (ADF&G) for their advice. Jeff Williams and Carl Lundquist are also thanked for their valuable assistance in the field. Finally, I would like to thank fellow students Aaron Martin, Bessie Green, Bruce Medhurst, Cassie Mellon, Chrissy Apodaca, Dan Rinella, Joshua Peirce and Meagan Boltwood of Dr. Wipfli's lab for their advice and support.

General Introduction

Landscape modeling is a rapidly developing field with many applied uses (Duncan & Lach, 2006; Kollarits *et al.,* 2006). Foremost amongst these uses is identification and conservation prioritization of high-value fish and wildlife habitats at broad spatial scales. The decision-making process of resource managers can be greatly facilitated by information on the spatial distribution of these resources (Johnson & Gage, 1997; Klemas, 2001; Mortberg *et al*., 2007).

Landscape modeling of aquatic organisms and their habitats in Southeast Alaska is in its incipient stages, although fish occurrence has been modeled successfully with landscape predictors elsewhere (Porter *et al*., 2000; Torgerson & Close, 2004; Van Zyll de Jong & Cowl, 2005). For this reason, any modeling efforts undertaken in this region may potentially yield useful information and potential models for the specific species in question, and may additionally serve as precedent for future modeling of the distribution of other aquatic organisms, including fishes. For instance, landscape factors found useful in modeling one organism in a geographic region may be transferable to both closely-related and disparate taxa (MacNally & Fleishman, 2002; Gutierrez *et al*., 2005). Likewise, modeling methods found useful in a geographic sub-region may be transferable across additional regions (Fielding & Haworth, 1995; Magnuson *et al*., 1998).

Resources directly consumed or otherwise utilized by humans are of obvious management importance. However, there has been increasing focus in recent times on ecosystem components that, while not of direct economic benefit to humans, play important roles in supporting populations of organisms from which we do obtain economic benefit (Gutierrez *et al*., 2005; Quist *et al*., 2006). The distribution of such non-target organisms may indicate habitats available to other commercially or recreationally important species (Willett, 2001; MacNally & Fleishman, 2002).

Determining presence of these non-game species or suites of species may in turn be helpful in classifying habitats in terms of their relative importance to humans. The presence or absence of particular species in a habitat can have many ecological ramifications important in describing the biotic attributes of the habitat (Sergio *et al*., 2004; Pearman *et al*., 2006). This is likely particularly true in the case of a species such as the threespine stickleback, which is directly consumed by species important to humans. The occurrence of such trophic resources stands alongside other physical and biotic attributes to provide a rounded description of habitat characteristics. When all habitat characteristics are considered together—both physical and biotic—they can be used to classify and assess habitat value, which can in turn be fed into resource management decision making processes (Theobald *et al*., 2000; Ostwald *et al*., 2002).

.

Direct exploitation of stickleback by humans is at most minimal. However, stickleback are consumed by a host of vertebrate and invertebrate consumers, including many of direct benefit to humans. Piscine consumers of threespine stickleback include cutthroat (*Onchorhyncus clarki*) and steelhead trout (*O. mykiss*) and coho (*O. kisutch*) and Atlantic salmon (*Salmo salar*), while mammalian consumers include otter, mink, and seal (Reimchen, 1994). Birds known to rely on stickleback as a dietary component include belted kingfishers, great blue herons, and many species of loon and grebe (Reimchen, 1994; Ruggles, 1994). The case for the threespine stickleback's role as a vital trophic link and ecological importance in aquatic ecosystems is strong.

While stickleback are ecologically important as a prey item, they are also of scientific and conservation value as an evolutionary 'supermodel' (Gibson, 2005). Isolated stickleback populations display rapid phenotypic adaptation to newly colonized habitats (Bell *et al*., 2004; Kristjansson, 2005). Body armor reduction, especially a reduction in the number of lateral plates, has been the most commonly documented adaptation, although other adaptations include traits associated with trophic morphology, life history, reproductive behavior, and coloration (Foster *et al*., 2003). Such adaptive divergence associated with local environmental conditions suggests the possibility that postglacial landscapes may harbor a large number of isolated populations with unique morphologies of scientific and conservation interest (Foster *et al*., 2003). Gaining knowledge of where such populations might reside is the first step in their identification and conservation.

Threespine stickleback occur across a wide geographic range, from Baja California to the Arctic coast in the Pacific, and from Chesapeake Bay to Baffin Island in the Atlantic (Morrow, 1980). Within this range, they display a variety of life history strategies—including marine, anadromous, riverine, and lacustrine forms (Foster *et al*., 2003). Although information on the broad-scale geographic distribution of stickleback exists, there has been very little investigation of environmental factors associated with stickleback occurrence at smaller scales—Hagen & Gilbertson (1972) provide an exception—and none in Southeast Alaska.

The distribution of other fish species in freshwater has been successfully described and predicted using landscape-level variables for stream (Kruse *et al*., 1997; Oakes *et al*., 2005), and lake (Olden & Jackson, 2001; Hershey *et al*., 2006) habitats. Lake studies in other geographic regions have indicated elevation, lake surface area, distance to nearest source habitat, and slope of lake outlet stream as important predictors of fish species occurrence (Matuszek & Beggs, 1988; Magnuson *et al*., 1998; Hershey *et al*., 2006).

Models of species occurrence are binary classifiers in which presence and absence (or when absence has not been assessed, presence and 'available') data records are analyzed against environmental correlates. The output of a species occurrence model is a set of occurrence probabilities (where $0 \geq P \geq 1$) associated with each combination of the environmental variables. Many statistical methods for species occurrence modeling exist. The most prevalent of these methods until recently was logistic regression, a generalization of linear regression (hence termed a ‘generalized linear model’ or ‘GLM’) that accounts for the non-linearity and non-normal error distribution of binary data. Although logistic regression is still widely used, a number of competing methods have seen increased use in species occurrence modeling. These methods include (but are not limited to—see Elith *et al*., 2006) generalized additive models (GAM), classification trees (CT), and artificial neural networks (ANN). With the increase in the number of methods being used has come the challenge of selecting amongst them, for they all have merit dependent on specific study objectives (Segurado & Araujo, 2004; Pearson *et al*., 2006).

In this study, the presence of threespine stickleback in small lakes of Southeast Alaska is modeled with mostly landscape-level (GIS) variables. A calibration set of lakes (n=36) in the northern part of Southeast Alaska was used for building stickleback occurrence models. Model stability and predictive capability were then assessed via leave-one-out (n-fold) cross-validation, and prediction to a geographically separate set of lakes (n=18) in the southern part of Southeast Alaska.

There are two main objectives to this study:

1) Determine the landscape factors most useful in modeling stickleback distribution in small lakes of Southeast Alaska.

2) Compare modeling methods and levels of modeling complexity to determine which modeling strategies are most effective in modeling stickleback distribution in small lakes of Southeast Alaska.

In addition, there are three secondary objectives:

1) Compare the distribution of salmonid fishes in small lakes with the distribution of stickleback in these lakes.

2) Discuss how results from the small lakes chosen for this study may apply to other lentic waterbodies.

3) Make recommendations for future species occurrence sampling programs in the Southeast Alaska region

The first main objective is addressed in chapter one by performing univariate and multivariate analyses using GLM, and measuring the explanatory and predictive performance attributed to individual environmental predictor variables. The second main objective is addressed in chapter two by building models using the GLM, GAM, CT, and ANN methods mentioned above and comparing their explanatory and predictive performance.

The secondary objectives are addressed in the chapter three of the thesis. The comparison of salmonid distributions with that of stickleback is performed using simple GLM models. The distribution of 0.5-5 ha lakes is compared to that of other lake sizes in the northern area, and the utility of models generated in this study is discussed in relation to these other lake sizes. Lastly, some simple examples from the current work are used to illustrate improvements that might be made to future species occurrence study designs in lakes of Southeast Alaska.

Study area

Southeast Alaska is characterized by its rugged terrain, with countless mountain peaks greater than 1000 m, even though no point in the region is greater than 100 km from saltwater (Figure 1). The maritime climate of the region is moist and moderate, with average rainfall exceeding 250 cm in many portions of the region and moderate annual temperatures. There are over 25,000 lakes and ponds in the region (USFS, 2003a).

Southeast Alaska was entirely covered in ice 15,000 years ago, with the exception of a few unglaciated mountain peaks and isolated refugia on some of the islands of the outer coast (Mann, 1986). Furthermore, the little ice age (~350-150 years BP) experienced glacial advance—particularly out of mainland valleys—which shapes the distribution of flora and fauna found in these valleys to the present day. The land has experienced large amounts of postglacial uplift as it rebounds from the weight of the ice mass under which it was once covered. This has resulted in terrains once inundated with marine waters now existing at as high as ~230 m elevation (Hastings, 2005).

The study area chosen within this context consists of northern (calibration) and southern (prediction) areas. The northern area comprises that part of the mainland near the town of Juneau that is bounded by the Berners River to the north and the Taku River to the south, inclusive of the portions of these two river valleys that are within the borders of Alaska (Figure 1). Most of the eastern portion of this area is in the Boundary Range Icefields ecological subsection (Nowacki *et al.*, 2001), and is covered in glaciers. The western portion of the northern area is within the Stephens Passage glaciomarine terrace ecological subsection, while the southeastern corner of the study area is within the Stikine-Taku River Valleys subsection. The southern area consists of that portion of Revillagigedo Island that falls outside of Misty Fjords National Monument. It is composed of a number of ecological subsections, including Traitor's Cove Metasediments, Bell Island Granitics, the Behm Island Complex, and other smaller intrusions (Nowacki *et al.*, 2001).

Because small lakes are not consistently mapped in the lakes GIS coverage used for site selection (USFS 2003a), a minimum lake size was chosen of 0.5 ha to avoid selecting lakes via GIS that do not actually exist on the ground. This lower surface area threshold also excluded many small waterbodies such as muskeg and alpine ponds which seemingly (although not investigated) have a very low probability of stickleback presence. To ensure that sampling effort could be undertaken at each lake for a minimal time period yet still adequately assess fish presence/absence, an upper lake surface area threshold of 5 ha was chosen.

Lakes in this 0.5-5 ha size range comprise 34% of the total number of lakes in the northern study area, with the majority of all mapped lakes (54%) being smaller (<0.5 ha) and 12% being larger (>5 ha). A similar size distribution is seen in lakes of the southern study area (Table 1).

Table 1. Proportional representation of lake size classes in the two study areas.

Lake size (ha)	Northern area	Southern area
<0.5	141(54%)	496(56%)
0.5-5	90(34%)	279(32%)
5	32(12%)	104(12%)

Lake selection was stratified by elevation in order to represent a range of lake elevations in the study. The rationale for stratifying study lakes by elevation was related to the natural history of the region, specifically the postglacial uplift or terrains as evidenced by marine deposits existing currently at elevations up to 230 m, and the continued uplift of the land to present (Larsen *et al.*, 2004). Hastings (2005) suggested that the former glacial depression of landforms allowed colonization by fish of lower-elevation habitats, but that high-elevation lakes and streams may have been, and are currently, inaccessible to fish due to the extreme high-gradient of their connections with source, marine populations.

Ninety lakes of surface area 0.5-5 hectares exist in the northern study area and 279 in the southern study area as identified in the Tongass National Forest lakes GIS coverage (USFS, 2003a). To ensure independence of sampling units, lakes which were directly connected to each other by surface waters were not included in the potential pool. From the remaining lakes, 36 study lakes in the northern region and 18 in the southern region were identified via stratified random selection based on six elevation quantiles (which divided the number of lakes in this size class into six equal-sized groups) (Table 2):

Table 2. Elevation-defined strata

	Elevation range (m)	
Stratum	Northern area	Southern area
1	0-20.6	0-67.0
2	20.6-32.2	67.0-117.4
3	32.2-103.6	117.4-154.5
4	103.6-323.6	154.5-240.2
5	323.6-732.4	240.2-510.5
6	>732.4	>510.5

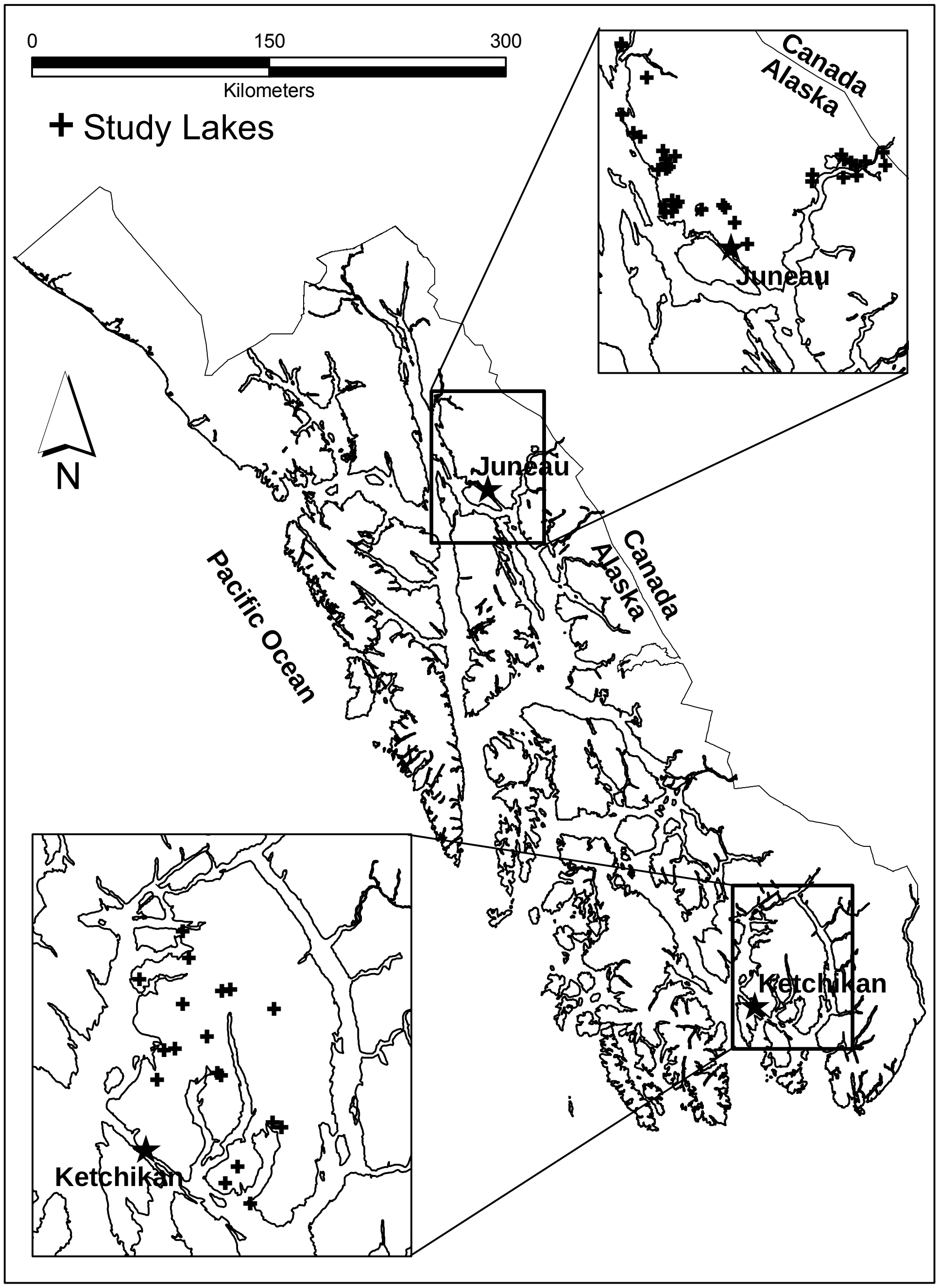


Figure 1. Southeast Alaska showing the northern and southern study areas.

Stickleback detection

At each lake both minnow trapping and snorkeling were performed. Eight minnow traps were set evenly around the perimeter of each lake, generally in water < 1-m depth, although when lake margins were steep traps were set in deeper water (depth 1-5 m). Traps were baited with salmon eggs, and set for approximately 1.5 hours (range 1.3-2.0 hr) at each lake.

A snorkel survey was conducted around the entire perimeter of each lake at 1-m depth parallel to shore, with the snorkeler scanning to both sides and swimming at a rate <0.5m/s. The snorkeler skirted around the outside of macrophytes or pieces of large wood when such structure precluded snorkeling in water of one-meter depth. Secchi disc readings were taken at three locations in the lake just prior to snorkel surveys, at two lakes the secchi reading was < 1 m. At those two lakes, occurrence data is based on minnow trapping results alone.

Observations of other fish species in the above sampling efforts were also recorded. In addition, 45 minutes of angling was undertaken at each lake with light spinning gear and a Luhr Jensen™ 'super duper' artificial lure.

Chapter 1. Landscape correlates of threespine stickleback occurrence in small lakes of Southeast Alaska

1.1 Introduction

Landscape factors have been used to effectively model organismal distribution in many settings (Thuiller, 2003; Steel *et al*., 2004; Luoto *et al*., 2006). Landscape factors useful in explaining species occurrence fall into broad categories of physical and biotic landscape characteristics, measures of terrain ruggedness, proximity to source habitats, habitat patch size, and habitat suitability.

Prediction of species occurrence can yield information helpful in making management decisions (Zabel *et al*., 2002). Of particular use in this respect are landscape factors available across wide areas. Of lesser use are factors which, whether measured in the field or derived from GIS, are not available for all possible habitats. Model predictions in the latter case can only be used, in the worst case scenario, to predict occurrence at sites that have been visited in the field.

Landscape factors chosen for modeling should ideally have either a direct relationship with the species being modeled or an indirect relationship with a factor or factors known to have a direct relationship with the occurrence of the study species (Guisan & Thuiller, 2005). Direct relationships with threespine stickleback (or any other species) occurrence can fall broadly into two categories: 1) relationships with population persistence (habitat suitability over time), and 2) relationships with habitat accessibility. Factors associated with these two broad categories of relationship have been referred to as ‘extinction’- or ‘colonization’-related factors (Magnuson *et al*., 1998; Hershey *et al*., 2006).

The factors having a direct causal relationship with stickleback occurrence are largely unknown. Additionally, even if those factors were known, they may be difficult to obtain over landscape-level scales. For instance, even if it was known that stickleback preferred a certain type of submergent vegetation for their spawning nest, would it be possible to detect this vegetation remotely and therefore use it for prediction to lakes not visited? Similarly, if a 17-cm cascade over the rocks of a stream was known to be the height limit of an obstruction passable by stickleback, how would that be helpful information if the location of such fine-scale obstructions to migration is not known? This reasoning leads to an ideal choice of factors that are either readily available, or that can be made available via remote sensing, broadly across landscapes. And ideally, such factors will have a direct (or plausible indirect) relationship with the species being modeled.

Very little is known about stickleback migration abilities or habitat requirements in Southeast Alaska. However, it is known that stickleback in similar post-glacial regions colonized from the marine environment (Bell *et al*., 1993). Because Southeast Alaska was almost completely under ice 15,000 years ago, it appears most if not all freshwater fish populations in the region must have invaded in the post-glacial epoch (Hastings, 2005). The postglacial uplift of terrains that has taken place subsequent to glacial recession in Southeast Alaska means fishes colonizing from the marine environment may have historically had access to lakes that now exist at relatively high elevations and may be presently isolated from colonization. However, there are lakes that exist at extremely high elevations, and due to steep outlet streams, have never been accessible to fishes via a stream course. For such lakes, transport by birds seems the only colonization route, although this mode of dispersal has not been documented (T. Reimchen, pers. comm.)

For this study, 14 environmental factors with likely influence on stickleback occurrence (via extinction or colonization) were assessed for their use in explaining and predicting threespine stickleback occurrence in small lakes of Southeast Alaska using generalized linear models (GLM). This assessment included the utility of factors in explaining variability in occurrence data, as well as predicting stickleback occurrence in an independent set of lakes.

1.2 Methods

1.2.1 Landscape and environmental factors

The 14 environmental factors used in this study were largely obtained from GIS analysis, although four of the factors were measured or assessed in the field at the time of each lake sampling visit (Table 3). Lake elevation was obtained directly from a U.S. Forest Service lakes GIS coverage (USFS, 2003a). These elevation values were found consistent

Table 3. Environmental variables used to model stickleback occurrence in small lakes of Southeast Alaska, and the definition, source, and ecological category (sensu Magnuson *et al*., 1998; Hershey *et al*., 2006) of each variable. For adjacent landscape variables, the width (m) of the band around the lakeshore analyzed is in parenthesis.

Environmental Factor	Source	Category
Lake specific factors		
1) Lake elevation (m)	USFS lakes GIS coverage[1]	Colonization
2) Lake depth (m)	Field measured	Extinction
3) Lake perimeter (m)	USFS lakes GIS coverage[1]	Extinction
4) Lake surface area (ha)	USFS lakes GIS coverage[1]	Extinction
5) Lake substrate (binary)	Field measured assessment (mineral-dominated or organic-dominated)	Extinction
Adjacent landscape factors		
6) Wetlands coverage (200 m)(%)	National wetlands inventory[2] (NWI)	Extinction
7) Mean slope (100 m)(%)	Based on digital elevation model grid cells within 100 m of lake[1,3]	Extinction
8) Mean slope (1000 m)(%)	Based on digital elevation model grid cells within 1000 m of lake[1,3]	Extinction
9) Linear distance from lake to marine shoreline (m)	Based on USFS lakes and marine shoreline GIS coverages	Colonization
Inlet and Outlet stream factors		
10) Presence of inlet stream(s) (bin.)	Assessed in the field	Extinction
11) Presence of outlet stream (bin.)	Assessed in the field	Colonization
12) Length of outlet stream (m)	Digitized from aerial Photography	Colonization
13) Mean slope of outlet stream (%)	Based on digital elevation model grid cells intersected by digitized outlet stream[1,3]	Colonization
14) Mean slope of outlet stream (%)	Based on digital elevation model grid cells intersected by digitized outlet stream[1,3]	Colonization

[1] USFS (2003a); [2] USFWS (2006); [3] USGS (2004)

with independent measurements derived from a Shuttle Radar Tomography Mission digital elevation model (SRTM-DEM) (USGS, 2004) and from helicopter altimeter readings taken at the time of field visits. Lake perimeter and surface area were digitized from USFS ortho-rectified aerial photography (USFS, 2003b). The length and location of lake outlet streams were digitized using a combination of the ortho-rectified digital photography and 15-m contour lines generated from the SRTM-DEM data (USGS, 2004).

Although all lakes did not have outlet streams, a likely location (and length) of the flow path from the lake to saltwater was obtained using the aerial photography and SRTM-DEM data for all lakes. . This is because the lack of an outlet stream—defined for this study as either a surface water outlet at the time of visit or a scoured bed (>0.3-m wide) devoid of perennial vegetation—was not thought to preclude fish access to a lake over long periods of time. Information on the length and gradient of the most plausible flow path (i.e. during extreme high water events) was still thought valuable, even in the absence of a permanent lake outlet stream.

Although the length of a lake outlet stream was generally measured to saltwater, an exception was made for those lakes near the Taku River. The lower Taku River is a low-gradient (<1%) river in immediate proximity to saltwater (<20 km) that contains known populations of stickleback (K. Kissner, pers. comm.). In addition, tidal influence extends to Twin Glacier Creek (R. Host, pers. comm.), and so all study lakes flow either directly into saltwater or into the Taku River within 10 k of saltwater influence. Because of this saltwater influence and the presence of stickleback, it was thought that the lower Taku River represents a source habitat for colonizing stickleback similar to that of the marine environment.

The mean and maximum gradient of lake outlet streams was measured by summarizing the values of the SRTM-DEM for those grid cells intersected by the digitized stream course. Two other slope-related factors—the mean slope of terrain analyzed both within 100 m, and within 1000 m, of the lake shoreline—were obtained from summarizing the SRTM-DEM cell-values within these respective bands around the lake. The percentage of wetlands coverage (within 200 m of the lakeshore) was calculated using the National Wetlands Inventory GIS coverage (USFWS, 2006).

Lake substrate was characterized during lake sampling visits by observing the bottom of the lake and prodding the substrate with a graduated 1.5-m dowel stick. The depth of deposited organic material was measured every 50 m around the lake perimeter. If >50% of these measurements indicated > 0.3 m depth of organics, the entire lake was classified as having substrate dominated by organics. Otherwise, the lake was classified as having mineral substrates. Lake depth (m) was taken as the mean of three depth measurements collected with a hand-held depth sonar at evenly spaced points along the long axis of each lake.

1.2.2 Statistical analysis

Statistical analysis of the factors that explain and predict stickleback occurrence was performed using GLM (logistic regression) which uses the logit link function to convert environmental factor values and presence/absence data to a probability of occurrence (Hosmer & Lemeshow 2000):

$P = e^{u} / (1 + e^{u})$, where *u* takes the form of the more familiar linear regression equation—

$u = A + b_1 X_1 + b_2 X + ... + b_n X_{nv}$

where *A* = regression constant, b_n= regression coefficients, and X_m= independent variables. In using GLM, is not required that predictor variables be normally distributed, linearly related, or of equal variance (Tabachnick & Fidell, 2000; Porter *et al.*, 2000). Individual factor contribution to model performance was assessed for three stages of model assessment: 1) model calibration—how well a model explained the data it was based on (calibration data), 2) leave-one-out cross-validation—how well a model performed when iteratively built on n-1 of the calibration cases and used to predict to the single removed case, and, 3) prediction—how well a model built on the calibration set of lakes predicted the data in the southern, prediction set of lakes.

Two metrics were used to assess individual factor contribution to model performance. The first was McFadden's ρ^2, which is the proportion of the null deviance that each model explained (Porter *et al.*, 2000; Johnson *et al.*, 2002). A second metric, the area under the receiver-operating characteristic curve (Fielding & Bell, 1997; Elith & Burgman, 2002), was also used. AUC is a measure of how well a model discriminates between presence and absence records (Boyce *et al*., 2002). A model's AUC score is the rate of true positives to false negatives that a model yields across the entire range of possible threshold values between 0 and 1. This is why AUC is termed 'threshold independent' and considered more useful than other metrics available (Liu *et al*., 2005). The AUC score was used to compare model (and thus individual predictor) performance in cross-validation and prediction as well. The AUC value can be used to compare model performance in all three stages—calibration, cross-validation, and prediction—of model assessment.

Three strategies were used to elucidate the contribution of individual variables in explaining and predicting stickleback occurrence. The first and simplest strategy was to run univariate GLM models with one individual predictor at a time to assess the proportion of deviance (McFadden's ρ^2) explained by each individual predictor. Univariate model performance was also assessed on cross-validation and in prediction to the test set, using AUC as the performance metric.

The second strategy was to run all possible subsets of two- and four-factor models that could be obtained from combinations of the 14 total environmental predictors, and compare—between all predictors—the average amount of deviance explained when each predictor was included in a model. The number of factors included in the 'all subsets' models was limited to four to reduce computing time. The amount of deviance explained by models including each factor was scaled within each model category (1, 2, and 4 factor models) to allow comparisons between categories.

The third strategy was to investigate ‘secondary effects’, factors that may not explain significant variance in their own right, but decrease model deviance when included after primary effects are already in the model. In this strategy, each factor that explained at least 0.2 of the null deviance was included individually in models as a ‘main effect’. Then, each of the remaining 13 factors were included one-at-a-time into models with the ‘primary’ (>0.2 null deviance) factors already included. The mean amount of residual deviance explained upon inclusion of each of these ‘secondary effects’ was then compared across models containing each combination of main and secondary effect.
A final step concerned those factors which were considered primarily ‘extinction’-related or ‘colonization’-related (Magnuson *et al*., 1998; Hershey *et al*., 2006). Extinction factors are factors related largely to the suitability of a lake for sustaining a population for a long period of time. Colonization factors are those factors which are more closely associated with how accessible a lake is to colonization. The classification of each factor used in this study (extinction- or colonization-related) is included in Table 3.

Comparisons between those factors categorized as extinction- and colonization-related were performed by taking the mean of each category’s performance (using McFadden’s ρ^2) on univariate analyses. A similar comparison was also performed by building a single model using all extinction factors and another using all colonization factors and comparing the two.

All statistical analyses were performed in the ‘R’ statistical environment (R Development Core Team, 2006); all GLM models were built using the 'glm' function that comes with the base R program.

1.3 Results

1.3.1 Main effects

The factors of greatest explanatory importance were lake elevation, linear distance to marine shoreline, length of outlet stream, maximum outlet stream gradient, and mean outlet stream gradient (Figure 2). These factors explained the highest proportion of deviance individually—and also had significant non-zero ($p<0.05$) coefficients—in univariate models (Hosmer & Lemeshow, 2000). The proportion of deviance explained by univariate models ranged from a low of $\sim 4 \times 10^{-5}$ (univariate factor: mean slope in 100 m band around lake) to a high of 0.29 (univariate factor: lake elevation). Length of lake outlet stream and linear distance from lake to saltwater explained similar amounts of deviance (0.22 and 0.26 respectively); these two factors, not surprisingly, also show high collinearity (Pearson correlation coefficient = 0.84).

Comparisons of univariate performance in the calibration, cross-validation, and prediction stages (Figure 3) revealed that all five of the factors identified as important main effects above displayed relative stability on cross-validation, not showing a precipitous drop in AUC value from calibration to cross-validation. AUC score was expectedly lower on cross-validation for all factors, because models were predicting to a data record not included in model construction. Even so, the AUC score for some factors (presence of outlet stream, presence of inlet stream, mean slope in 100 m band around lake) dropped drastically on cross-validation relative to other factors. Such factors appear to have a spurious relationship with stickleback occurrence in the original, calibration dataset.

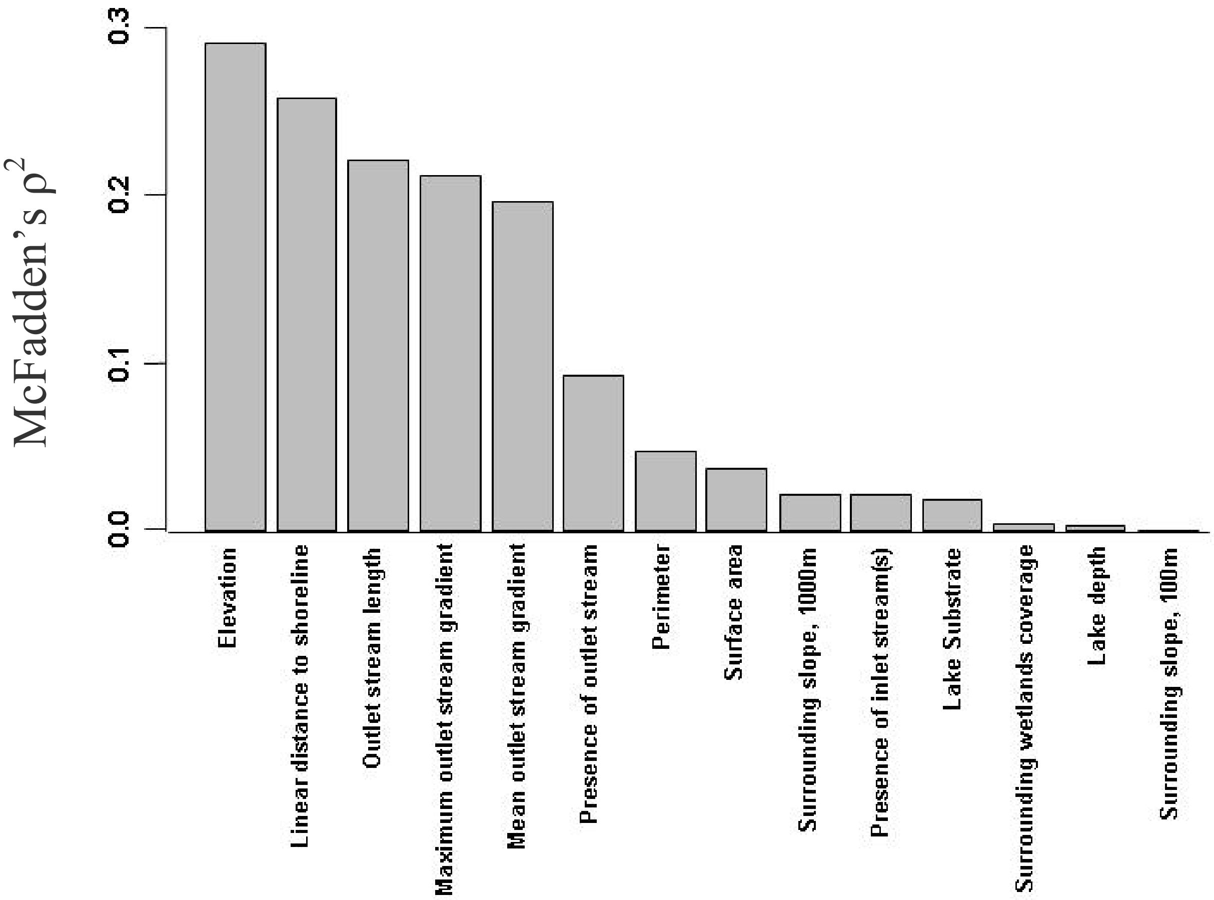


Figure 2. Proportion of null deviance (McFadden's ρ^2) explained by each univariate GLM model.

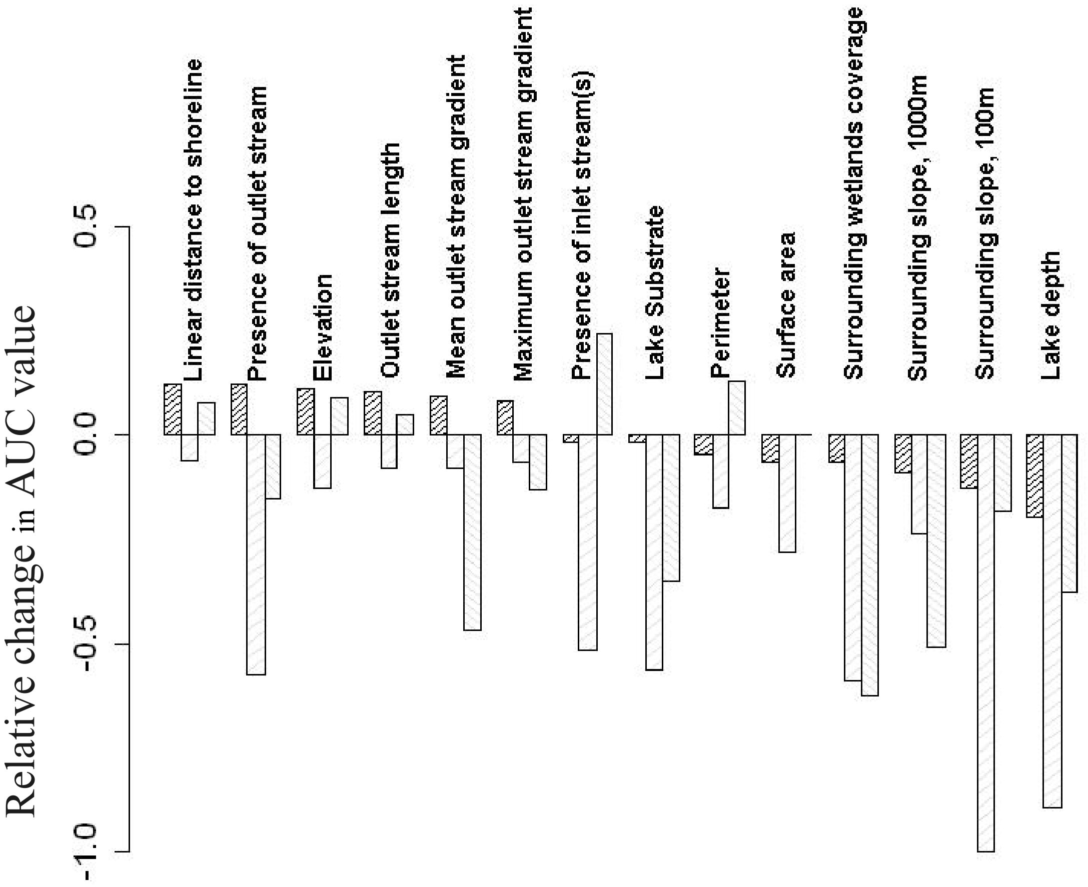


Figure 3. Stability of environmental variable importance on cross-validation and prediction to the test set as measured by area under the ROC curve (AUC): 1) calibration AUC value - mean AUC score (▩), 2) relative change in AUC value on cross-validation (▨), and, 3) relative change in AUC value on prediction to test set (▧).

Most of the individual factors that produced relatively useful univariate models continued to contribute to model performance when included in two-factor (bivariate) and four-factor models (Figure 4). Bivariate and four-factor models that included lake elevation displayed the highest McFadden's ρ^2 values. In contrast, lake outlet stream length explained a relatively small portion of the deviance in bivariate and four-factor models when compared to univariate models including lake outlet stream length alone. Using presence of lake outlet stream in bivariate and four-factor models increased McFadden's ρ^2, although the deviance outlet stream presence explained in univariate models was below the mean for the univariate models.

On prediction to the test set of lakes univariate models including linear distance to saltwater, lake elevation, and length of outlet stream displayed higher AUC values than on the original data. Similarly, univariate models including presence of an inlet stream and lake perimeter had AUC scores higher on prediction to the test set. Mean outlet stream gradient was the single individual effect (as identified above) that predicted stickleback occurrence in the test set relatively poorly.

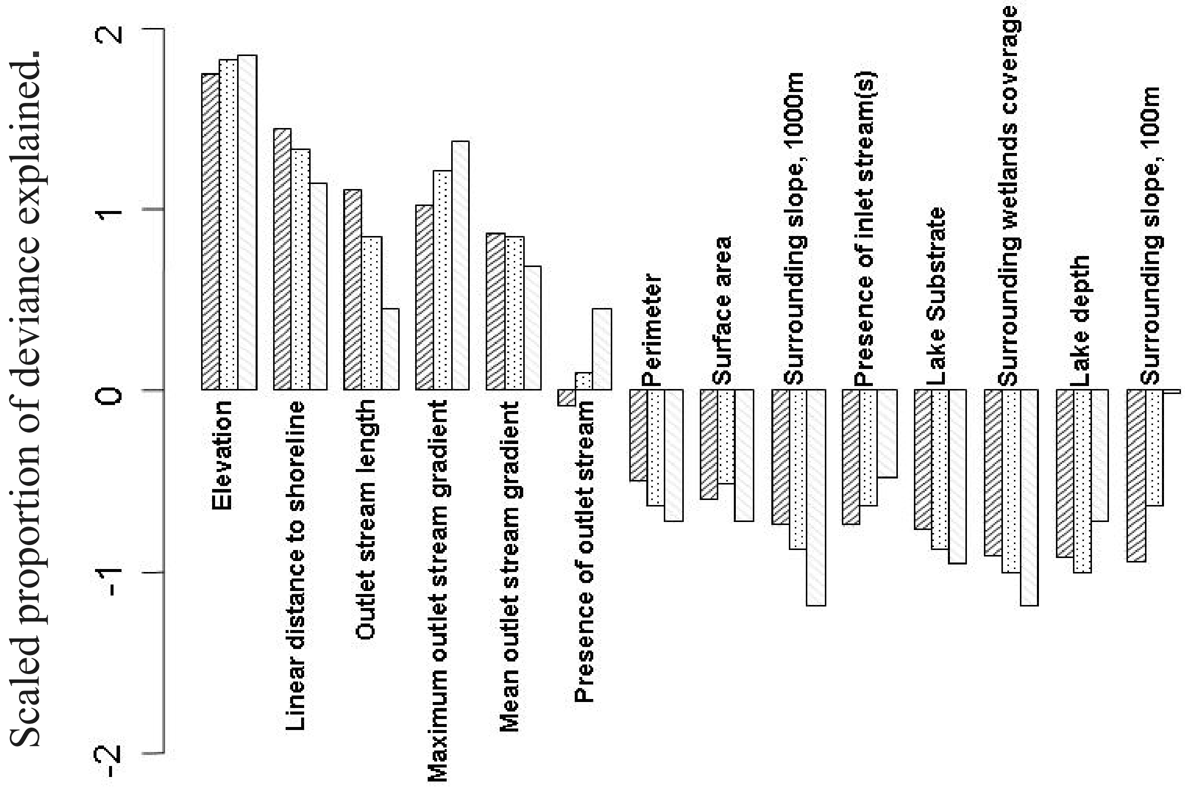


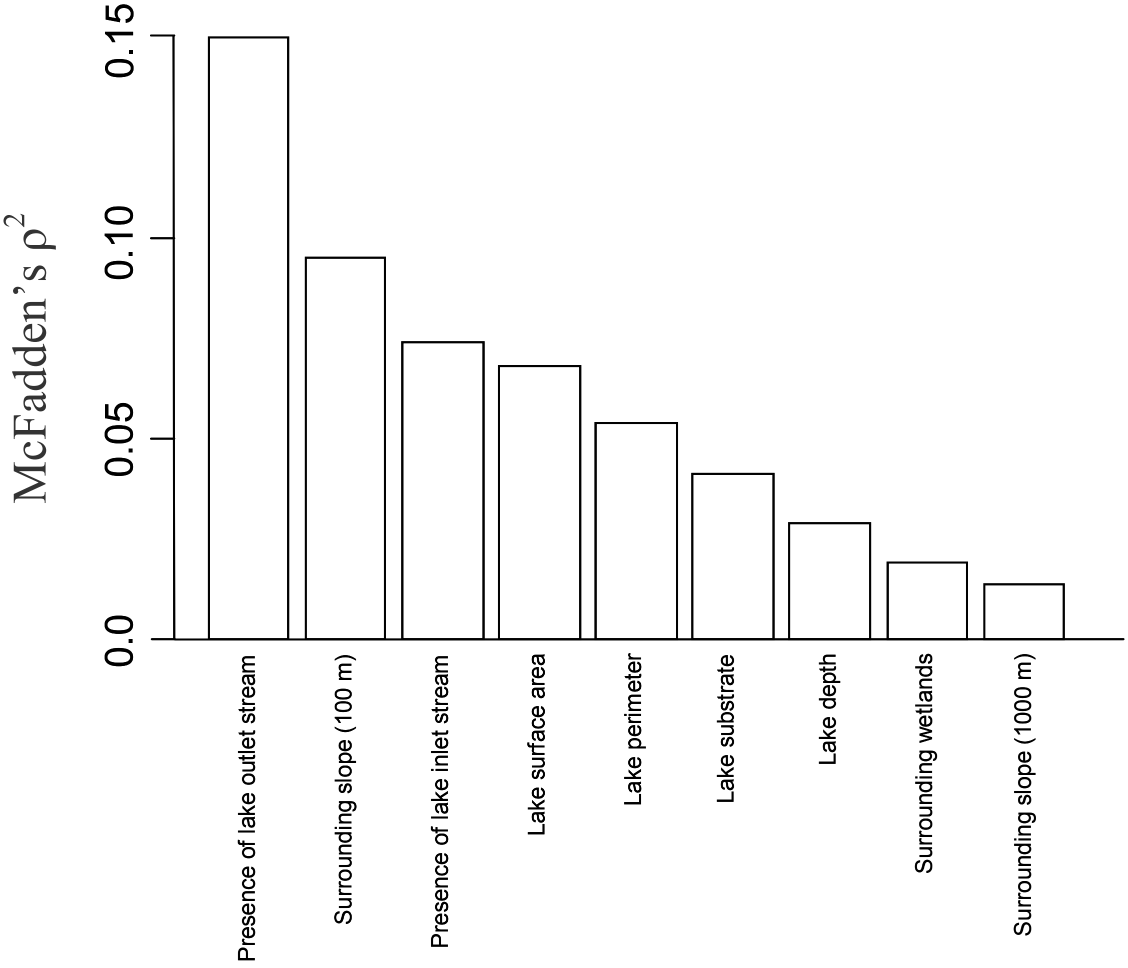
Figure 4. Scaled deviance explained by each environmental factor. For univariate models (▨), this is the scaled deviance of the single univariate model. For two- (░) and four-factor (□) models, this is the mean scaled deviance averaged across all possible models which include each of the 14 environmental variables.

1.3.2 Secondary effects

When secondary effects were added to models in which a main effect was already included, these secondary effects did not improve model performance dramatically. The one possible exception is the presence of outlet stream factor. The presence of a lake outlet stream, although not producing a model significant at the p-0.05 level individually, was the factor that decreased model deviance to the highest degree when added to each of the main effects above (a mean of 0.15 of the total model deviance) (Figure 5). The presence of an outlet stream also increased model AUC score on cross-validation (mean increase ~0.06) to a greater extent than any of the other potential 'secondary effects'. Inclusion of lake surface area and presence of an inlet stream both decreased model deviance and increased model cross-validation AUC value. However, these two factors did not decrease model deviance or increase cross-validation AUC values to a great extent (< 0.10 change in model deviance and < 0.05 change in cross-validation AUC value), nor did any other secondary factors.

1.3.3 Extinction-related vs. colonization-related factors

Colonization-related factors contributed to model explanatory performance more than extinction-related factors (Figure 5), explaining a larger portion of model deviance. This was the case when comparing both the mean deviance explained by univariate models comprised of each of the categories (Figure 5), and the total deviance explained in two models built with all of the factors of the respective categories included (Figure 6).



.

Figure 5. Mean proportion of deviance (McFadden's ρ^2) explained by inclusion of secondary effects in models, given prior inclusion of a primary (proportion of deviance explained > 0.20) effect.

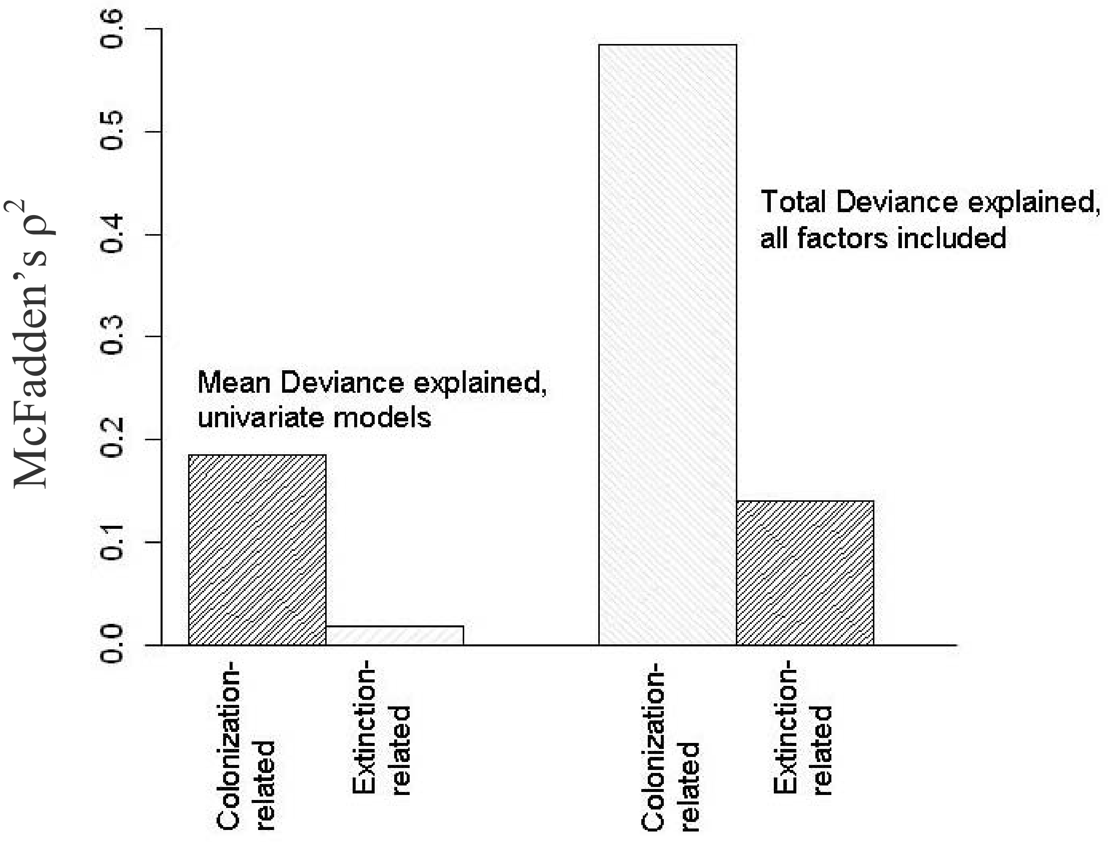


Figure 6. Mean and total proportion of null deviance (McFadden's ρ2) explained by extinction- and colonization-related environmental variables.

1.4 Discussion

1.4.1 Main effects

It is difficult to assess the individual roles of environmental predictors in an observational study, inherent in which are interactions between variables and possible missing covariates (Beier & Noss, 1998; Li & Wu, 2004; Parker *et al*., 2005). However, this chapter has attempted to do just that from a number of angles, using deviance explained on calibration—and model AUC values in 3 model assessment stages—to elucidate the factors that shape stickleback occurrence. The overall trend observed is that factors related to habitat accessibility—the colonization-related factors—override any of the factors related to population extinction.

The relative importance of extinction- versus colonization-related factors in shaping organism (including fish) distribution has been investigated in a variety of studies previously (Conroy *et al*., 1999; Magnuson *et al.*, 1998; Beisner *et al*., 2006; Bertolo & Magnan, 2006; Hershey *et al*., 2006). Beisner *et al*. (2006) illustrated for lakes the relative importance of extinction and colonization factors across varying motilities of subject organism, with the distribution of less vagile species being influenced to a larger degree by physical barriers to dispersal. Over the range of species size they considered—from phytoplankton to fish—fish were on the low vagility end, and thus most limited by isolation factors. Bertolo & Magnan (2006) stressed the importance of lake elevation and longitude, in shaping fish distribution in Canadian Shield lakes, and note that these factors reflect the extent to which fish could colonize lakes historically. Van Zyll de Jong *et al*. (2005) related the distribution of fishes at a regional scale in Newfoundland in part to the ability of fishes to access lakes. However, they noted that individual differences in lake environmental character become of greater importance at smaller spatial scales.

Attributing stickleback presence to individual predictors was not possible in this study due to collinearity between the predictors. An example of the problem of collinearity amongst factors is illustrated by lake elevation and linear distance to saltwater (Pearson correlation= 0.64). Distance to saltwater (and the closely related length of outlet stream), is apparently quite useful in explaining stickleback distribution. However, a mechanism by which this factor might limit stickleback range is unclear. For instance, if a very slow, low-gradient stream stretched inland for a long distance, it seems that stickleback could negotiate such a distance and colonize nearby habitats. Distance from saltwater may be an indirect factor; other factors, including elevation and outlet stream grade, are perhaps more direct indications of lake accessibility. Fransen *et al*. (2006) used elevation effectively to model fish distribution in streams of Washington state. However, they point out that even elevation is likely an indirect influence on distribution, instead proxying for a suite of factors including likelihood of migration barriers and stream temperature.

It appears that many lake habitats exist across the region that would be suitable for stickleback, but do not contain stickleback. Stickleback are not known as an environmentally sensitive species; this is reflected in the wide range of habitats stickleback are known to inhabit (Foster *et al*., 2003). Stickleback are apparently largely confined to low-lying, accessible habitats in Southeast Alaska, and excluded via migration barriers from other available, suitable habitats. Such a difference between the occupied habitats of a species and the total number of available habitats for that species can be referred to conceptually as the difference between the 'realized' and 'fundamental' niche of a species (Zaniewski *et al*., 2002; Brotons *et al*., 2004). Such a distinction in the literature has been made with the implication being that, given enough time for dispersal, the realized niche that a species actually occupies will approach that of the fundamental niche of all available habitats (Lehmann *et al*., 2002). Although this may be the case for threespine stickleback, the constraint of their dispersal to surface waters—apparently of moderate grade—makes it seem unlikely they will ever fully occupy all suitable habitats in the region.

The natural history of Southeast Alaska may assist in explaining the distribution of stickleback here (Hastings, 2005). The main colonization mode of stickleback is via surface waters (T. Reimchen, pers. comm.). Many of the habitats sampled in this study may not be presently, nor have been historically, accessible to colonization. The upper elevational extent of stickleback occurrence in the present study was 205 m above sea level. This coincides well with the maximum observed elevation of marine deposits in the immediate area (230 m) (Hastings, 2005) and suggests that stickleback may be limited more or less to this maximum. It is conceivable that stickleback could access and survive at much higher elevations. However, the high collinearity between lake elevation and maximum outlet stream gradient seen in the study lakes (0.85), suggests that lakes of high elevation with moderate, meandering outlet streams navigable by stickleback may be exceptional in this region. It is likely that outlet stream gradient conspires with elevation to limit stickleback populations largely to lower elevations.

1.4.2 Secondary effects

Presence of an outlet stream was related to stickleback occurrence, although outlet stream was not significant (p=0.07) by itself in a univariate model. Three lakes in the area of the Herbert River, north of the town of Juneau, lacked outlet streams and stickleback populations. Two of these three lakes are ‘intermorainal swale ponds’: locations of water impoundment between terminal moraines created ~140 years ago by little ice age advance of the Herbert Glacier (R. Carstensen, pers. comm.). One lake is a large (3.2 ha) kettle pond created roughly at the same time. All did not exist ~150 years ago, and all are relatively low-lying (elevation range: 50-100 m), but not in the floodplain of the Herbert River. These lakes can be contrasted with a fourth study lake that is located on the opposite side of the Herbert River at slightly lower elevation (in the floodplain), has an outlet stream, and contains stickleback. The Herbert River in the area of these lakes seems unlikely stickleback habitat, with cascades and high water velocities, and no stickleback were captured in very limited trapping performed in the Herbert River. Whether stickleback colonized this fourth lake via the Herbert River as it is situated today, or via marine waters at some historical point in time, is unknown. However, it seems because the three lakes in this area lacking outlet streams are apparently devoid of stickleback, the lack of an outlet stream may reduce the chances that a lake will harbor stickleback.

It is important to note that some of the other factors considered secondary effects here—presence of an inlet stream, lake surface area, mean slope around the lake, and others—might show significant primary relationships with stickleback occurrence upon more extensive sampling.

1.4.3 Summary

The distribution of threespine stickleback in small lakes represents the distribution of a prey resource for many other vertebrate (and invertebrate) consumers (Reimchen, 1994), and so in turn is likely related to the distribution of these consumers. A secondary objective of this study was to detect other fish species in these small lakes. Either cutthroat trout, Dolly Varden, or coho salmon co-occurred with stickleback in 18 of the 22 lakes that contained stickleback, and these salmonids were only found in 4 lakes not containing stickleback. The distribution of stickleback roughly approximates that of salmonid species directly consumed by man (for more discussion of this, see the 'general discussion and recommendations' section below). Stickleback likely influence salmonid populations in these shared habitats through competition and as a prey resource, though it is clear neither precludes the presence of the other.

Unique morphological populations of threespine stickleback occur in other post-glacial landscapes (Bell *et al*., 1993). These populations are of scientific and conservation concern from evolutionary biology and genetic conservation perspectives (Foster *et al*., 2003). Stickleback populations in Southeast Alaska have been largely unsampled, and so such potentially unique populations remain unidentified in this region. The identification of landscape correlates of stickleback populations in this study may help to guide future efforts to identify and potentially protect such populations of concern.

Chapter 2. Comparison of methods for modeling threespine stickleback occurrence in small lakes of Southeast Alaska

2.1 Introduction

Modeling species occurrence entails discrimination between habitat units in which a species occurs and in which it does not occur (Guisan & Zimmerman, 2000). An ideal species occurrence model does this perfectly, splitting multivariate space so that there is no overlap between presence and absence locations. Such an ideal model will never predict species occurrence probability to be greater at a site where a species does not occur than it will predict to a site where a species does occur (Fielding & Bell, 1997).

This ideal species occurrence model is rarely encountered. More common are models in which there is overlap in the environmental predictor combinations associated with species presence and absence. For these models, ‘false positives’ and ‘false negatives’ will be generated in the model output. Such errors correspond with model output probabilities greater at species absence locations than species presence locations (Fielding & Bell, 1997).

So, generally, models have error rates, and the object is to produce models that minimize both false positives and false negatives (Fielding & Bell, 1997). Minimizing one or the other is simple: produce a model that has an output of probability occurrence of 1 at all sites, or of 0 at all sites. Balancing the two types of errors is more difficult and is a main goal of species occurrence modeling (Fielding & Bell, 1997).

A further challenge consists of producing robust models; models that can be used to predict outside of the dataset on which they were first constructed. This ability is at least as important, if not more important, than model performance in predicting only to the data on which the model is based (Boyce *et al*., 2002). Very often, models are constructed with the objective of not merely explaining the original data, but predicting accurately to new, independent sites.

A number of methods exist for modeling species occurrence (Guisan & Zimmerman, 2000; Elith *et al*., 2006). In fact, the number of methods has increased greatly recently, due in part to the increased speed of the personal computer, which makes many of these methods feasible (Garzon *et al*., 2006). With this increase in available methods comes the difficulty of choosing amongst them.

There have been a number of efforts recently to compare the performance of different modeling methods side-by-side (Segurado & Araujo, 2004; Elith *et al*., 2006). These efforts have identified some strengths and weaknesses of the various methods. However, some of these method comparisons have produced equivocal results (Moisen & Frescino, 2005; Pearson *et al*., 2006), and some authors have suggested not choosing a single method, but instead routinely using multiple methods to analyze the same dataset

(Thuiller, 2003). Alternatively, others have suggested that the modeling method should be chosen on a case-by-case basis, depending on the characteristics of the specific dataset being analyzed (Segurado & Araujo, 2004; Pearson *et al*., 2006).

In this study, threespine stickleback occurrence in small lakes of Southeast Alaska was modeled with mostly landscape-level (GIS) variables. A calibration set of lakes (n=36) in the northern part of the region was used for building models using four methodologies widely used for species presence modeling: the generalized linear model (GLM), generalized additive model (GAM), classification tree (CT), and artificial neural network (ANN). For each modeling method, a ‘simple’, ‘intermediate’, and ‘complex’ model was constructed. Model stability and predictive capability were then assessed via leave-one-out cross validation, and prediction to a geographically separate set of lakes (n=18) in the southern part of the region. The performances of the four modeling methodologies, and 3 levels of model complexity, were then compared for each of three—calibration, cross validation, and prediction—stages of model assessment.

2.2 Methods

Before model construction, all variables showing collinearity (Pearson correlation coefficient > 0.70) in pairwise tests were examined. For pairs of variables showing collinearity > 0.70, the variable retained either made more sense as a plausible correlate of stickleback occurrence or was more readily available for all lakes in the study region, and therefore more useful for predictive purposes. All remaining variables were then tested for significance in univariate GLM's, and retained only if they significantly reduced deviance at the $p<0.25$ level (Hosmer & Lemeshow, 2000). All statistical analyses were performed using 'R' (R Development Core Team, 2006).

For each of the four modeling methods—GLM, GAM, CT, and ANN, a 'simple', 'intermediate', and 'complex' model were constructed. These three levels of model complexity correspond with three model sizes—two factors, three factors, and five factors were included in models of each respective complexity level. For CT models, there is a variation from these model sizes that is discussed in the CT section below.

2.2.1 Simple models

The simple (two-factor) model for each method included lake elevation and linear distance to saltwater. These two factors are the most readily available across the landscape of Southeast Alaska, as the first already exists as an attribute in the USFS 'lakes' coverage (USFS, 2003a) and the second is obtainable through relatively simple GIS processing. These two factors have the added benefit of being available at all points on the landscape irrespective of whether a lake is mapped at that point or not. This is convenient because not all lakes in the size range investigated in this study are mapped. So these two factors can be used to predict probabilities of stickleback occurrence for small lakes throughout the landscape, even if they are not mapped.

Additionally, lake elevation and linear distance to shoreline are two of the most useful explanatory variables in the pool of factors considered. When used individually in univariate GLMs, these two factors are the two out of the total pool of factors that explain the largest proportion of deviance (0.29 and 0.26, respectively). A bivariate model including lake elevation and linear distance to saltwater ranks sixth out of all possible combinations of factors in bivariate GLMs (91 total possible). And both of these factors appear in a three-factor model selected by stepwise forward selection using AICc (AIC with small sample correction) (Burnham & Anderson, 2002).

2.2.2 Intermediate models

The intermediate (three-factor) model for each method included those factors in a GLM chosen by a stepwise forward procedure using AICc: lake elevation, linear distance to saltwater, and presence of a lake outlet stream. So the intermediate model is the simple model with presence of lake outlet stream added as a third factor.

2.2.3 Complex models

The complex (five-factor) model for each method included all those factors that showed deviance reduction in univariate GLM's significant at $p<0.25$ (Hosmer & Lemeshow, 2000): 1) lake elevation, 2) linear distance to saltwater, 3) presence of lake outlet stream, 4) mean slope of lake outlet stream, and 5) lake surface area.

2.2.4 Generalized linear models (GLM)

GLM are characterized by a link function which allows modeling of non-normal response variables such as presence/absence data. The GLM used here—and commonly with presence/absence data—is logistic regression, which uses the logit link to model the probability of presence:

$P = e^u / (1 + e^u)$, where *u* takes the form of the more familiar linear regression equation—

$u = A + b_1 X_1 + b_2 X + ... + b_n X_{nv}$

where *A* = regression constant, b_n= regression coefficients, and X_m= independent variables. It is not required that predictor variables be normally distributed, linearly related, or of equal variance (Porter *et al*., 2000; Tabachnick & Fidell, 2000). A potential drawback of GLM is the necessary a priori assumption of linearity or—upon predictor transformation—some other type of parametric response. This results in limited model flexibility in contrast to the other data-driven, non-parametric methods used in this study (Segurado & Araujo, 2004).

2.2.5 Generalized additive models (GAM)

GAM (Hastie & Tibshirani, 1990) have been used extensively in modeling species distribution. Generalized additive models are known for their flexibility and data-driven nature, which allow them to model complex ecological relationships (Granadeiro *et al*., 2004). In GAM, the smoothed form of one factor at a time is estimated via maximum likelihood, according to the following relationship (Hastie and Tibshirani, 1990):

$$E[y] = g^{-1}\left(\beta_0 + \sum_k S_k(x_k)\right)$$

The residual deviance is iteratively explained with subsequent significant factors, with model error decreasing upon each subsequent smooth. GAM have compared favorably to GLM in a number of species distribution studies (Franklin, 1998; Pearce & Ferrier, 2000; Thuiller, 2003). GAM strength lie in their ability to model highly non-linear and non-monotonic relationships. A crucial step in using GAM is choosing smoothers that result in a reasonable number of effective degrees of freedom (Thuiller *et al.*, 2003). The ‘mgcv’ library of R (Wood, 2006) automates the degrees of freedom associated with each parameter via a cross-validation routine in which the ‘un-biased risk estimator’ (UBRE) criterion is minimized by balance between gains from reduced deviance and the estimated degrees of freedom for the entire model, serving the same purpose as an AIC score . The smooth of each variable included in models for this study was approximated with a cubic spline smoother.

2.2.6 Classification trees (CT)

The work of Breiman (1984) paved the way for the modern increase in the use of CT for modeling species presence (De’ath, 2002; Thuiller *et al.*, 2003; Hershey *et al.*, 2006). CT use a recursive partitioning approach to the data, with each subsequent split causing the maximum possible decrease in the total deviance (impurity) of the tree (De'ath & Fabricius, 2000). The final tree represents a balance between homogeneity within each 'node' of the tree and complexity of the entire tree (Figure 7):

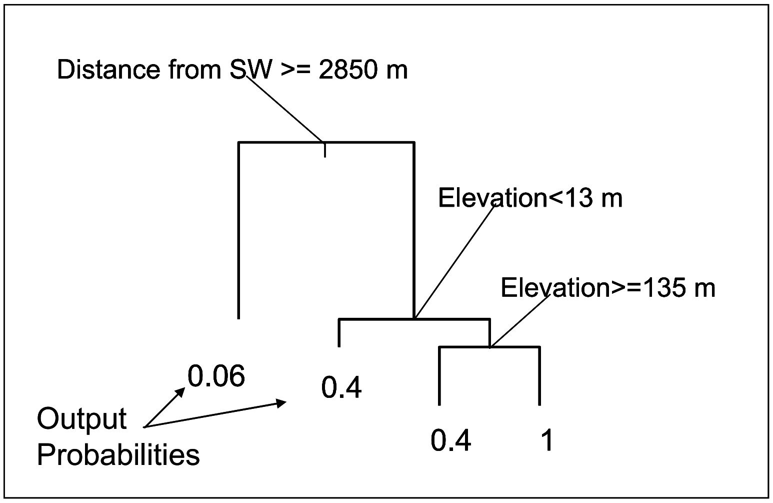


Figure 7. Example of classification tree used in the present study.

CT allow for complex interactions amongst predictors, are easy to conceptualize, have no distributional assumptions, are robust to the presence of outliers, and handle categorical predictors well (Vayssieres *et al.*, 2000; Turgeon & Rodriguez, 2005). They have been used with success in modeling stream and lake fish distributions (Magnuson *et al.*, 1998; Turgeon & Rodriguez, 2005; Hershey *et al.*, 2006). The CT used here was constructed using the 'rpart' library of R (Therneau & Atkinson, 2006), which follows the work of Breiman *et al.* (1984) closely.

The factors used in the intermediate and complex CT models constructed for this study differed from those factors used with the other modeling methods. The reason for this is that the rpart CT algorithm, when given a choice of any of the 5 factors made available to the algorithm to split the data, continually split the data using lake elevation and linear distance to saltwater, rather than using any of the three remaining factors. For this reason, the intermediate and complex CT models used here are built only with the 'elevation' and 'linear distance' factors, and other model parameters were varied to add model complexity. The parameters varied in the R function 'rpart' to create more complex CT models were the CP, minbucket, and minsplit parameters (Therneau & Atkinson, 2006).

2.2.7 Artificial neural networks (ANN)

Artificial neural networks have as their basis an algorithm designed to mimic mammalian brain function (Bishop, 1995). ANN are increasingly being used to model species occurrence (Lusk *et al*., 2002; Olden, 2003; Olden *et al*., 2006). ANN consist of layers comprised of neurons. In species occurrence modeling, the number of layers is typically 3: an input layer, a 'hidden' layer, and an output layer (Figure 8). The number of neurons in the input layer corresponds to the number of environmental predictors used in the model. The number of neurons in the hidden layer varies but is

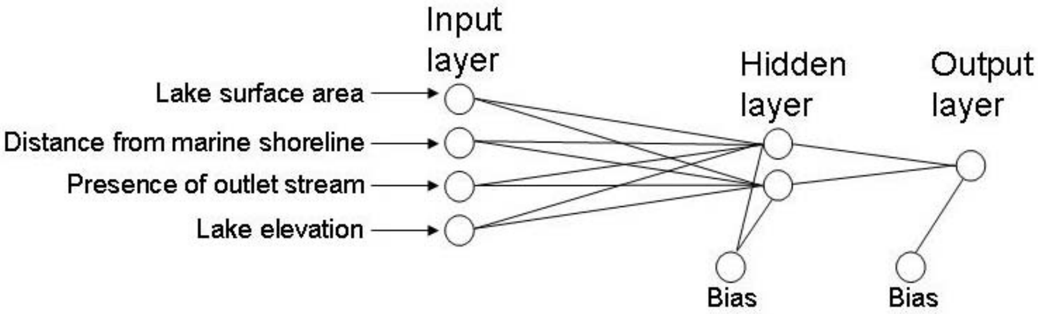


Figure 8. Architecture of artificial neural network (ANN) models used in this study.

usually optimized by cross-validating ANNs built with a range of neurons in the hidden layer. The number of neurons in the output layer also varies, but for species occurrence modeling, one neuron is often used (Olden & Jackson, 2002). Each ANN model run produces a slightly different result due to the heuristic nature of ANN. Therefore, ANN solutions were calculated 10 times for each model and the mean used for predictive purposes (Segurado & Araujo, 2004).

2.2.8 Model evaluation

The area under the receiver-operating characteristic curve (AUC) (Fielding & Bell, 1997; Elith & Burgman, 2002), was used to evaluate all models. AUC is a measure of how well a model discriminates between presence and absence records (Fielding & Bell, 1997). A model's AUC score is the rate of true positives to false negatives that a model yields across the entire range of possible threshold values between 0 and 1. This is why AUC is considered threshold independent and is considered effective in relation to other model evaluation metrics (Liu *et al*., 2005). The AUC score was used to compare model (and thus individual predictor) performance in cross-validation and prediction as well. The AUC value can be used to compare model performance in all three stages—calibration, cross-validation, and prediction—of model assessment.

2.2.9 Geographic prediction

Simple (two-factor) models were used to predict stickleback occurrence across Revillagigedo Island using the following steps in ArcGIS and 'R':

1) a GIS raster file containing SRTM-DEM derived elevations for all 30X30m pixels across the island was converted to a point feature shapefile.

2) a GIS raster file was produced using the 'euclidean distance' tool (Spatial Analyst toolbox) with source set as the marine shoreline (from the USFS 'shore' GIS coverage). This raster was converted to a point feature shapefile.

3) The two shapefiles were spatially joined, and the attribute table was imported into the 'R' environment.

4) For 'simple' models from each modeling method, the predict function for each method was used, with the target prediction object being the GIS-output 'elevation' and 'distance from shoreline' values.

5) The output probabilities were then exported from 'R' to a .dbf file, brought back into ArcGIS, joined (by a unique identifier) to the 'elevation' point shapefile, and reconverted to a raster dataset.

2.2.10 Variable importance

For all models and complexity levels, relative variable importance was assessed by comparing the AUC value of the full model with that of the model with each factor excluded in turn. This was performed for the calibration, cross-validation, and prediction stage. The comparison value is expressed as ($AUC_{full\ model}$- $AUC_{reduced\ model}$)/ $AUC_{full\ model}$.

2.3 Results

The differences in performance of the four modeling methods used here were generally not clear-cut, but there were a number of notable trends in model performance (Figure 9). Every modeling method and level of complexity predicted stickleback occurrence better than random chance. All models had lower $AUC_{xvalidation}$ (mean 0.81) than $AUC_{calibration}$ (mean 0.94) across all modeling methods and model complexity levels. This was expected, as rather than the models just predicting to the data upon which they were constructed, models were predicting to the left-out record in this stage. Surprisingly, mean $AUC_{prediction}$ nearly equaled $AUC_{calibration}$ across all models (0.92 and 0.94 respectively). The predictive performance of all models on the test set of lakes was higher than expected. AUC values of >0.9 are subjectively considered very useful (Fielding &

Bell, 1997). More details on the relative performance of the four modeling methods follow.

2.3.1 Generalized linear models (GLM)

The GLM models of all three complexity levels tended to show stability upon cross-validation (Figure 8), with $AUC_{calibration}$- $AUC_{x\text{-}validation}$ less (~0.08) than the other three modeling methods. Unlike the other modeling methods, intermediate and complex GLM models showed better performance than the simple GLM model on cross-validation. Complex GLM model AUC value on prediction to the test set was higher than for the other complex models: $AUC_{GLM} > AUC_{CT} > AUC_{GAM} > AUC_{ANN}$.

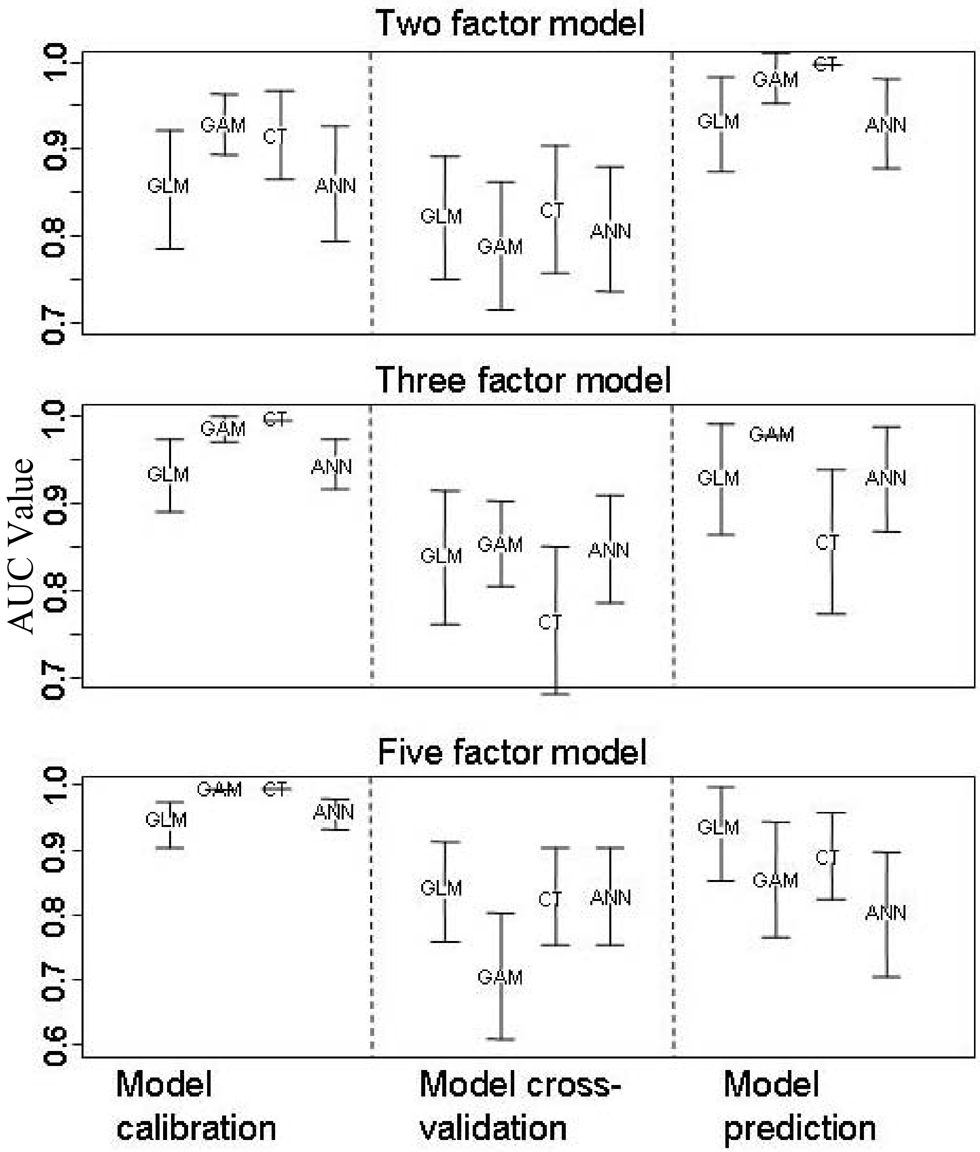


Figure 9. Performance (AUC value) of the four modeling methods and three model complexity levels on calibration, cross-validation, and prediction to the test set.

2.3.2 Generalized additive models (GAM)

GAM models—especially the intermediate and complex levels—tracked the original data more closely than GLM models, performing similarly in the calibration stage to the CT models, and discriminating stickleback occurrence lakes at a higher rate than both GLM and ANN in calibration. The simple and complex GAM models had AUC_{xval} lower than that of the other methods, while the intermediate GAM model AUC_{xval} was equal to that of ANN (0.85) and greater than the GLM and CT intermediate models. Mean GAM performance dropped off most between calibration and cross-validation, ($AUC_{calibration}$-AUC_{xval}) (mean 0.18). The complex GAM model had the lowest $AUC_{prediction}$ value (0.71) of any model in any stage of assessment, reflecting a tendency of the flexible GAM method to overfit the calibration data.

2.3.3 Classification trees (CT)

CT models tracked the calibration data closely, and then dropped off in performance somewhat on cross-validation. The CT simple model had AUC=1.0 on prediction to the test set, perfectly discriminating between occurrence and non-detection sites. However, GAM intermediate AUC for prediction to the test set was lower (0.86) than the other methodologies and complex showed performance intermediate between that of the other complex models.

2.3.4 Artificial neural networks (ANN)

ANN simple and intermediate models predicted to the test set with equal accuracy (AUC = 0.93) and with higher accuracy than the complex model (AUC=0.80). ANN models. Mean AUC_{xval} for ANN models of all complexity was equal to that of GLM and higher than both CT and GAM.

2.3.5 Model complexity

The relative performance of the three levels of model complexity (mean across all modeling methods) followed this pattern:

Calibration: $AUC_{complex}$ (0.98) > $AUC_{intermediate}$ (0.97) > AUC_{simple} (0.88)
Cross Validation: $AUC_{intermediate}$ (0.83) > AUC_{simple} (0.81) > $AUC_{complex}$ (0.80)
Prediction: AUC_{simple} (0.96) > $AUC_{intermediate}$ (0.92) > $AUC_{complex}$ (0.87)

Complex models performed best on calibration, but their performance dropped off on cross-validation and prediction. The trend with simple models was generally the opposite, they did not track the original data closely but were relatively robust on cross-validation and prediction. GLMs were an exception to this pattern, the most complex GLM outperforming the other complex models on prediction with AUC (0.93), which was equal to the AUC of GLM simple and intermediate models.

2.3.6 Variable importance

The importance of the five variables included in models as expressed by ($AUC_{full\ model}$-$AUC_{reduced\ model}$)/ $AUC_{full\ model}$ indicated the relatively high contribution of linear distance to saltwater and lake elevation in explaining stickleback distribution (Figure 10-12). Lake area, presence of lake outlet stream, and mean slope of outlet stream, when excluded from complex models, actually increased model prediction accuracy to the test set, as well as in cross-validation in some instances (Figure 12).

Modeling methods differed with respect to the relative contribution of each factor to model performance. GLM predicted most accurately to the test set of lakes when all five factors were included in the model, while exclusion of lake area, presence of lake outlet stream, or mean slope of outlet stream from ANN increased AUC value on prediction to the test set (Figure 12). In the calibration stage, exclusion of linear distance from saltwater had little effect on the AUC value of the CT simple model, but a greater effect on each of the other simple models (Figure 10). In contrast, lake elevation was much more important to simple model performance in the prediction stage of model performance (Figure 12).

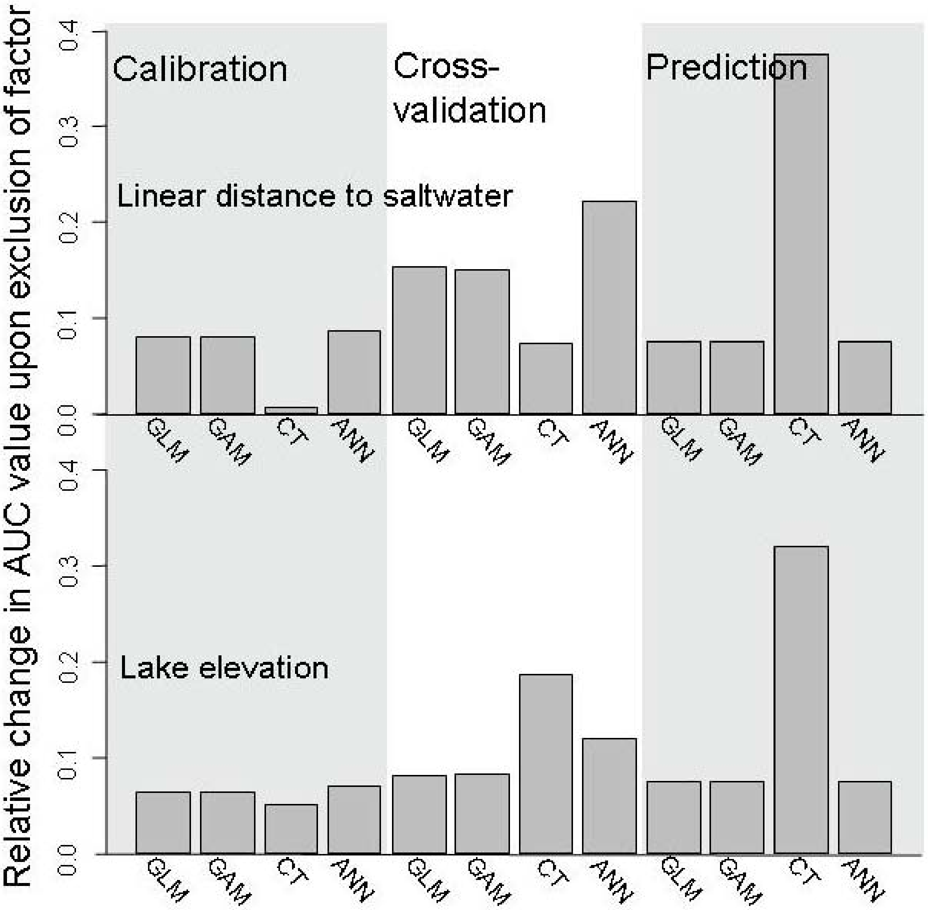


Figure 10. Relative change in AUC Value upon inclusion of each factor in simple models: (AUCfull model- AUCreduced model)/ AUCfull model for each stage of model performance—calibration, cross-validation, and prediction.

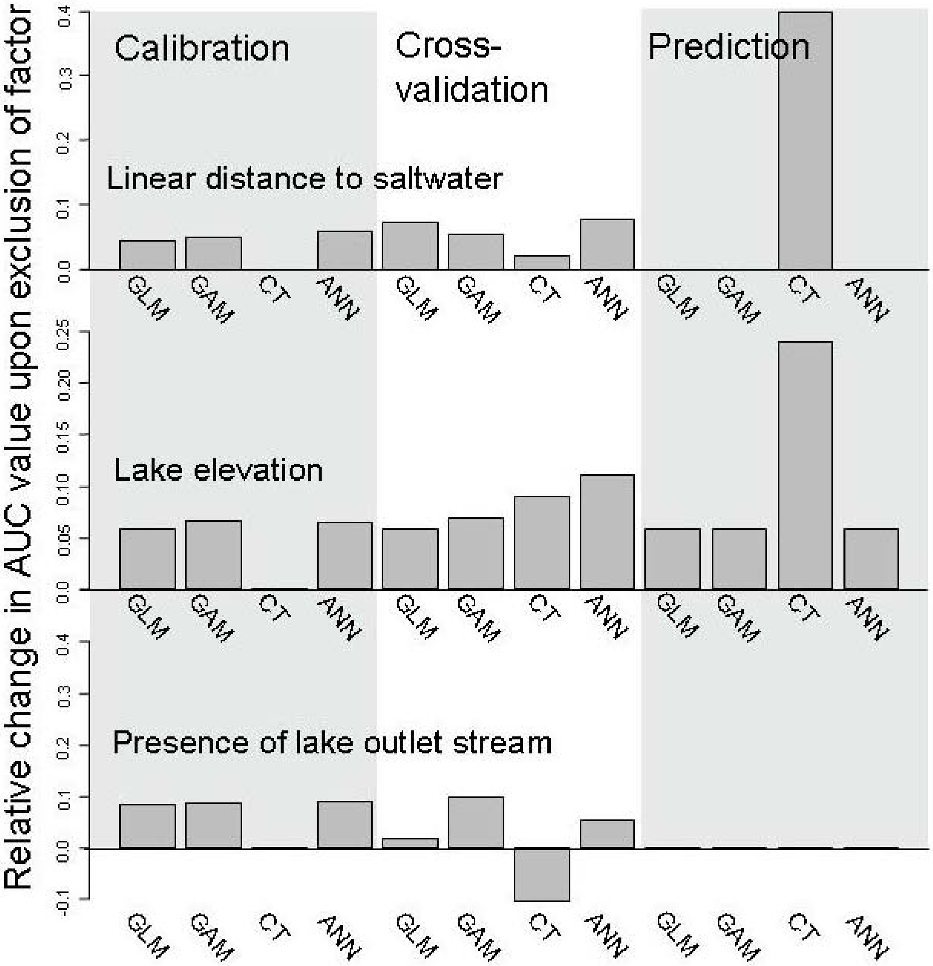


Figure 11. Relative change in AUC Value upon inclusion of each factor in intermediate models: ($AUC_{full\ model}$- $AUC_{reduced\ model}$)/ $AUC_{full\ model}$ for each stage of model performance—calibration, cross-validation, and prediction.

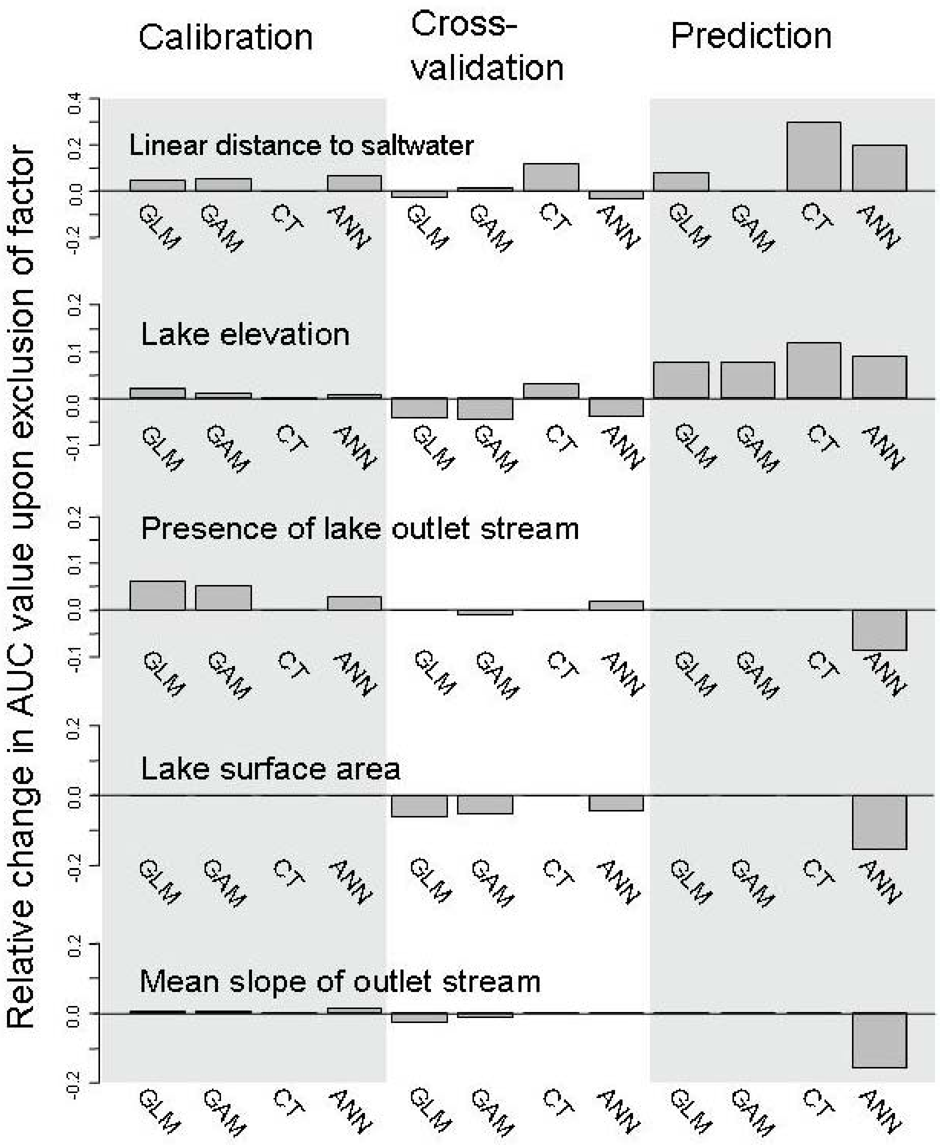


Figure 12. Relative change in AUC Value upon inclusion of each factor in complex models: ($AUC_{\text{full model}}$- $AUC_{\text{reduced model}}$)/ $AUC_{\text{full model}}$ for each stage of model performance—calibration, cross-validation, and prediction.

2.4 Discussion

All modeling methods explained and predicted stickleback presence fairly well, with AUC values for all methods, complexities, and model assessment stages exceeding 0.7. The parametric GLM modeling method is a conservative choice, as the *a priori* determination of response shape limits the extent to which GLMs overfit data (Vaughan & Ormerod, 2005). The other three methods are empirical approaches and more prone to overfitting, but this can be avoided in all methods by limiting the number of factors used in models (Vaughan & Ormerod, 2005). The GAM and CT simple models had the highest predictive performance of all methods and model complexities in prediction to the test set. ANN simple and intermediate models performed robustly on cross-validation and prediction; the complex ANN model fell off in predictive performance relative to the other modeling methods. Parsimony seems the key in successfully building robust models for the three non-parametric modeling methods used here.

Guisan & Zimmerman (2000) suggest that shrinkage rules can be used to limit the number of variables included in models, citing the m/10 rule (Harrell *et al*., 1996) as an upper limit, where m is the number of observations in the minority of a binary dataset. In the current study, m would be the number of lakes in the calibration dataset where stickleback were present divided by 10, 14/10= 1.4. Using this rule for the current dataset would result in models with maximum size of a single variable. While this rule seems a bit conservative, it does generally coincide with the finding that simple models appear appropriate, given the small sample size of the present study.

The results here generally coincide with the suggestion that added insight can be gained by using multiple modeling methods, but there may not be a clear choice of method that outperforms all others (Thuiller, 2003). Although GLM, GAM, CT, and ANN have all performed favorably in one application or another (Olden & Jackson 2001; Segurado & Araujo, 2003; Thuiller, 2003; Moisen & Frescino, 2005), it may be ill-conceived to pick a single method at the exclusion of others. Rather, a strategy of using multiple methods side-by-side and then being able to compare and contrast results has the benefit of multiple results that can support each other or provide alternate possibilities. Hosmer & Lemeshow (2000) suggest a subjective scale for assessing model performance based on AUC value, with >0.9 being excellent, 0.8-0.9 being good, 0.7-0.8 adequate, and 0.6-0.7 poor. The results from the current study generally fall into the good-excellent range, and fall within the range of AUC values found by previous studies (Table 4).

Output probabilities from the simple model of each of the four modeling methods showed the same general trends (Figure 13). The results of model performance across modeling methods and stages of model assessment indicate that simple models including two widely available landscape factors (lake elevation and linear distance to saltwater) perform robustly on cross-validation and prediction. The high performance of such simple models (across all methods), and the availability of these two factors across the region allows prediction to all points on the terrain of the southern prediction area. In doing this, the caveats exist that:

1) The eastern portion of Revillagigedo Island was not sampled due to land status that prohibited lake access, so predictions to that area of the island are, strictly speaking, outside of the range of model inference.

2) AUC scores on prediction were high, suggesting model validity. However the prediction set of lakes was small (n=18) with only four lakes in which stickleback occurred. It is possible that models would not predict as accurately given a larger sample of stickleback occurrence records from lakes on Revillagigedo Island.

3) The 'lake elevation' and 'linear distance to saltwater' factors allow prediction across all points on the terrain of Revillagigedo Island. This, obviously, does not indicate stickleback occurrence at all points on the terrain, only the probability of occurrence given the existence of a small lake.

4) These probabilities are only predictions from the northern calibration lakes to the southern study area. The models performed well on prediction to the southern set of lakes (AUC range 0.93-1.0), but further testing is required to verify model predictive accuracy.

Table 4. Literature AUC values for species occurrence modeling.

	Organism(s)	Modeling	Model	
Author, year	modeled	method	performance stage	Mean AUC (range)
Dobrowski *et al*., 2006	Plants (19 spp.)	GAM	x-validation (5-fold)*	0.74(0.58-0.84)
Jensen *et al*., 2005	Blue crab	GAM	calibration	0.85(0.81-0.91)
			x-validation (interannual)[†]	0.71(0.44-0.86)
Luoto *et al*., 2006	Butterflies (98 spp.)	GAM	x-validation (4-fold)*	0.79(0.48-0.99)
Suarez-Seaone *et al*., 2002	Birds (3 spp.)	GAM	calibration	0.92(0.90-0.96)
			x-validation (10-fold)*	0.91(0.88-0.95)
Brotons *et al*., 2004	Birds (30 spp.)	GLM	calibration	0.87(0.70-1.0)
			prediction (70%-30%)[‡]	0.82(0.57-0.94)
Pearson *et al*., 2006	Proteaceae (4 spp.)	ANN	prediction (70%-30%)[‡]	0.95(0.89-0.99)
		CT	prediction (70%-30%)[‡]	0.90(0.86-0.95)
		GAM	prediction (70%-30%)[‡]	0.96(0.92-0.99)
		GLM	prediction (70%-30%)[‡]	0.96(0.92-0.99)
Thuiller, 2003	Trees (61 spp.)	GLM	calibration	0.95(0.82-0.99)
			prediction (70%-30%)[‡]	0.94(0.82-1.0)
			prediction (independent set)	0.95(0.82-0.99)
`		GAM	calibration	0.96(0.84-0.99)
			prediction (70%-30%)[‡]	0.94(0.84-1.0)
			prediction (independent set)	0.95(0.84-0.99)
		CT	calibration	0.94(0.83-0.99)
			prediction (70%-30%)[‡]	0.87(0.54-0.96)
			prediction (independent set)	0.92(0.82-0.98)
		ANN	calibration	0.97(0.87-1.0)
			prediction (70%-30%)[‡]	0.95(0.82-0.99)
			prediction (independent set)	0.96(0.85-0.99)

*Cross validation performed as noted, in contrast with the leave-one-out (n-fold) cross-validation performed in the present study.

[†] Cross validation performed from one model calibration year to other model testing years.

[‡] Model built with 70% of original records and prediction made to the 30% remaining records.

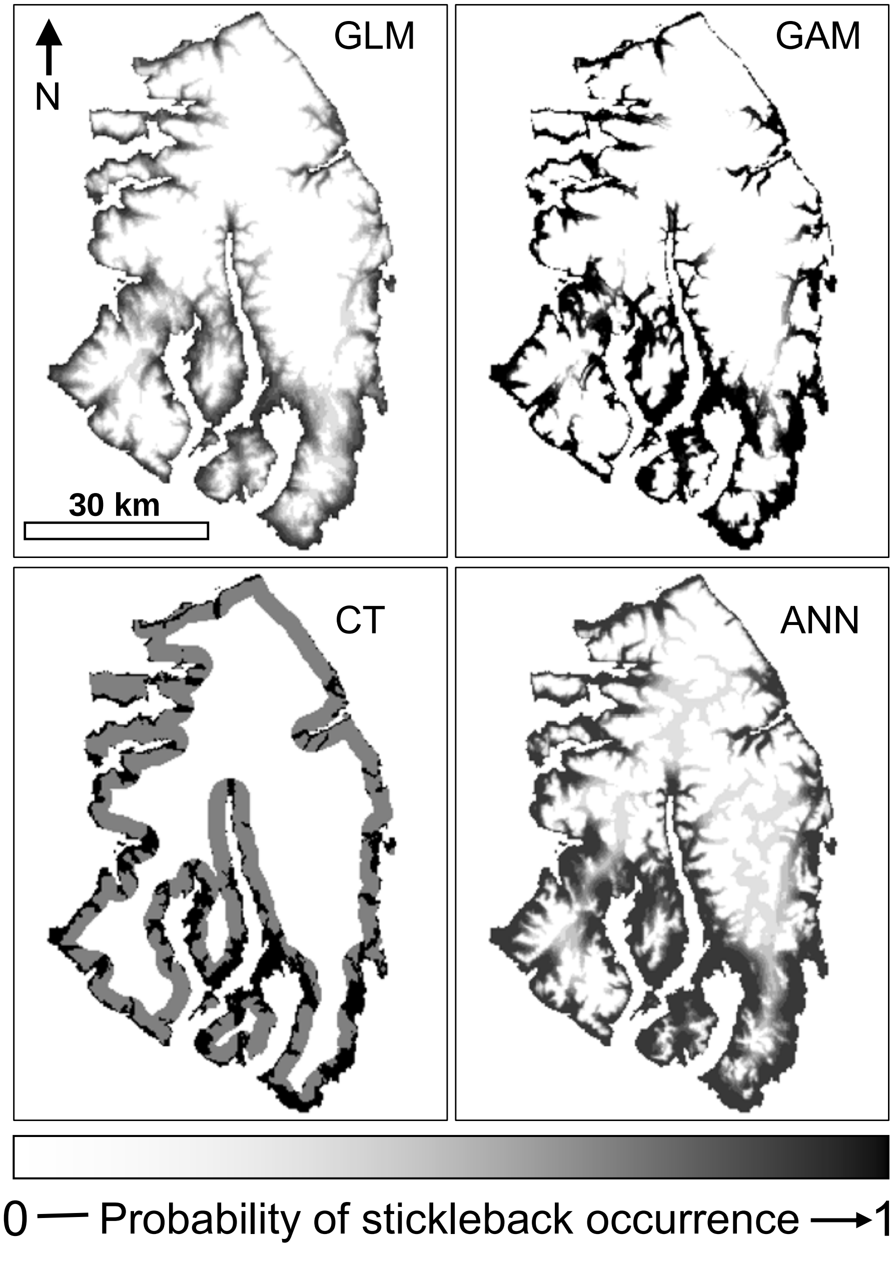


Figure 13. Predicted probabilities of stickleback occurrence from the simple model (two factor model with lake elevation and distance from saltwater included) of each modeling method applied to the landscape of Revillagigedo Island.

Each modeling methodology differed slightly in the pattern of predicted distributions (Figure 13). For instance, ANN, and to a lesser extent GLM, predicted probabilities remain at levels above zero for a considerable distance inland. GAM and CT, however, predict a sharp drop-off in probability so that both methods predicted stickleback occurrence probability ~0.0 anywhere roughly 10 km or more from saltwater. Of course, it is impossible to tease apart the effect of elevation and linear distance, for the two factors show collinearity (Pearson correlation coefficient= 0.64). But the utility of these simple models is evident, at least for predicting occurrence probability over large spatial scales.

CT, and to a lesser extent GAM, produced models with output probabilities typically either very low ($0<P<0.05$) or very high ($0.95<P<1.0$) on cross-validation and prediction (see output probability graphs, Appendix A, Figures 1-3). Although both methods had generally high AUC values on cross-validation and prediction, such dichotomization of response may not be a realistic representation of how stickleback occurrence probabilities vary across environmental gradients. GLM and ANN models yielded results that may be more realistic, showing a gradual trend between low- and high-probability lakes. It may be that providing the CT and GAM methods with a larger dataset would allow both these methods to build more complex response curves (GAM) or trees (CT). The performance of the four different methods used here in relation to sample size is not clear. Hernandez *et al*. (2006) compared the performance of four different species distribution modeling methods over a range of sampling sizes, and found differences across methods, although none of the methods they evaluated were used in the present study. Published information comparing the methods used here across sample sizes appears to be lacking.

Another characteristic of both CT and GAM predictions to the test set (Appendix A, Figure 3) is their tendency to produce a number of false positives and no false negatives. In all lakes in the prediction set that were assigned low probability by these two methods, stickleback were not detected. However CT and GAM did not perform as well when they predicted a probability of occurrence of ~1.0. When probabilities of near 1.0 were predicted by these methods, in approximately 50% of lakes stickleback were actually absent. The exception is the relatively high performance of the CT simple model, which in fact had the highest performance of any model in prediction to the test set.

It appears all modeling methods performed adequately in this study, particularly when the number of factors was limited. It might be that with a larger sample size, an increased number of environmental predictors would significantly improve model fit and predictive capability. The small sample size used here does not support a large number of model parameters, and a parsimonious modeling strategy is most appropriate, particularly when using the CT and GAM methods.

2.4.1 Summary

Landscape factors were used here to effectively model the occurrence of threespine stickleback. Such widely available attributes such as lake elevation and lake distance from saltwater can be very useful in modeling and predicting broad-scale patterns of distribution. The successful use of these attributes for modeling stickleback occurrence encourages further use of such simple landscape models for modeling the distribution of other fish species in this region.

The choice of method for modeling fish distribution can be a difficult one. In this study, simple empirical models (GAM, CT, and ANN) with two environmental factors included were successfully used. Three- and five-factor GLM were also effective in all stages of model assessment. Future fish occurrence studies in Southeast Alaska might benefit from larger sample sizes, which may support more complex empirical models.

Chapter 3. General Discussion and Recommendations

3.2 General Discussion

Stickleback absence in some low-lying lakes close to saltwater represented variance for which models did not account. Water temperature is perhaps a missing covariate that might be beneficial to include in future modeling efforts. One small lake in the Taku River floodplain, but at the toe of a steep side slope, had cold water temperature (8.0ºC) at the time of sampling relative to all other low-lying lakes sampled (a number of alpine lakes had lower water temperatures). The water source of this lake appears to come from a steep, adjacent alluvial fan and associated groundwater upwell. In addition, this particular lake was quite small, so apparently water did not have a long residence time in the lake in which to warm. The possibility exists that stickleback avoid cold waters when possible.

Stickleback were not detected in two other small lakes in the northern study area that were at low elevation and in proximity to saltwater. These lakes were shallow, and neither had an outlet stream. Lake depth was not a significant correlate of stickleback occurrence in this study, but that does not rule out the possibility that it may influence occurrence patterns, as it does for lake trout (Hershey *et al*., 2006), and arctic char (Van Zyll de Jong *et al*., 2005; Hershey *et al*., 2006) in other high latitude lakes. Although water temperature was not extreme at either lake when taken at the time of visit (16ºand

17ºC respectively), a smaller pond in the area reached at least 27ºC (S. Pyare, pers. comm.) during the summer of 2005. While it is not known whether 27ºC is within the thermal limits of threespine stickleback, the possibility of an upper thermal limit for stickleback exists. Both warm and cold water temperatures can potentially affect fish species distribution in freshwater. Swales (2006) found the upper temperature of North American lakes and reservoirs harboring rainbow trout populations (*Onchorynchus mykiss*) to be around 21ºC, while the influence of water temperature on stream fish distribution has been demonstrated for a number of fish species (Dunham *et al*., 2003; de la Hoz Franco & Budy, 2005; Smith & Kraft, 2005). Temperature data taken during site visits in the present study did not reveal any correlation with stickleback presence probability (unpublished data), but it may be that measurements resulting from more intensive lake temperature sampling would show some relationship with stickleback presence.

Winterkill also presents a possible limit on habitat suitability for fish in lakes (Rahel, 1984). Causes can include anoxia (Rahel, 1984, Jackson *et al*., 2001) and the possibility (especially in small lakes at high latitudes) of lakes freezing solid (Hershey *et al*., 2006). In northern Sweden, isolated lakes that experience anoxic conditions in winter are relatively species poor (Ohman *et al* 2006), and Paszkowski & Tonn (2000) suggest that wintertime anoxia is a filter on fish communities in northern Alberta. Factors that potentially mitigate the effects of winter anoxia relate to availability of refuge, either via deep water (Jackson *et al*., 2001) or inlet streams (Tonn & Magnuson, 1982). Assessing the extent of anoxic lake conditions in southeast Alaska—and the possible landscape-level covariates of dissolved oxygen levels—would likely aid in explaining fish distribution patterns in the region. Any such assessment would greatly benefit from intensive DO sampling through time, and such sampling was not possible in the present study.

Previous studies on lake (Olden & Jackson, 2001; Hershey *et al*., 2006) and stream (Rieman & McIntyre, 1995; Porter *et al*., 2000) fish distribution have shown the utility of landscape factors for predicting occurrence. The present study confirms the usefulness of such factors in prediction, with most of the important predictors in this dataset (lake elevation, lake area, and outlet stream length) being large-scale landscape factors. Field-measured factors (presence of outlet stream being the exception) generally did not explain stickleback occurrence well. Concordance with the literature was found in the importance of lake surface area (Rago & Wiener, 1986), elevation (Hershey *et al*., 2006), and lake connectivity (Magnuson *et al*., 1998) for explaining lake fish distribution. The lack of significance of lake depth seen here diverges from some published findings (Jackson *et al*., 2001; Hershey *et al*., 2006).

Stickleback were not found at an elevation of greater than 205 m in this study. They are known to occur at an elevation of 215 m in Upper Slate Lake (E. Kline, pers. comm.), which is near the northern, calibration set of lakes. The maximum elevation at which a natural (unstocked) lake fish population is known to occur in Southeast Alaska is in

Lower Texas Lake in the far southern portion of the region at 710 m (Baade, 1961). There is no definite barrier to fish passage between Lower Texas Lake and the saltwater, at least for the anadromous coho salmon documented there. There is a high-elevation (410 m) lake on Baranof Island in northern Southeast Alaska which contains a documented population of resident sculpins. The outlet stream from this lake is of extremely high gradient and is likely to represent a barrier to fish migration. Such high-elevation fish populations appear to be anomalous; the likelihood of stickleback and other fish species presence in lakes of the region seems to have a strong negative correlation with elevation.

The relatively high importance of elevation as a predictor of fish distribution found here concurs with a number of previous studies of both stream and lake fishes. Elevation is significantly associated with the distribution of small-bodied fishes (including nine-spine and brook stickleback), in Canadian Shield lakes (Bertolo & Magnan, 2006). These authors also found elevation significant in explaining the occurrence of five other large-bodied fish species. Hershey *et al*. (2006) found lake elevation of varying importance in explaining the presence of five fish species in northern Alaska—ranging from not explanatory for arctic grayling, to very important in explaining arctic char presence. Although in studies by Hershey *et al*. (2006), Magnuson *et al*. (1998) and the present study, lake elevation is classified as a colonization-related variable, it is worthy of note that many factors—including ionic composition (D’Arcy and Carignan, 1997; Kratz *et al*., 1997), temperature (Cavalli *et al*., 1997; Edmundson & Mazumder, 2002), and productivity (Magnuson *et al.*, 1998)—may covary with elevation. Thus, it is impossible in such field studies to put a finger on a single, proximate cause of species presence or absence.

Southeast Alaska has experienced dramatic post-glacial rebound (Hastings, 2005). As a result, the potential exists that basins which historically were inundated with marine waters now exist as lakes far above saltwater. The phenomenon of historic fjords uplifting to become isolated lakes has been observed in post-glacial landscapes elsewhere—including in eastern Canada (Girard and Angers, 2006), northwest Russia (Snyder *et al.*, 1997; Corner *et al.*, 1999), southern Sweden (Sandgren and Snowball, 2001), southern Finland (Sarmaja-Korjonen and Hyvarinen, 2002), and central Norway (Solem *et al.*, 1997). Such historic indundation and subsequent lake isolation is thought to have greatly influenced fish distributions in postglacial regions. For salt-water tolerant species such as the torrent sculpin *Cottus gobio*, anadromous salmonids, and the ninespine stickleback (*Pungitius pungitius*), historic marine inundation may have increased the number of habitats colonizable (Power *et al.*, 1973; Scott and Crossman, 1973; Kontula and Vainola, 2004). However, in other instances, such as for the saltwater-intolerant longnose dace in eastern Canada, historic low-lying seas may have been a barrier to colonization (Girard and Angers, 2006).

The maximum known elevation of marine deposits in southeast Alaska of 230 m (Hastings, 2005) may roughly indicate an elevational cutoff for natural fish populations, as fish access to such high elevation lakes may be compromised by high-gradient outlet streams. Exceptions might include lakes such as the above-mentioned Lower Texas Lake, which is situated at high elevation, but has a moderate-gradient outlet allowing fish populations to colonize. The high collinearity between lake elevation and maximum outlet stream gradient seen in the lakes of the present study (0.85), suggests that such a situation may be uncommon in Southeast Alaska given the abrupt terrain. Elevation appears to covary with a number of factors that may influence stickleback distribution in the region.

Many of the small lakes in this study show evidence of current or past beaver damming at the outlet of the lake. The presence of a beaver dam at a lake outlet indicates that perhaps at times in the past, not nearly as much water was impounded in the lake basin. At the extreme, it is possible that some of these small lakes owe their existence to beaver dams, without which they might not impound water. Therefore, even if the current position of a lake is below the possible elevation limit of historic marine inundation, this historic period of potential colonization may predate the age of the lake itself.

Small lakes can be considered ephemeral landscape features due to sediment and organic matter deposition in their basins (Kalff, 2002). An important facet of stickleback distribution in lakes is not only where populations persist presently, but which habitats will persist longest into the future. A small lake with an inlet stream(s) actively transporting sediment may have a life expectancy less than that of a large, deep lake that does not experience a large input of sediment (Wetzel, 2001; Kalff, 2004). Such habitat changes through time will likely reshape stickleback distribution in small lakes from its present state.

The results of this study suggest the utility of landscape models for successfully predicting stickleback occurrence. The large elevational extent sampled across yielded factors (including, but not limited to, lake elevation) that were useful in explaining model deviance and predicting to a geographically separate set of lakes. At this scale, the colonization-related factors clearly are important in discriminating occurrence from non-detection lakes. At smaller scales there may be other, biotic factors that shape organismal distribution (Trani, 2002). Such factors may only become evident upon sampling a much larger number of lakes than was sampled here. However, this study can serve as a springboard towards further investigating other factors that may shape the distribution of stickleback, and other fishes, in the region.

3.2 Salmonid occurrence

As indicated previously (Chapter 1, Section 1.4.3), the distribution of salmonid species at a large scale roughly mirrored that of stickleback. Across all sites sampled (northern and southern areas, n=54), 82% of stickleback-bearing lakes contained salmonid species. Conversely, in stickleback-absence lakes, 88% were also devoid (as evidenced by sampling) of salmonids. It appears that, although salmonid distribution patterns differ somewhat from that of stickleback, competition and/or predation are not limiting the range of either species in this region.

Focusing on the calibration lakes, further evidence of the parallels in distribution pattern between stickleback and salmonids is illustrated by the fact that the importance of some of the environmental factors most useful in explaining stickleback distribution is maintained when applied to salmonids (Figure 14). For instance, lake elevation, which ranked first in explained deviance (McFadden's ρ^2) amongst predictors of stickleback (Figure 14a), consistently ranked one or two in explaining the most deviance in univariate GLM models for salmonid taxa (Figures 14 b-d). Lake elevation only dropped from the top five rankings of all factors (univariate GLM model McFadden's ρ^2 and p-value) in its significance explaining anadromous (coho) salmonid presence (rank=8, Figure 14h). This drop in rank is at odds with the fact that lake elevation explains by far the most deviance (McFadden's ρ^2) of all models explaining anadromous status (= presence of coho salmon). The relatively high p-value (0.14) of lake elevation in explaining coho presence (Figure 14h) may be due to, 1) lake elevation being outranked by factors with which it is collinear (mean Pearson correlation with lake elevation for the seven factors = 0.42), and/or, 2) lack of data (coho were only present in 7 of 36 calibration lakes). A larger sample may have led to a lower p-value for lake elevation in explaining lake anadromous status.

A paucity of data for resident salmonid (n=9) and coho presence (n=7) lakes in the calibration data may be one reason that other important factors in explaining stickleback occurrence seem to explain deviance in the salmonid data (Figures 14c-d),

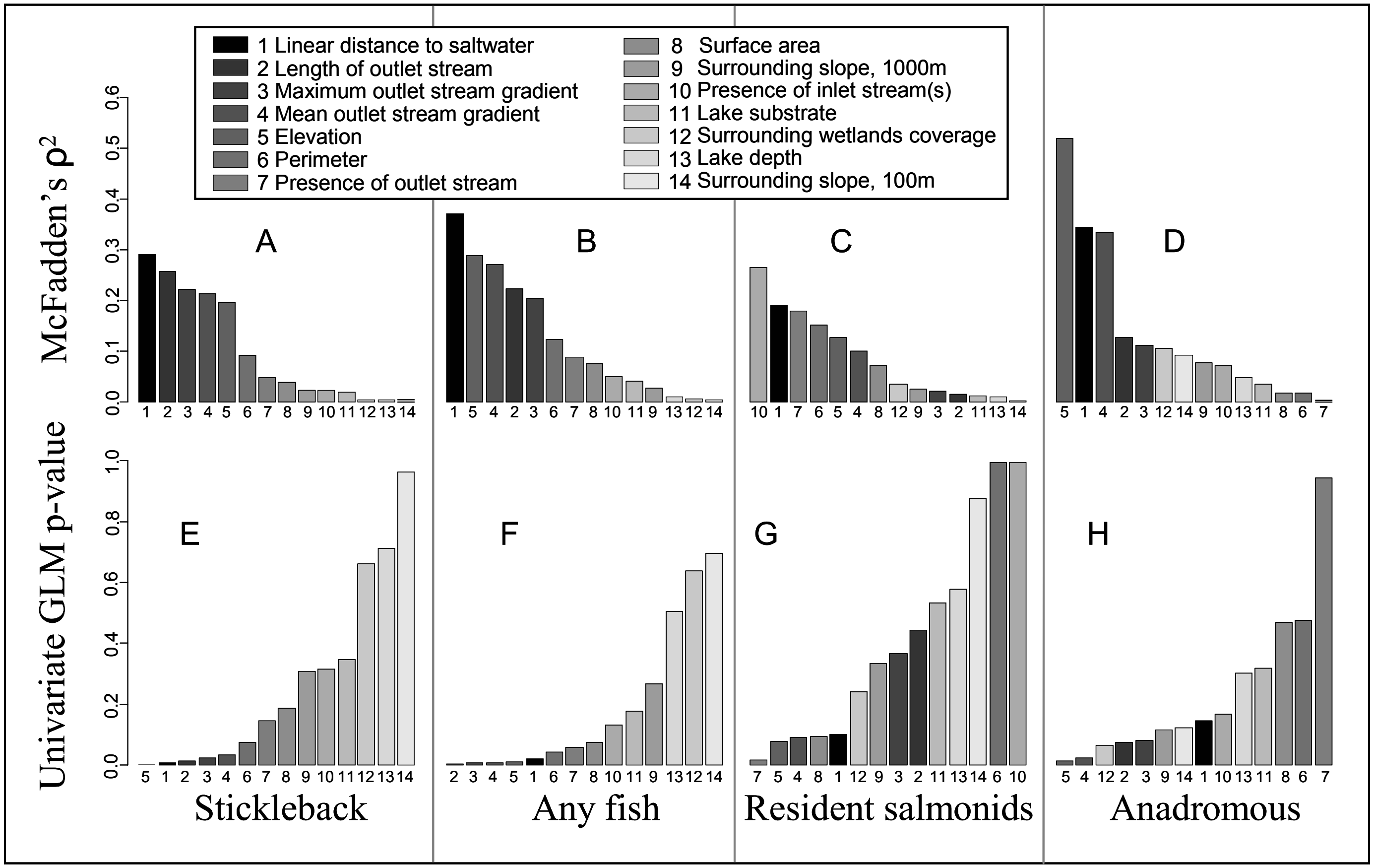


Figure 14. Changes in rank of univariate model McFadden's ρ2 and p-values when applied to other fish taxa sampled in this study (models correspond to each of the 14 environmental factors). Any Fish= stickleback, Dolly Varden, cutthroat trout, or coho salmon. Resident salmonids= Dolly Varden or cutthroat trout. Anadromous= coho salmon.

but that this does not translate to high levels of univariate model significance (Figures 14g-h). However, with exceptions (examples: length of outlet stream and maximum stream gradient for resident salmonids) (Figure 14c), important factors in explaining stickleback deviance hold for the salmonid taxa (Figure 14a-d).

Patterns in distribution of salmonids differ from that of stickleback in a number of ways. For resident salmonids, presence of an inlet stream seems to be an important factor, accounting for more deviance than any other factor. In all lakes in the northern, calibration set containing Dolly Varden or cutthroat trout, an inlet stream or streams were documented. In the southern set of lakes, the association between inlet streams and resident salmonids persisted, but not as strongly; four of six lakes with Dolly Varden or cutthroat had an inlet stream(s) in these lakes. Coho salmon were predicted marginally well by the percentage of wetlands coverage in the lake surrounding area (p=0.06) (Figure 14h). This wetlands factor never appeared in stickleback occurrence models. In addition, elevation was even more important as a factor for coho occurrence than stickleback, explaining over half the deviance (0.52) (Figure 14d), and being significant well below the 0.05 level (p=0.01) (Figure 14h). Coho were found mainly at very low elevations in the floodplain of the Taku River in the northern area, with an elevation of 32 m being the maximum at which coho were observed. Coho were not observed in the southern set of lakes. This may be because the floodplain landscape position was not prevalent in this area; random stratified site selection in this area yielded no lakes situated in the floodplain. It may be that anadromous fishes, including coho, are constrained to low-lying, floodplain habitats to an even greater extent than are stickleback (Figure 14, 15).

3.3 Generalizing to the total population of lakes

As stated previously, there were logistical constraints in the size class of lake visited in the present study. The question exists, how do the current results generalize to lentic water bodies not represented in this study? Lakes both lesser (<0.5 ha) and greater (>5 ha) in surface area than those sampled here show a skew in distribution towards lower elevations (Table 5).

Table 5. Elevational distribution of study lakes in relation to the entire population of lakes in the northern study area.

Elevation	Number of lakes (proportion of lakes in stratum)			
stratum	<0.5 ha	0.5-5 ha	>5 ha	Total
0-20.6	42(0.63)	15(0.22)	10(0.15)	67
20.6-32.2	12(0.44)	15 (0.56)	0 (0.0)	27
32.2-103.6	34(0.67)	15(0.29)	2(0.04)	51
103.6-323.6	19(0.45)	15(0.45)	8(0.19)	42
323.6-732.4	19(0.46)	15(0.37)	7(0.17)	41
>732.4	15(0.43)	15(0.43)	5(0.14)	35
Total (proportion)	141(0.54)	90(0.34)	32(0.12)	263

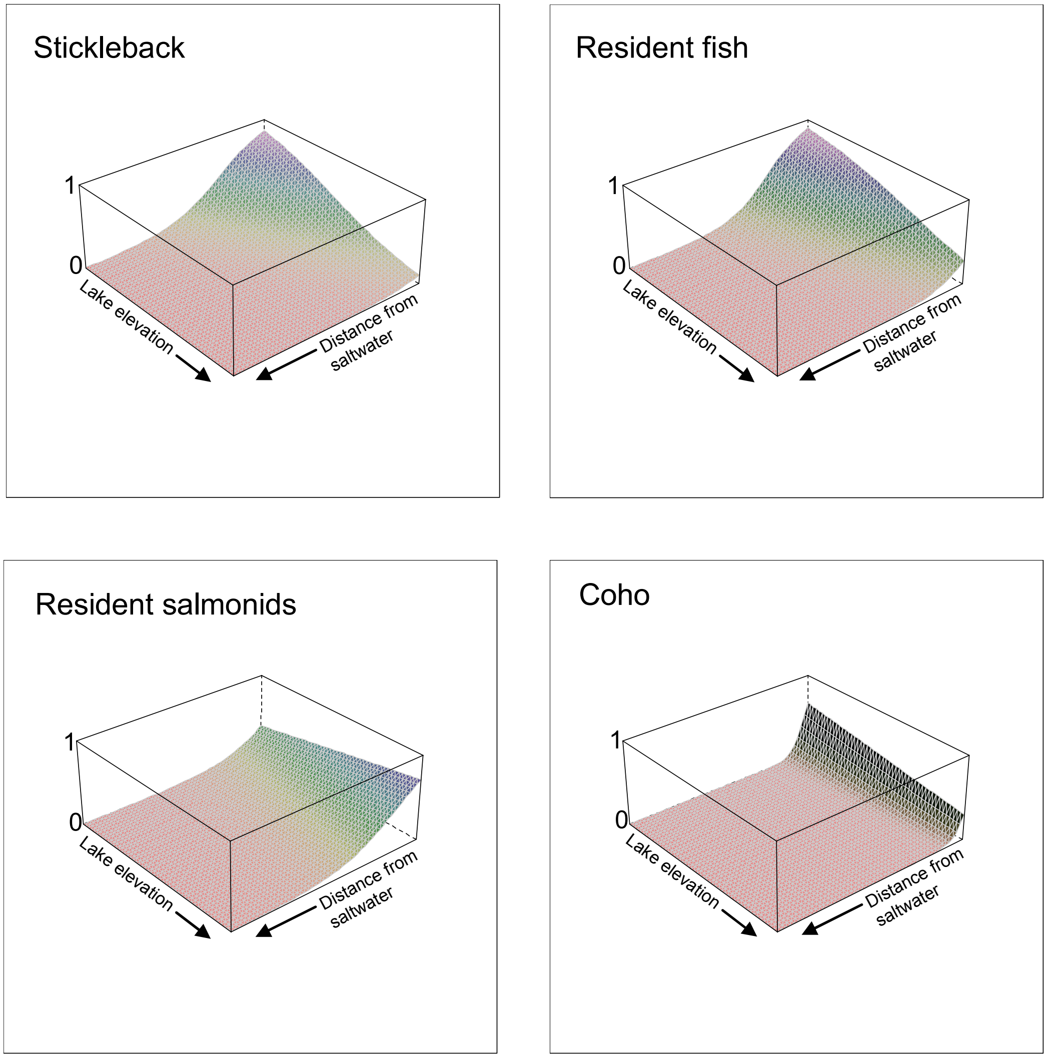


Figure 15. Comparison of simple (two-factor) models with lake elevation and distance from saltwater included for each of four fish taxa modeled.

Whether this skew in elevational distribution translates into a higher proportion of the entire lake population containing stickleback and/or other fishes is not known. Of study lakes 0.5-5 ha in surface area, 39% contained stickleback, 44% contained any fish species, and 19% contained coho salmon. As elevation shows strong correlation with the presence of all fish species, a higher proportion of lakes both smaller (<0.5 ha) and larger (>5 ha) than the study lakes might be expected to support fish populations. However, other factors may confound this relationship between lake size and fish species occurrence:

> 1) Smaller lentic waterbodies may be less likely associated with outlet streams. Lake surface area was correlated with outlet stream presence (Pearson correlation coefficient= 0.36). Whether this association holds true for lakes of other sizes is not known, but there is some evidence that it does. Of large lakes (>5 ha) in the calibration area, 81% are associated with (intersect) a USFS mapped stream arc (USFS 2003a), compared to smaller (0.5-5 ha) lakes of the (48%), and even smaller <0.5 ha lakes (5%). While this seemingly supports the idea that larger lakes are more likely to have an outlet stream, note that information gathered during site visits resulted in an assessment of 'outlet stream presence' for 75% (27 of 36) study lakes in the 0.5-5 ha range, well above the 48% rate based on outlet streams mapped (USFS 2003a) for lakes of this size class. The presence or absence of lake outlet streams—a factor associated with lake accessibility to

colonizing fishes, and potentially with other lake biotic and physical attributes—is not readily assessed from available GIS information.

2) Evidence from the present study regarding the role of lake surface area in shaping fish distribution is ambiguous. Lake surface area was included in 'complex' models in the present study, due to a p-value <0.25 (0.19) in a univariate GLM (Table 6). However, the proportion of deviance explained (McFadden's ρ^2) by surface area in a univariate stickleback GLM (0.038) (Table 6) is very small when contrasted with that explained by linear distance from saltwater (0.26), a nearly 7-fold difference. Similarly, a univariate GLM for coho occurrence using elevation (although not being significant; p-value =0.47) explained 20 times the deviance that surface area did (elevation McFadden's ρ^2= 0.34, surface area McFadden's ρ^2=0.017). The coefficient of surface area for coho occurrence is negative as well, indicating less likelihood of occurrence for larger lakes (although again note that the model p-value of 0.47 did not approach any conventional level of significance).

3) Evidence from the literature suggests that lake size may promote fish species occurrence. A strong trend in species richness vs. available area of habitat has been an observed trend in community ecology for many years (Preston, 1960; MacArthur & Wilson, 1963) including in lakes (Barbour & Brown, 1974; Tonn & Magnuson, 1982). A logical extension of this trend is that individual taxa may be

less likely present in smaller habitats; this has borne true for some species in lakes (Beauchamp *et al*., 1992; Hershey *et al*., 2006). It may be that for the small sample size and somewhat narrow range of lake sizes considered here, this general relationship was not clearly evident, but that it does exist over the entire size range of lakes in the region.

Table 6. Lake surface area univariate GLM parameters for fish taxa occurrence

Fish taxon	Estimated Coefficient	p-value	McFadden's ρ^2
Stickleback	0.52	0.19	0.038
Any fish	0.77	0.07	0.074
Resident salmonids	0.71	0.09	0.071
Coho	-0.41	0.47	0.017

3.4 Recommendations for future fish species occurrence modeling.

3.4.1 Sample size

A key recommendation to come from this research involves sample size. The utility of large sample sizes in species occurrence modeling can be illustrated in an example based on the stickleback modeling performed in the present study. For stickleback, univariate GLM models with lake elevation as the sole predictor resulted in coefficients of 0.766 for the intercept and 7.6 X 10^{-3} for the elevation term. As stated previously (Chapter 1, Section 1.2.2, the output probabilities of occurrence from logistic regression models are produced by the following equation:

$$P = e^{u} \big/ (1 + e^{u}) \text{ , where } u = A + b_1 X_1 + b_2 X + \ldots + b_n X_{nv} \text{ ;}$$

And so, substituting, we have, $P = e^{0.766 + 7.6x10^{-3}\,Elevation} \Big/ \left(1 + e^{0.766 + 7.6x10^{-3}\,Elevation}\right)$

This equation was the result of maximum likelihood estimation performed in GLM, and does not represent a true, underlying distribution of stickleback occurrence probabilities for the study area. However, because it has support from the data in the present study, it is a good starting point for generating some hypothetical data to illustrate the effect of sample size on occurrence models. Specifically, we can now generate random data which results from this underlying relationship between elevation and stickleback occurrence. Initially, we start from a mock dataset generated across 36 lakes (as in the present study). From 10 GLM models built from random data output from the above equation, the amount of variation in model fit is dramatic (Table 7):

Table 7. Variation in the significance and explanatory power of lake elevation in 10 trial univariate model runs. Models based on 36 hypothetical lakes.

Trial	Elevation term p-value	McFadden's ρ^2
1	0.04	0.29
2	0.02	0.37
3	0.06	0.17
4	0.02	0.38
5	0.11	0.22
6	0.03	0.46
7	0.02	0.25
8	0.006	0.49
9	0.03	0.49
10	0.06	0.17

The significance of the elevation term varies from (0.006-0.11) and the deviance explained over almost a 3-fold range (0.17-0.49). Contrasting this with models built on 100 theoretical lakes in the same fashion (Table 8):

Table 8. Variation in the significance and explanatory power of lake elevation in 10 trial model runs. Models based on 100 hypothetical lakes.

Trial	Elevation term p-value	McFadden's ρ^2
1	1.6×10^{-4}	0.30
2	3.0×10^{-4}	0.26
3	7.2×10^{-4}	0.37
4	8.5×10^{-4}	0.30
5	1.3×10^{-3}	0.31
6	1.9×10^{-3}	0.35
7	1.2×10^{-4}	0.42
8	3.0×10^{-4}	0.45
9	5.1×10^{-5}	0.41
10	1.2×10^{-4}	0.37

Here, the significance is much higher (p-values much smaller) and the proportion of deviance explained by models built on this larger dataset ranges over a less than 2-fold range (0.26-0.45). (10 models run on a theoretic sample of 1000 lakes results in p-values all ~0, with McFadden's ρ^2 ranging from 0.28-0.39.)

Increased sample sizes lead to greater model certainty as well as greater confidence in relative variable importance (Pearce & Ferrier, 2000; Stockwell & Peterson, 2002). At detection probabilities >0.5, greater sample size is suggested, as opposed to greater sampling effort at each site (Tyre *et al*., 2003). For these reasons, a sampling strategy as was performed in the present study (ie. short sampling visit and maximized number of sites) appears appropriate, but more sites sampled would be beneficial. However, although an intuitive sense suggests detection probability was greater than 50% in this study, repeat sampling trials at a number of sites would be useful to support this. With this caveat in mind, more sites visited may result in greater insight into relative importance of predictors, as well as increased model precision.

As discussed previously, model predictive utility, above simply building precise models, is generally desired in species occurrence modeling. In the current study, stickleback models were tested in the 18 southern lakes. This southern test set was chosen intentionally to be geographically separate from the calibration set, with the goal of testing model transferability. This general goal was only partially met by predicting to these southern lakes, because 18 lakes is not a large sample, and so the possibility of high model predictive performance by chance existed. This can be illustrated by theoretical trials similar to those above, but with the randomly generated data being the test data for 18 hypothetical lakes. For comparison, data from a hypothetical test set of 36 lakes is computed from the same underlying GLM as used above (Table 9):

Table 9. Comparison of AUC values for 10 trial model predictions to hypothetical test lakes.

	AUC value	
Trial	18 test lakes	36 test lakes
1	0.78	0.75
2	0.96	0.95
3	0.84	0.82
4	0.89	0.82
5	0.71	0.84
6	1.0	0.91
7	0.88	0.94
8	0.82	0.90
9	0.87	0.75
10	0.82	0.83

AUC values based on 18 hypothetical lakes range from 0.71-1.0, while for 36 hypothetical lakes AUC values range from 0.75-0.95. In 1000 trials, the standard deviation in AUC values from 18 lakes and from 36 lake test sets differs (0.11 and 0.07 respectively), as do the minimum AUC value for the two test sets (0.32-0.47). Although no minimum test set size can be suggested from these hypothetical test trials, the trend towards decreased variation in results with larger datasets, both in the calibration and prediction sets, is evident.

3.4.2 Choice of modeling method

The choice of modeling method is not a clear cut issue; all methods performed suitably in the present study. One clear difference was that GLM models were not as prone to overfit the original, calibration dataset as the other methodologies used, particularly when 3-5 terms were included in the model. A limitation of GLM is that they are constrained to follow the limited range of shapes allowed within logistic response. Only the parameters of the model can change, but the model fitted values will always follow an s-shaped logistic curve. This is in contrast with GAM, CT, and ANN, which all can assume a shape dictated (to a further extent) by the data at hand (Segurado & Araujo, 2004; Olivier & Wotherspoon., 2005). Occurrence records generated by both a linear model and non-linear model (sine curve) illustrate this behavior (Figure 16). GLM is adequate when occurrence likelihood is expected to vary linearly across environmental gradients. However, this is often not the case, and logistic regression does not have the flexibility to deal with more complex responses (Suarez-Seaone *et al.*, 2002; Olivier & Wotherspoon, 2005). A downside to the use of GAM, CT, and ANN is that many are already familiar with the more conventional GLM method, and gaining technical expertise in implementing these methods may be time consuming. However, when extensive datasets are available, and occurrence appears to interact with environmental factors in a complex manner, the more flexible methodologies are recommended. One further note regards the rather disjointed occurrence probabilities that are output from CT models. Although use of CT is not discouraged for future work, especially due to ease of use and interpretability, the ‘quantum’ steps that output probabilities take along environmental gradients seem an unrealistic representation of how occurrence probabilities may be distributed in the natural world (Figure 16). The other modeling methods (GLM and ANN) used here tend to output probabilities that change gradually as environmental factors vary; this may be a more realistic perspective.

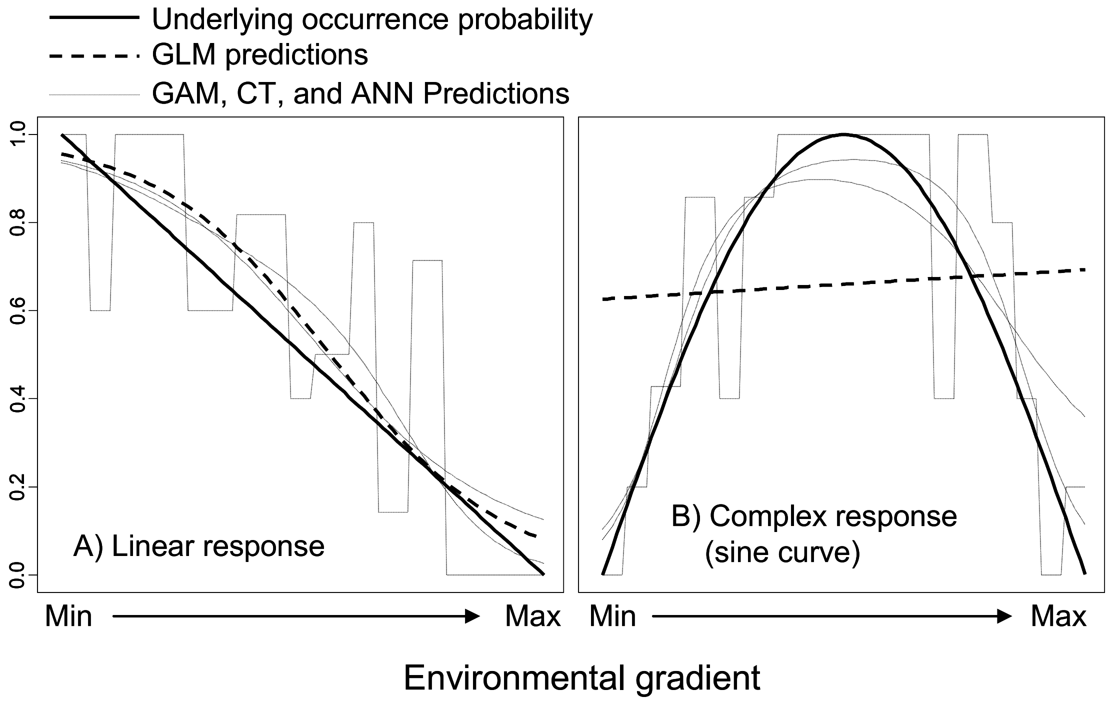


Figure 16. Comparison of occurrence model response curves for mock data distributed in a linear (A) and non-linear (B) pattern.

3.4 Summary of recommendations.

This section presents a concise summary of the findings, corresponding to secondary objective three, from the present study—organized by the primary or secondary objective under which the findings fall— with implications and recommendations for future fish distribution research in Southeast Alaska.

Primary objective 1: Landscape correlates of stickleback distribution (Chapter 1)

Colonization-related factors (e.g., lake elevation, slope of lake outlet stream) were apparently much more influential than extinction-related factors (depth, lake surface area, surrounding wetlands) in shaping stickleback distribution (Chapter 1, Section 1.3.3). For this reason, future species occurrence studies for both stickleback and other fish species conducted at large spatial scales may benefit from focusing on such factors. Additionally, because the colonization-related factors developed here serve as indirect indicators of potential barriers to fish migration, it is recommended that factors be developed that create a more direct tie to actual, ground–verified migration barriers (Chapter 1, Section 1.1). For instance, development of an index that describes the relationship between lake elevation and slope of a lake outlet stream to the likelihood of their actually existing a migration barrier(s) along the stream course could lead to models that use the likelihood of a barrier as an independent variable. This may more accurately explain and predict fish occurrence than the coarse-scale slope factor (with no ground verification) used in the present study.

While colonization-related factors were most useful in the present study, extinction-related factors describing habitat suitability may be of greater use if they are more fully developed. An example of such development would be sampling of water chemistry parameters through time, especially water temperature and pH. These parameters were measured during site-visits only in the present study. Because their variability through time was unknown, they were not included in analyses. However, both have been found influential in fish species occurrence studies conducted elsewhere (Rago & Wiener, 1986; Matuszek *et al*., 1990; Hershey *et al*., 2006).

Primary objective 2: Comparison of methods for modeling stickleback occurrence (Chapter 2)

The results from this study were somewhat equivocal regarding choice of modeling method (Chapter 2, Section 2.4). Simple models with two (and at the most three) factors included outperformed more complex (five-factor) models (Chapter 2, Section 2.4). While this is true for the small number of lakes sampled in this study, a larger sample size may support more complex models (Chapter 3). GLM models, because of their *a priori* response shape, lend themselves to these small sample sizes (Chapter 3, Section 3.4.1). However, greater sample sizes may support the complex response output of GAM, CT, and ANN (Chapter 3, Section 3.4.1).

Secondary objective 1: Comparison of salmonid and stickleback distribution (Chapter 3, Section 3.3.2)

Salmonids and stickleback overlap greatly in their distribution in Southeast Alaska, and it appears that the colonization-related factors helpful in explaining and predicting stickleback distribution work well for salmonids as well (Chapter 3, Section 3.2). Therefore it is suggested that such colonization-related factors (lake elevation, slope and presence of outlet stream, outlet stream length and its proxy, linear distance from saltwater) be considered first in modeling salmonids at large spatial scales. However, extinction-related factors may be important as well, especially at finer scales (Chapter 3, Section 3.1), and so should not be ruled out as potentially influential.

Secondary objective 2: Applying results of the present study to other lentic waterbodies (Chapter 3, Section 3.3).

There was no strong relationship found in the present study between lake size and stickleback distribution for the size range of lakes considered here (Chapter 1, Section 1.4). However, both smaller (<0.5 ha) and larger (> 5 ha) lentic waterbodies show a skew in distribution towards lower elevation in the northern, calibration study area (Chapter 3, Section 3.3), which may promote likelihood of fish species occurrence in these lakes. Further studies of species occurrence over a wider range of lake sizes should be conducted to verify or refute the utility of the correlates important for the 0.5-5 ha size range considered here.

References

Baade B.T. (1961) *Lake fish summary*. Alaska Department of Fish and Game, Division of Sport Fish, Douglas, AK, USA.

Barbour C.D. & Brown J.H. (1974) Fish species diversity in lakes. *American Naturalist*, **198**, 473-489.

Beauchamp J.J., Christensen S.W. & Smith E.P. (1992) Selection of factors affecting the presence of brook trout (*Salvelinus fontinalis*) in Adirondack lakes: a case study. *Canadian Journal of Fisheries and Aquatic Sciences*, **49**, 597-608.

Beier P. & Noss R.F. (1998) Do habitat corridors provide connectivity? *Conservation Biology*, **12**, 1241-1252.

Beisner B.E., Peres-Neto P.R., Lindstrom E.S., Barnett A., & Longhi M.L. (2006) The role of environmental and spatial processes in structuring lake communities from bacteria to fish. *Ecology*, **87**, 2985-2991.

Bell M.A., Orti G., Walker J.A. & Koenings J.P. (1993) Evolution of pelvic reduction in threespine stickleback fish: a test of competing hypotheses. *Evolution*, **47**, 905-914.

Bell M.A., Aguirre W.E., & Buck N.J. (2004) Twelve years of contemporary armor evolution in a threespine stickleback population. *Evolution*, **58**, 814-824.

Bertolo A., & Magnan P. (2006) Spatial and environmental correlates of fish community structure in Canadian Shield lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, **63**, 2780-2792.

Bishop C.M. (1995) *Neural Networks for Pattern Recognition*. Oxford University Press Inc, New York, USA.

Boyce M.S., Vernier P.R., Nielsen S.E. & Schmiegelow F.K.A. (2002) Evaluating resource selection functions. *Ecological Modelling*, **157**, 281-300.

Breiman L., Friedman J.H., Olshen A. & Stone C.G. (1984) *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA, USA.

Brotons L., Thuiller W., Araujo M.B. & Hirzel A.H. (2004) Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, **27**, 437-448.

Burnham K.P. & Anderson D.R. (2002) *Model Selection and Multimodel Inference: A Practical and Information-theoretic Approach*. 2nd edn. Springer Science, New York.

Cavalli L., Miquelis A. & Chappaz R. (2001) Combined effects of environmental factors and predator-prey interactions on zooplankton assemblages in five high alpine lakes. *Hydrobiologia*, **455**, 127-135.

Conroy C.J., Demboski J.R. & Cook J.A. (1999) Mammalian biogeography of the Alexander Archipelago of Alaska: a north temperate nested fauna. *Journal of Biogeography*, **26**, 343-352.

Corner G.D., Yevzerov V.Y., Kolka V.V. & Moller J.J. (1999) Isolation basin stratigraphy and Holocene relative sea-level change at the Norwegian-Russian border north of Nikel, northwest Russia. *Boreas*, **28**, 146-166.

D'Arcy P. & Carignan R. (1997) Influence of catchment topography on water chemistry in southeastern Quebec Shield lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, **54**, 2215-2227.

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

De'ath G. & Fabricius K. (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, **81**, 3178-3192.

de la Hoz Franco, E.A. & Budy P. (2005) Effects of biotic and abiotic factors on the distribution of trout and salmon along a longitudinal stream gradient. *Environmental Biology of Fishes*, **72**, 379-391.

Dobrowski S.Z., Greenberg J.A., Ramirez C.M. & Ustin S.L. (2006) Improving image derived vegetation maps with regression based distribution modeling. *Ecological Modelling*, **192**, 126-142.

Duncan S.L. & Lach D.H. (2006) Privileged knowledge and social change: effects on different participants of using geographic information systems technology in natural resource management. *Environmental Management*, **38**, 267-285.

Dunham J., Schroeter R. & Rieman B. (2003) Influence of maximum water temperature on occurrence of Lahontan cutthroat trout within streams. *North American Journal of Fisheries Management*, **23**, 1042-1049.

Edmundson J.A. & Mazumder A. (2002) Regional and hierarchical perspectives of thermal regimes in subarctic, Alaskan lakes. *Freshwater Biology*, **47**, 1-17

Elith J. & Burgman M. (2002) Predictions and their validation: rare plants in the Central Highlands, Victoria, Australia. In: *Predicting Species Occurrence: Issues of Accuracy and Scale* (Eds J.M. Scott, P.J. Heglund & M.L. Morrison), pp. 303-314. Island Press, Covelo, CA, USA.

Elith J., Graham, C.H., Anderson, R.P., Dudik M., Ferrier S., Guisan A., Hijmans R.J., Huettmann F., Leathwick J., Lehmann A., Li J., Lohmann L.G., Loiselle B.A., Manion G., Moritz C., Nakamura M., Nakazawa Y., Overton J.M., Peterson T., Phillips S.J., Richardson K., Scachetti-Pereira R., Schapire R., Soberon J., Williams S., Wisz M.S. & Zimmermann N.E. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 1-23.

Fielding A.H. & Bell J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **1**, 38-49.

Fielding A.H. & Haworth P.F. (1995) Testing the generality of bird-habitat models. *Conservation Biology*, **9**, 1466-1481.

Foster S.A., Baker J.A. & Bell M.A. (2003) The case for conserving the threespine stickleback: protecting an adaptive radiation. *Fisheries*, **28**, 10-18.

Franklin J. (1998) Predicting the distribution of shrub species in southern California from climate and terrain-derived variables. *Journal of Vegetation Science*, **9**, 733-748.

Fransen B.R., Duke S.D., McWethy L.G., Walter J.K. & Bilby R.E. (2006) A logistic regression model for predicting the upstream extent of fish occurrence based on geographical information systems data. *North American Journal of Fisheries Management*, **26**, 960-975.

Garzon M.B., Blazek R., Neteler M., Sanchez de Dios R., Ollero H.S. & Furlanello C. (2006) Predicting habitat suitability with machine learning models: the potential area of *Pinus sylvestris* L. in the Iberian Peninsula. *Ecological Modelling*, **197**, 383-393.

Gibson G. (2005) The synthesis and evolution of a supermodel. *Science*, **307**, 1890-1891.

Girard P. & Angers B. (2006) The impact of postglacial marine invasions on the genetic diversity of an obligate freshwater fish, the longnose dace (*Rhinichthys cataractae*), on the Quebec peninsula. *Canadian Journal of Fisheries and Aquatic Sciences*, **63**, 1429-1438.

Granadeiro J.P., Andrade J. & Palmeirim J.M. (2004) Modelling the distribution of shorebirds in estuarine areas using generalized additive models. *Journal of Sea Research*, **42**, 227-240.

Guisan A. & Zimmermann N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147-186.

Guisan A. & Thuiller W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993-1009.

Gutierrez D., Fernandez P., Seymour A.S. & Jordano E. (2005) Habitat distribution models: are mutualist distributions good predictors of their associates? *Ecological Applications*, **15**, 3-18.

Hagen D.W. & Gilbertson L.G. (1972) Geographic variation in environmental selection in *Gasterosteus aculeatus* L. in the Pacific Northwest, America. *Evolution*, **26**, 32-51.

Harrell F.E., Lee K.L. & Mark D.B. (1996) Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, **15**, 361–387.

Hastie T.J. & Tibshirani R.J. (1990) *Generalized Additive Models*. Chapman and Hall, London.

Hastings K. (2005) Long term persistence of isolated fish populations in the Alexander Archipelago. Phd. Dissertation, University of Montana, USA. 217 pp.

Hernandez P.A., Graham C.H., Master L.L. & Albert D.L. (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, **29**, 773-785.

Hershey A.E., Beaty S., Fotino K., Keyse M., Mou P.P., O'Brien W.J., Ulseth A.J., Gettel G.A., Lienesch P.W., Luecke C., McDonald M.E., Mayer C.H., Miller M.C., Richards C., Schuldt J.A. & Whalen S.C. (2006) Effect of landscape factors on fish distribution in arctic Alaskan lakes. *Freshwater Biology*, **51**, 39-55.

Hosmer D.W. & Lemeshow S. (2000) *Applied Logistic Regression*. Wiley and Sons, New York.

Jackson D.A., Peres-Neto P.R. & Olden J.D. (2001) What controls who is where in freshwater fish communities—the roles of biotic, abiotic, and spatial factors. *Canadian Journal of Fisheries and Aquatic Sciences*, **58**, 157-170.

Jensen O.P., Seppelt R., Miller T.J. & Bauer L.J. (2005) Winter distribution of blue crab *Callinectes sapidus* in Chesapeake Bay: application and cross-validation of a two-stage generalized additive model. *Marine Ecology Progress Series*, **299**, 239-255;

Johnson L.B. & Gage S.H. (1997) Landscape approaches to the analysis of aquatic ecosystems. *Freshwater Biology*, **37**, 113-132.

Johnson C.M., Johnson L.B., Richars C. & Beasley V. (2002) Predicting the occurrence of amphibians: an assessment of multiple-scale models. In: *Predicting Species Occurrence: Issues of Accuracy and Scale* (Eds J.M. Scott, P.J. Heglund & M.L. Morrison), pp. 157-170. Island Press, Covelo, CA, USA.

Kalff J. (2002) *Limnology*. Prentice Hall, New Jersey, USA.

Klemas V.V. (2001) Remote sensing of landscape-level coastal environmental indicators. *Environmental Management*, **27**, 47-57.

Kollarits S., Kuschnig G., Veselic M., Pavicic A., Soccorso C. & Aurighi M. (2006) Decision-support systems for groundwater protection: innovative tools for resource management. *Environmental Geology*, **49**, 840-848.

Kontula T. & Vainola R. (2004) Molecular and morphological analysis of secondary contact zones of *Cottus gobio* in Fennoscandia: geographical discordance of character transitions. *Biological Journal of the Linnean Society*, **81**, 535-552.

Kratz T.K., Webster K.E., Bowser C.J., Magnuson J.J. & Benson B.J. (1997) The influence of landscape position on lakes in northern Wisconsin. *Freshwater Biology*, **37**, 209-217.

Kristjansson B.K. (2005) Rapid morphological changes in threespine stickleback, *Gasterosteus aculeatus*, in freshwater. *Environmental Biology of Fishes*, **74**, 357-363.

Kruse C.G., Hubert W.A. & Rahel F.J. (1997) Geomorphic influences on the distribution of Yellowstone cutthroat trout in the Absaroka Mountains, Wyoming. *Transactions of the American Fisheries Society*, **126**, 418-427.

Larsen C.F., Motyka R.J., Freymueller J.T., Echelmeyer K.A. & Ivins E.R. (2004) Rapid uplift of southern Alaska caused by recent ice loss. *Geophyisical Journal International,* **158**, 1118-1133.

Lehmann A., Overton J.M. & Leathwick J.R. (2002) GRASP: generalized regression analysis and spatial prediction. *Ecological Modelling*, **157**, 189-207.

Li H. & Wu J. (2004) Use and misuse of landscape indices. *Landscape Ecology*, **19**, 389-399.

Liu C., Berry P.M., Dawson T.P. & Pearson R.G. (2005) Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, **28**, 385-393.

Luoto M., Heikkinen R.K., Poyry J. & Saarinen K. (2006) Determinants of the biogeographical distribution of butterflies in boreal regions. *Journal of Biogeography*, **33**, 1764-1778.

Lusk J.J., Guthery F.S. & DeMaso S.J. (2002) A neural network model for predicting northern bobwhite abundance in the Rolling Red Plains of Oklahoma. In: *Predicting species occurrence: issues of accuracy and scale* (Eds J.M. Scott, P.J. Heglund & M.L. Morrison), pp. 345-356. Island Press, Covelo, CA, USA.

MacArthur R. & Wilson E.O. (1963) An equilibrium theory of insular zoogeography. *Evolution*, **17**, 373-387.

MacNally R. & Fleishman E. (2002) Using "indicator" species to model species richness: model development and predictions. *Ecological Applications*, **12**, 79-92.

Magnuson J.J., Tonn W.M., Banerjee A., Toivonen J., Sanchez O. & Rask M. (1998) Isolation vs. extinction in the assembly of fishes in small northern lakes. *Ecology*, **79**, 2941-2956.

Mann, D. (1986) Wisconsin and Holocene glaciation of Southeast Alaska. In: *Glaciation in Alaska, the Geologic Record*. (Eds T. D. Hamilton, K.M. Reed & R.M. Thorson), pp. 237-265. Alaska Geological Society, Anchorage, AK, USA.

Matuszek J.E. & Beggs G.L. (1988) Fish species richness in relation to lake area, pH, and other abiotic factors in Ontario lakes. *Canadian Journal of Fisheries and Aquatic Research*, **45**, 1931-1941.

Moisen G.G. & Frescino T.S. (2005) Comparing five modeling techniques for predicting forest characteristics. *Ecological Modelling*, **157**, 209-225.

Morrow JE. (1980) *The Freshwater Fishes of Alaska*. Alaska Northwest Publishing Co., Anchorage, AK, USA.

Mortberg U.M, Balfors B. & Knol W.C. (2007) Landscape ecological assessment: a tool for integrating biodiversity issues in strategic environmental assessment and planning. *Journal of Environmental Management*, **82**, 457-470.

Nowacki, G., Shephard, M., Krosse, P., Pawuk, W., Fisher, G., Baichtal, J., Brew D.A., Kissinger, E. & Brock, T. (2001) *Ecological Subsections of Southeast Alaska and Neighboring Areas of Canada*. U.S. Department of Agriculture, Forest Service, Technical Publication No. R10-TP-75, 306 pp.

Oakes R.M., Gido K.B., Falke J.A., Olden J.D. & Brock B.L. (2005) Modelling of stream fishes in the Great Plains, USA. *Ecology of Freshwater Fish*, **14**, 361-374.

Ohman J., Buffam I., Englund G. & Blom A. (2006) Associations between water chemistry and fish community composition: a comparison between isolated and connected lakes in northern Sweden. *Freshwater Biology*, **51**, 510-522.

Olden J.D. & Jackson D.A. (2001) Fish-habitat relationships in lakes: gaining predictive and explanatory insight by using artificial neural networks. *Transactions of the American Fisheries Society*, **130**, 878-897.

Olden J.D. & Jackson D.A. (2002) A comparison of statistical approaches for modeling fish species distributions. *Freshwater Biology*, **47**, 1976-1995.

Olden J.D. (2003) A species-specific approach to modeling biological communities and its potential for conservation. *Conservation Biology*, **17**, 854-863.

Olden J.D., Joy M.K. & Death R.G. (2006) Rediscovering the species in community-wide predictive modeling. *Ecological Applications,* **16**, 1449-1460.

Olivier F. & Wotherspoon S.J. (2005) GIS-based application of resource selection functions to the prediction of snow petrel distribution and abundance in East Antarctica: comparing models at multiple scales, *Ecological Modelling*, **189**, 105-129.

Ostwald M. (2002) GIS-based support tool system for decision-making regarding local forest protection: illustrations from Orissa, India. *Environmental Management*, **30**, 35-45.

Parker T.H., Stansberry B.M., Becker C.D. & Gipson P.S. (2005) Edge and area effects on the occurrence of migrant forest songbirds. *Conservation Biology*, **19**, 1157-1167.

Paszkowski C.A. & Tonn W.M. (2000) Community concordance between the fish and aquatic birds of lakes in northern Alberta, Canada: the relative importance of environmental and biotic factors. *Freshwater Biology*, **43**, 421-437.

Pearce J. & Ferrier S. (2000) Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, **133**, 225-245.

Pearman P.B., Penskar M.R., Schools E.H. & Enander H.D. (2006) Identifying potential indicators of conservation value using natural heritage occurrence data. *Ecological Applications*, **16**, 186-201.

Pearson R.G., Thuiller W., Araujo M.B., Marinez-Meyer E., Brotons L., McClean C., Miles L., Segurado P., Sawson T.P. & Less D.C. (2006) Model-based uncertainty in species range prediction. *Journal of Biogeography*, **33**, 1704-1711.

Porter M.S., Rosenfeld J. & Parkinson E.A. (2000) Predictive models of fish species distribution in the Blackwater drainage, British Columbia. *North American Journal of Fisheries Management*, **20**, 349-359.

Power G., Pope G.F. & Coad B.W. (1973) Postglacial colonization of the Matamek River, Quebec, by fishes. *Journal of the Fisheries Research Board of Canada*, **30**, 1586–1589.

Preston F.W. (1960) Time and space and the variation of species. *Ecology*, **41**, 611-627.

Quist M.C., Hubert W.A. & Rahel F.J. (2006) Concurrent assessment of fish and habitat in warmwater streams in Wyoming. *Fisheries Management and Ecology*, **13**, 9-20.

R Development Core Team (2006) *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rago P.J. & Wiener J.G. (1986) Does pH affect fish species richness when lake area is considered? *Transactions of the American Fisheries Society*, **115**, 438-447.

Rahel F.J. (1984) Factors structuring fish assemblages along a bog lake successional gradient. *Ecology*, **65**, 1276-1289.

Reimchen T.E. (1994) Predators and morphological evolution in threespine stickleback. In: *The evolutionary biology of the threespine stickleback* (Eds M.A. Bell & S.A. Foster), pp. 240-275. Oxford University Press, New York, USA.

Rieman B.E. & McIntyre J.D. (1995) Occurrence of bull trout in naturally fragmented habitat patches of varied size. *Transactions of the American Fisheries Society*, **124**, 285-296.

Ruggles A.K. (1994) Habitat selection by loons in southcentral Alaska. *Hydrobiologia*, **279/280**, 421-430.

Sandgren P. & Snowball I. (2001) The late Weichselian sea level history of Kullen Peninsula in northwest Skane, southern Sweden. *Boreas*, **30**, 115-130

Sarmaja-Korjonen K. & Hyvarinen H. (2002) Subfossil littoral Cladocera as indicators of brackish-water Littorina transgression of the Baltic Basin in a small lake in Finland. *Boreas*, **31**, 356-361

Scott W.B. & Crossman E.J. (1973) Freshwater fishes of Canada. *Bulletin of the Fisheries Research Board of Canada* 184. 966 pp.

Segurado P. & Araujo M.B. (2004) An evaluation of methods for modeling species distributions. *Journal of Biogeography*, **31**, 1555-1568.

Sergio F., Marchesi L. & Pedrini P. (2004) Integrating individual habitat choices and regional distribution of a biodiversity indicator and top predator. *Journal of Biogeography*, **31**, 619-628.

Smith T.A. & Kraft C.E. (2005) Stream fish assemblages in relation to landscape position and local habitat variables. *Transactions of the American Fisheries Society*, **134**, 430-440.

Snyder J.A., Forman S.L., Mode W.N. & Tarasov G.A. (1997) Postglacial relative sea-level history: sediment and diatom records of emerged coastal lakes, north-central Kola Peninsula, Russia. *Boreas*, **26**, 329-346.

Solem J.O., Solem T., Aagaard K. & Hanssen O. (1997) Colonization and evolution of lakes on the central Norwegian coast following deglaciation and land uplift 9500 to 7800 years B.P. *Journal of Paleolimnology*, **18**, 269-281.

Steel E., Fiest B.E., Jensen D.E., Pess G.R., Sheer M.B., Brauner J.B. & Bilby R.E. (2004) Landscape models to understand steelhead (*Oncorhynchus mykiss*) distribution and help prioritize barrier removals in the Willamette basin, Oregon, USA. *Canadian Journal of Fisheries and Aquatic Sciences*, **61**, 999-1011.

Stockwell D.R.B. & Peterson A.T. (2002) Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, **148**, 1-13.

Suarez-Seaone S., Osborne P.E. & Alonso J.C. (2002) Large-scale habitat selection by agricultural steppe birds in Spain: identifying species-habitat responses using generalized additive models. *Journal of Applied Ecology*, **39**, 755-771.

Swales S. (2006) Review of factors affecting the distribution and abundance of rainbow trout (*Oncorhynchus mykiss* Walbaum) in lake and reservoir systems. *Lake and Reservior Management*, **22**, 167-178.

Tabachnick B. & Fidell L. (2000) *Using Multivariate Statistics*, 4th edn. Pearson Allyn & Bacon, New York.

Theobald D.M., Hobbs N.T., Bearly T., Zack J.A., Shenk T. & Riebsame W.E. (2000) Incorporating biological information in local land-use decision making: designing a system for conservation planning. *Landscape Ecology*, **15**, 35-45.

Therneau T.M. & Atkinson B. (2006*) rpart: Recursive Partitioning Software.* R package version 3.1-29.

Thuiller W. (2003) BIOMOD—optimizing prediction of species distributions and projecting potential future shifts under global change. *Global Change Biology***,** **9**, 1353-1362.

Thuiller W., Araujo, M.B. & Lavorel S. (2003) Generalized models vs. classification tree analysis: predicting spatial distributions of plant species at different scales. *Journal of Vegetation Science*, **14**, 669-680.

Tonn W.M. & Magnuson J.J. (1982) Patterns in the species composition and richness of fish assemblages in northern Wisconsin lakes. *Ecology*, **63**, 1149-1166.

Torgerson C.E. & Close D.A. (2004) Influence of habitat heterogeneity on the distribution of larval Pacific lamprey (*Lampetra tridentate*) at two spatial scales. *Freshwater Biology*, **49**, 614-630.

Trani M.K. (2002) The influence of spatial scale on landscape pattern description and wildlife habitat assessment. In: *Predicting Species Occurrence: Issues of Accuracy and Scale* (Eds J.M. Scott, P.J. Heglund & M.L. Morrison), pp. 141-155. Island Press, Covelo, CA, USA.

Turgeon K. & Rodriguez M.A. (2005) Predicting microhabitat selection in juvenile Atlantic salmon *Salmo salar* by the use of logistic regression and classification trees. *Freshwater Biology*, **50**, 539-551.

Tyre A.J., Tenhumberg B., Field S.A., Niejalke D., Parris K. & Possingham H.P. (2003) Improving precision and reducing bias in biological surveys: estimating false-negative error rates. *Ecological Applications*, **13**, 1790-1801.

USFS (2003a) *Tongass National Forest lakes GIS coverage*. URL: http://www.fs.fed.us/r10/tongass/gisinfo/pages/metadata/water.html, United States Forest Service, Region 10, Juneau AK, USA [accessed on 9 Jan 2007].

USFS (2003b) *Tongass National Forest aerial photogrammetry*. URL: http://www.fs.fed.us/r10/tongass/gisinfo/pages/metadata/imag.html , United States Forest Service, Region 10, Juneau AK, USA [accessed on 9 Jan 2007].

USFWS (2006) *U.S. Fish and Wildlife Service, Classification of Wetlands and Deepwater Habitats of the United States*. FWS/OBS-79/31., U.S. Fish and Wildlife Service, Branch of Habitat Assessment, Washington, DC.

USGS (2004) *1 Arc Second SRTM Elevation Data, Reprocessed to GeoTIFF*. The Global Land Cover Facility. College Park, Maryland, USA. URL: http://www.landcover.org.

Van Zyll de Jong M.C. & Cowx I.G. (2005) Association between biogeographical factors and boreal lake fish assemblages. *Fisheries Management and Ecology*, **12**, 189-199.

Vaughan I.P. & Ormerod S.J. (2005) The continuing challenges of testing species distribution models. *Journal of Applied Ecology*, **42**, 720-730.

Vayssieres M.P., Plant, R.E. & Allen-Diaz B.H. (2000) Classification trees: an alternative non-parametric approach for predicting species distributions. *Journal of Vegetation Science*, **11**, 679-694.

Wetzel, R.G. (2001) *Limnology: Lake and River Ecosystems*. 3rd edn. Academic Press, San Diego.

Willett T.R. (2001) Spiders and other arthropods as indicators in old-growth versus logged redwood stands. *Restoration Ecology*, **9**, 410-420.

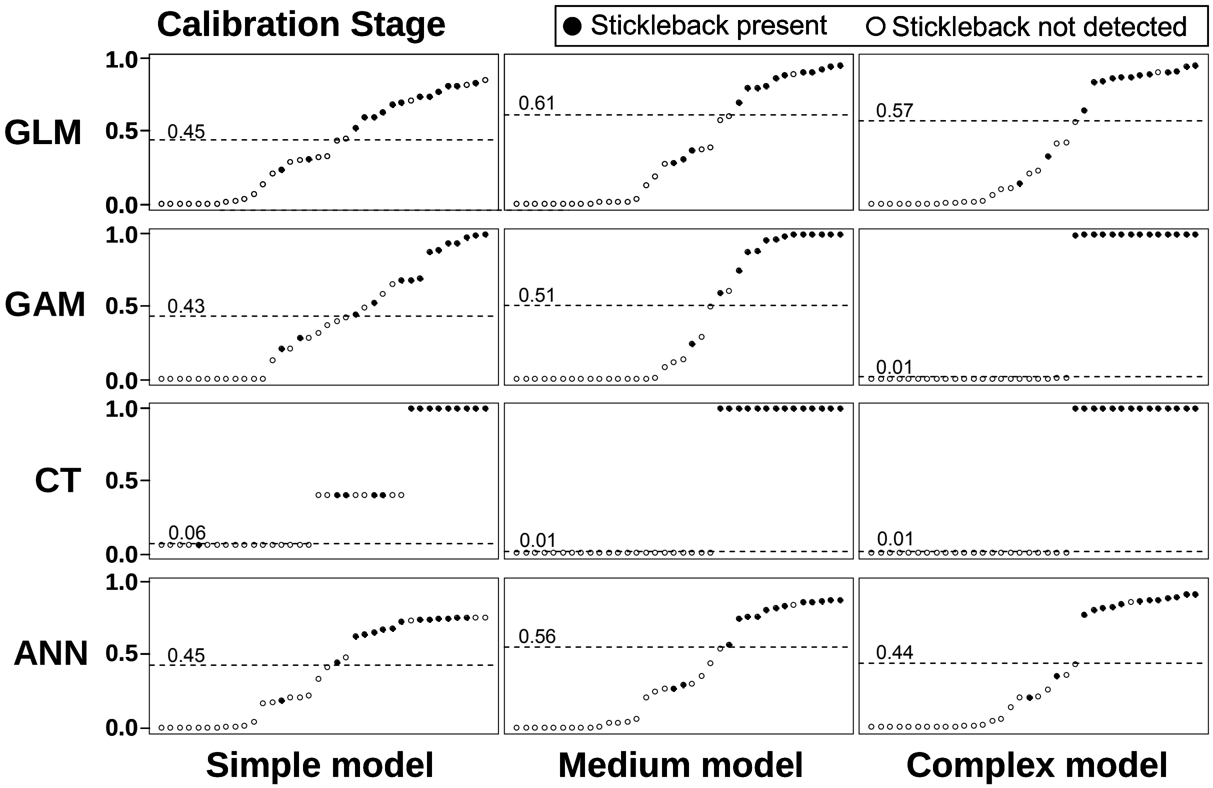
Wood S.N. (2006) *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC,

Zabel C.J., Roberts L.M., Mulder B.S., Stauffer H.B., Dunk J.R., Wolcott K., Solis D., Gertsch M., Woodbridge B., Wright A., Goldsmith G. & Keckler C. (2002) A collaborative approach in adaptive management at a large-landscape scale. In: *Predicting Species Occurrence: Issues of Accuracy and Scale* (Eds J.M. Scott, P.J. Heglund & M.L. Morrison), pp. 241-254. Island Press, Covelo, CA, USA.

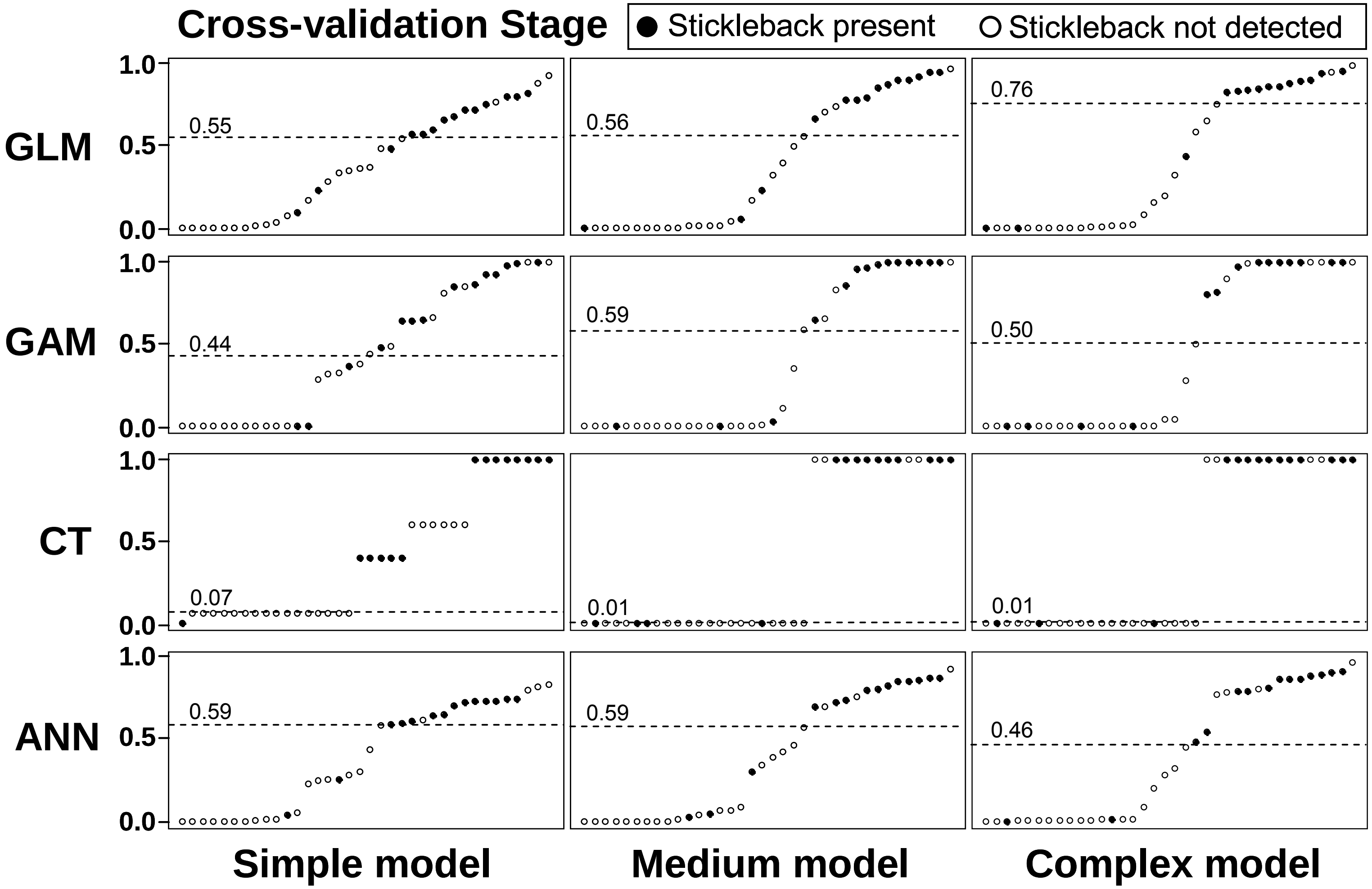
Zaniewski A.E., Lehmann A. & Overton J.M. (2002) Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157**, 261-280.

Appendices

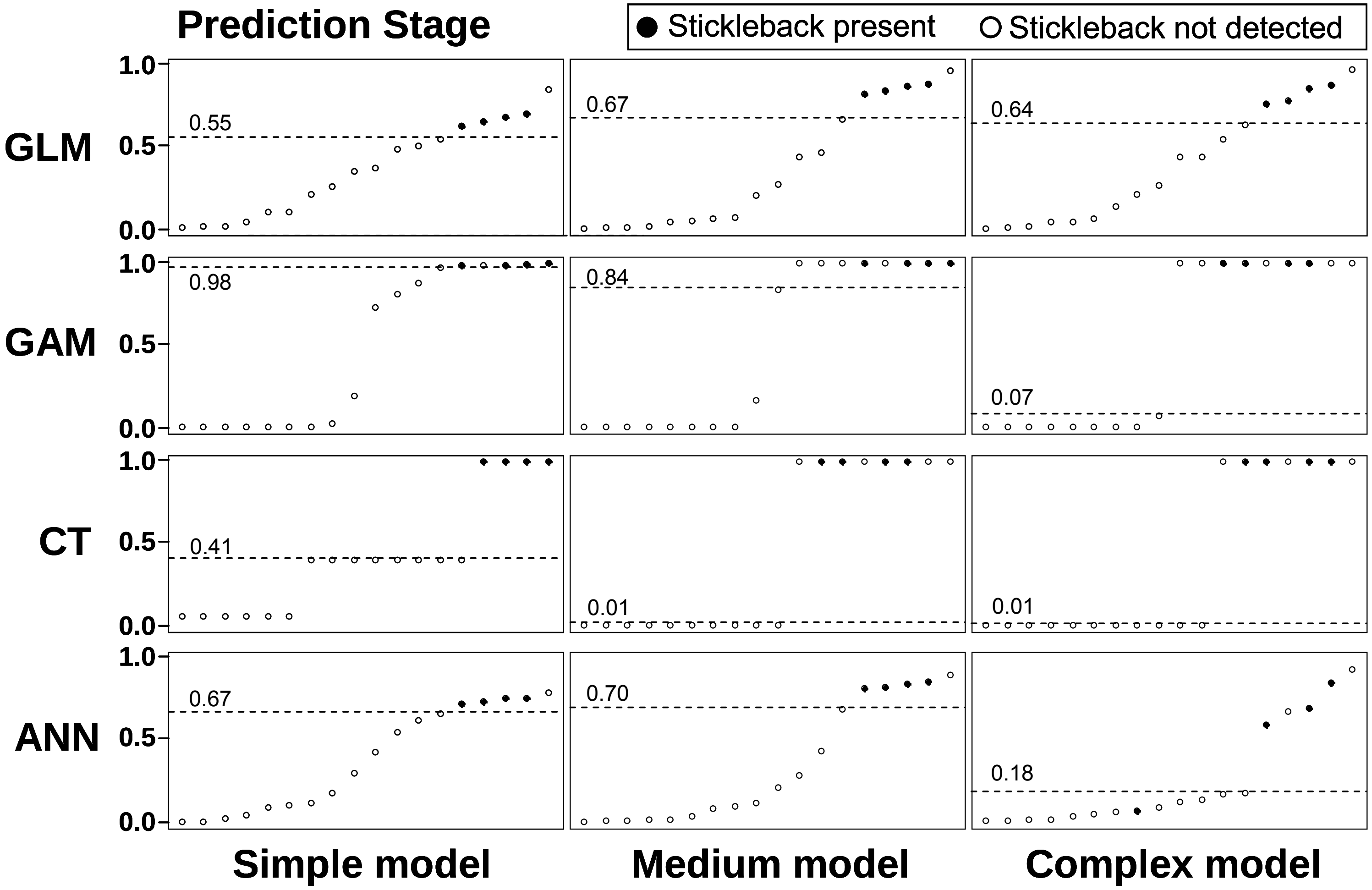
Appendix A. Output probabilities from the four modeling methods and three levels of model complexity on calibration, cross-validation, and prediction to the set of test lakes.



Appendix A, Figure 1. Model output probabilities for the calibration stage of model assessment.



Appendix A, Figure 2. Model output probabilities for the cross-validation stage of model assessment.



Appendix A, Figure 3. Model output probabilities for the prediction stage of model assessment.

Appendix B. Environmental attribute values, geographic location, and fish presence records for the 54 lakes sampled.

Appendix B, Table 1. Environmental attribute values of the 54 lakes sampled.

Lake ID	Area (N= Northern, S=Southern)	Surface area (ha)	Linear distance to saltwater (m)	Perimeter (m)i	Lake elevation (m)	Lake depth (m)	Presence of inlet stream (s) (bin)	Presence of outlet stream (s) (bin)	Lake substrate (bin)	Wetlands coverage (%)	Maximum outlet stream gradient (%)	Minimum outlet stream gradient (%)	Slope in lake surrounding area (100 m buffer)	Slope in lake surrounding area (1000 m buffer)	Length of outlet stream (m)
1	N	0.35	700	310	3	0.7	0	0	1	1	3	1.4	0.8	10	490
2	N	1.1	130	530	2.6	0.5	0	0	1	0.91	3	1.6	2	8.7	100
3	N	2.4	5000	580	760	15	1	1	0	0.04	42	6.3	9.7	57	4600
4	N	3.2	550	780	110	16	0	1	0	0	21	5.3	6.6	28	2300
5	N	3.3	1300	700	150	6	0	1	0	0.1	33	6.2	20	36	2100
6	N	3.2	2200	1200	210	2.9	1	1	1	0.16	29	6.9	8.6	31	2800
7	N	0.9	6300	590	66	1.3	1	1	0	0.11	19	5.2	3.9	29	8600
8	N	0.2	1200	210	30	2.8	1	1	0	0.47	3.3	1.2	2.1	61	1600
9	N	0.49	3900	320	850	0.93	1	0	0	0.01	68	16	8.4	34	7600
10	N	0.73	7800	460	730	1.7	1	1	0	0	65	8.4	9.4	21	13000
11	N	1	5200	570	51	4.3	1	1	0	0.41	23	5.8	4.6	9.4	7100
12	N	2.4	2400	940	32	0.73	1	1	0	0.61	23	3.5	6.8	19	7600
13	N	1.6	910	550	120	13	1	1	0	0.07	19	5.2	12	9.3	1500
14	N	0.97	1800	450	14	9.3	1	1	1	0.95	11	1.9	0.9	21	3700
15	N	0.37	4600	300	54	2.5	0	0	0	0	14	3.8	3.2	21	6800
16	N	0.78	830	520	32	0.9	1	1	1	0.23	14	2.7	3.9	18	4000
17	N	1.4	2500	950	26	3.9	1	1	0	0.63	7.3	1	3.7	14	3000
18	N	1.4	4400	500	48	3.6	0	0	1	0.23	16	4.1	3.7	28	6600
19	N	3	4600	930	48	4	1	1	0	0.22	22	3.6	4.7	14	7800
20	N	0.66	3700	760	31	1.3	0	1	0	0.17	21	4.2	1.7	30	6700
21	N	0.6	2300	330	190	9.8	0	0	1	0	41	5.9	5	6.1	6900

Lake ID	Area (N= Northern, S=Southern)	Surface area (ha)	Linear distance to saltwater (m)	Perimeter (m)i	Lake elevation (m)	Lake depth (m)	Presence of inlet stream (s) (bin)	Presence of outlet stream (s) (bin)	Lake substrate (bin)	Wetlands coverage (%)	Maximum outlet stream gradient (%)	Minimum outlet stream gradient (%)	Slope in lake surrounding area (100 m buffer)	Slope in lake surrounding area (1000 m buffer)	Length of outlet stream (m)
22	N	0.96	1700	530	12	1.1	1	1	0	0.51	3.5	0.58	2.8	46	2100
23	N	1.1	610	510	9.8	4.1	0	1	0	0.55	3.5	1.4	1.3	45	880
24	N	0.7	1200	400	29	0.7	1	1	0	0.55	6.5	2.1	7.4	45	1300
25	N	0.54	660	500	10	1.7	1	1	0	0.32	7.9	3.8	2.3	48	750
26	N	0.18	3300	250	490	0.9	1	1	0	0.5	32	7.5	8.3	36	11000
27	N	0.4	4800	250	500	5.4	0	0	1	0.97	30	6.4	2	32	7200
28	N	0.71	5000	360	880	9.8	1	1	0	0	60	22	7.9	52	7000
29	N	1.6	630	570	240	6.1	1	1	0	0.44	52	18	9.1	31	900
30	N	0.96	4000	640	520	1.8	0	0	1	0.96	18	6.6	1.7	24	4900
31	N	0.72	5600	370	810	1.8	1	1	0	0	56	11	7.8	18	7700
32	N	0.55	180	400	23	3.2	0	0	0	0.21	2	1	2	45	170
33	N	1.2	3200	500	390	2.9	0	1	1	0.47	23	5.1	5.7	29	12000
34	N	1.6	970	860	370	5.6	0	1	1	0.59	62	16	7.8	5	1800
35	N	2.4	7700	730	880	28	1	1	0	0.08	85	8.5	15	2.7	9800
36	N	1.2	5500	410	980	22	1	1	0	0	57	8.8	10	24	9000
37	S	1.2	2900	580	60	3	0	0	1	0.84	14	3.1	2.9	18	6500
38	S	1	200	590	11	1.4	1	1	1	0.76	2.1	1.6	6.3	12	54
39	S	2	860	1000	110	3.2	1	1	0	0.18	15	9.2	14	28	490
40	S	1.1	1500	590	80	4.7	1	1	1	0.75	20	5.4	12	27	3100
41	S	0.27	10000	240	150	1	0	1	1	0.66	22	4.5	10	48	12000
42	S	0.65	1500	500	150	1.3	0	1	1	0.82	20	6.7	11	23	2300
43	S	2.5	5100	660	200	5.1	0	1	1	0.95	18	5.7	12	32	6100
44	S	1.5	1300	560	310	1.7	1	1	1	0.62	40	6.6	12	34	4300
45	S	2.6	6100	830	270	4.2	0	1	0	0.74	36	3.8	10	37	8000
46	S	5	5300	990	710	30	1	1	0	0.25	46	7.4	22	51	6900
47	S	0.46	4700	390	540	1.1	1	1	0	0.94	34	7.3	11	45	9800

Lake ID	Area (N= Northern, S=Southern)	Surface area (ha)	Linear distance to saltwater (m)	Perimeter (m)i	Lake elevation (m)	Lake depth (m)	Presence of inlet stream (s) (bin)	Presence of outlet stream (s) (bin)	Lake substrate (bin)	Wetlands coverage (%)	Maximum outlet stream gradient (%)	Minimum outlet stream gradient (%)	Slope in lake surrounding area (100 m buffer)	Slope in lake surrounding area (1000 m buffer)	Length of outlet stream (m)
48	S	1.9	1300	650	63	5	1	1	1	0.4	19	6	12	24	1700
49	S	1.4	900	700	78	4.3	1	1	1	0.59	18	4.8	7.6	24	2300
50	S	0.76	1400	640	230	4.2	0	1	0	0.39	24	6.9	12	35	2800
51	S	2	2300	1400	180	3.9	1	1	1	0.41	14	4.3	9.1	27	4100
52	S	2.2	2700	670	370	7	0	1	0	0.57	23	6.2	22	35	4100
53	S	0.49	1300	370	140	4.6	0	0	0	0.84	12	4.7	19	25	1900
54	S	0.85	2100	440	290	2.5	0	1	1	0.98	27	6.1	9.7	31	2600

Appendix B, Table 2. Geographic location and fish species occurrence data for the 54 lakes sampled (Datum: NAD27 Alaska).

Lake ID	Area (N= Northern, S=Southern)	Latitude	Longitude	Stickleback	Dolly Varden	Cutthroat trout	Coho salmon
1	N	58.85113	134.97701	0	0	0	0
2	N	58.84623	134.97547	0	0	0	0
3	N	58.76219	134.84351	0	0	0	0
4	N	58.66487	134.96667	1	0	0	0
5	N	58.61602	134.90784	1	0	0	0
6	N	58.60659	134.87320	1	0	1	0
7	N	58.56970	134.75455	0	0	0	0
8	N	58.57091	133.64419	1	0	0	1
9	N	58.56356	133.85258	0	0	0	0
10	N	58.55566	134.69362	0	0	0	0
11	N	58.54842	134.73974	1	1	1	0
12	N	58.55096	133.83959	1	0	1	1
13	N	58.54604	133.73241	1	0	1	0
14	N	58.54410	133.80276	1	0	1	1
15	N	58.53585	134.73787	0	0	0	0
16	N	58.53614	133.78133	1	0	0	0
17	N	58.53685	133.63232	1	0	1	1
18	N	58.52727	134.73171	0	0	0	0
19	N	58.52724	134.72349	0	1	0	0
20	N	58.52287	134.73813	0	0	0	0
21	N	58.51970	134.77807	0	0	0	0
22	N	58.51193	133.99791	0	1	1	1
23	N	58.50884	133.77488	1	0	0	0
24	N	58.50404	133.84351	1	1	0	1
25	N	58.49262	134.00204	1	0	0	0
26	N	58.43954	134.70577	0	0	0	0
27	N	58.43504	134.67720	0	0	0	0
28	N	58.42853	134.44958	0	0	0	0
29	N	58.42624	134.74527	0	0	0	0
30	N	58.42414	134.68690	0	0	0	0
31	N	58.42208	134.43897	0	0	0	0
32	N	58.41656	134.55546	1	0	0	1
33	N	58.41081	134.70323	0	0	0	0
34	N	58.40694	134.73875	0	0	0	0
35	N	58.38375	134.38856	0	0	0	0
36	N	58.32747	134.32481	0	0	0	0
37	S	55.82679	131.52764	0	0	0	0

Lake ID	Area (N= Northern, S=Southern)	Latitude	Longitude	Stickleback	Dolly Varden	Cutthroat trout	Coho salmon
38	S	55.77049	131.50736	0	0	0	0
39	S	55.72778	131.69619	1	0	0	0
40	S	55.70022	131.35457	1	0	1	0
41	S	55.69788	131.38644	0	0	0	0
42	S	55.67220	131.53595	0	0	0	0
43	S	55.65535	131.19242	0	0	0	0
44	S	55.60157	131.44775	0	0	0	0
45	S	55.57787	131.57127	0	0	0	0
46	S	55.57564	131.61188	0	0	0	0
47	S	55.52316	131.41537	0	0	0	0
48	S	55.51740	131.40119	1	0	1	0
49	S	55.51304	131.64003	1	0	1	0
50	S	55.41368	131.21427	0	1	0	0
51	S	55.40265	131.18078	0	1	0	0
52	S	55.32242	131.34907	0	0	1	0
53	S	55.28863	131.39809	0	0	0	0
54	S	55.24275	131.30816	0	0	0	0