

**Effects of Sampling Sufficiency and Model Selection on Predicting the Occurrence of  
Stream Fish Species at Large Spatial Extents**

By

Kirk L. Krueger

Dissertation submitted to the Faculty of  
Virginia Polytechnic Institute and State University  
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Fisheries and Wildlife Sciences

Paul L. Angermeier, Chair

C. Andrew Dolloff

Patricia A. Flebbe

Donald J. Orth

Eric P. Smith

January 23, 2009

Blacksburg, Virginia

**Keywords:** predicting occurrence, sampling sufficiency, stream fish, multiple logistic regression, artificial neural network, classification tree, spatial extent, resolution

# **Effects of Sampling Sufficiency and Model Selection on Predicting the Occurrence of Stream Fish Species at Large Spatial Extents**

Kirk L. Krueger

(ABSTRACT)

Knowledge of species occurrence is a prerequisite for efficient and effective conservation and management. Unfortunately, knowledge of species occurrence is usually insufficient, so models that use environmental predictors and species occurrence records are used to predict species occurrence. Predicting the occurrence of stream fishes is often difficult because sampling data insufficiently describe species occurrence and important environmental conditions and predictive models insufficiently describe relations between species and environmental conditions. This dissertation 1) examines the sufficiency of fish species occurrence records at four spatial extents in Virginia, 2) compares modeling methods for predicting stream fish occurrence, and 3) assesses relations between species traits and model prediction characteristics.

The sufficiency of sampling is infrequently addressed at the large spatial extents at which many management and conservation actions take place. In the first chapter of this dissertation I examine factors that determine the sufficiency of sampling to describe stream fish species richness at four spatial extents across Virginia using sampling simulations. Few regions of Virginia are sufficiently sampled, portending difficulty in accurately predicting fish species occurrence in most regions. The sufficient number of samples is often large and varies among regions and spatial scales, but it can be substantially reduced by reducing errors of sampling omission and increasing the spatial coverage of samples.

Many methods are used to predict species occurrence. In the second chapter of this dissertation I compare the accuracy of the predictions of occurrence of seven species in each of three regions using linear discriminant function, generalized linear, classification tree, and artificial neural network statistical models. I also assess the efficacy of stream classification methods for predicting species occurrence. No modeling method proved distinctly superior. Species occurrence data and predictor data quality and quantity limited the success of predictions of stream fish occurrence for all methods. How predictive models are built and applied may be more important than the statistical method used.

The accuracy, generality (transferability), and resolution of predictions of species occurrence vary among species. The ability to anticipate and understand variation in prediction characteristics among species can facilitate the proper application of predictions of species occurrence. In the third chapter of this dissertation I describe some conservation implications of relations between predicted occurrence characteristics and species traits for fishes in the upper Tennessee River drainage. Usually weak relations and variation in the strength and direction of relations among families precludes the accurate prediction of predicted occurrence characteristics. Most predictions of species occurrence have insufficient accuracy and resolution to guide conservation decisions at fine spatial grains. Comparison of my results with alternative model predictions and the results of many models described in peer-reviewed journals suggests that this is a common problem. Predictions of species occurrence should be rigorously assessed and cautiously applied to conservation problems. Collectively, the three chapters of this dissertation demonstrate some important limitations of models that are used to predict species occurrence. Model predictions of species occurrence are often used in lieu of sufficient species occurrence data. However, regardless of the method used to predict species occurrence most predictions have relatively low accuracy, generality and resolution. Model predictions of species occurrence can facilitate management and conservation, but they should be rigorously assessed and applied cautiously.

## **Acknowledgments**

I can never compensate my wife, Heather Muehler for her patience, understanding, and help. I cannot overstate her amazing commitment, patience, support, and forgiveness. Her boundless energy, intensity, and drive are only matched by her compassion.

I also thank Paul Angermeier, whose insightful advice and steady encouragement made this work possible. Paul and my committee members Andy Dolloff, Pat Flebbe, Don Orth, and Eric Smith were always available to provide invaluable direction, encouragement and challenges. They continue to provide exceptional examples of the conduct of science.

Laura Delarche Roghair and Scott Klopfer provided GIS and database assistance that made this work possible. Conversations with Travis Brenden, Tim Copeland, Darroch Whitaker and Scott Klopfer taught me as much as any formal coursework. John Copeland's dedication to science, curiosity, patience and good humor has supported my efforts and serve as an example for other natural resource managers. Tim Quinn and Dave Price have supported my efforts and continue to encourage my progress as a scientist.

I thank all the people who have supported my efforts and endured my curiosity. Notably, my teachers Wayne Hubert, Thomas Collins, John Peterka, Todd Koel, and Donna Bruns-Stockrahm and my parents Donnel and Ardella, sisters Suzanne and Carissa, and friends Clayton and Michelle Derby have consistently encouraged my curiosity and tolerated my critical assessments.

I also thank the Virginia Department of Game and Inland Fisheries and the United States Geological Survey for providing funding for much of this work.

## Table of Contents

General Introduction.....	1
Chapter 1. Estimating sufficient sampling effort for large-scale stream fish surveys.....	4
INTRODUCTION.....	4
METHODS.....	6
Study Regions and Environmental Attributes.....	6
Data Collection.....	8
Sampling Simulations.....	9
Sampling Sufficiency Assessment.....	10
Predicting Sufficient Sampling Effort.....	10
Assessing Sampling Strategies.....	10
RESULTS.....	12
Study Regions and Environmental Attributes.....	12
Data Collection.....	12
Sampling Simulations.....	13
Sampling Sufficiency Assessment.....	13
Predicting Sufficient Sampling Effort.....	14
Assessing Sampling Strategies.....	14
DISCUSSION.....	15
Data Attributes.....	15
Sampling Simulations.....	17
Sampling Sufficiency Assessment.....	18
Predicting Sufficient Sampling Effort.....	19
Assessing Sampling Strategies.....	21
CONCLUSIONS.....	22
ACKNOWLEDGMENTS.....	23
LITERATURE CITED.....	24
Chapter 2. Using remotely sensed data to predict occurrence of fish species in stream reaches of Virginia: a comparison of modeling methods.....	40
INTRODUCTION.....	40
METHODS.....	45
Region and Species Selection.....	45
Variable Selection.....	46
Modeling Methods.....	47
Model Evaluation.....	48
Modeling Effort.....	48
Sample Location.....	49
Critical Value Selection.....	49
Prediction Performance.....	49
Model Complexity.....	50
RESULTS.....	50
Variable Selection.....	50
Model Evaluation and Comparison.....	51
Modeling Effort.....	51
Sample Location.....	51
Critical Value Selection.....	52
Prediction Performance.....	53
Model Complexity.....	53

DISCUSSION .....	54
Variable Selection and Assessment.....	54
Model Evaluation and Comparison .....	55
Modeling Effort.....	55
Sample Location.....	55
Critical Value Selection.....	56
Prediction Performance .....	56
Model Complexity.....	57
CONCLUSIONS .....	59
ACKNOWLEDGMENTS.....	60
LITERATURE CITED.....	61
Chapter 3. Conservation implications of relations among the accuracy, generality and resolution of predictions of occurrence and species traits for fishes in the upper Tennessee River drainage.	76
INTRODUCTION .....	76
METHODS.....	80
Study Area.....	80
Predicting Fish Occurrence .....	80
Fish Occurrence Data .....	80
Fish Occurrence Predictors.....	81
Model Complexity and Selection .....	82
Model Complexity and Occurrence Characteristics .....	83
Relations Between Prediction Characteristics and Species Traits.....	84
System-wide Predictions and Assessment.....	84
RESULTS.....	85
Model Complexity and Selection .....	85
Model Complexity and Occurrence Characteristics .....	86
Relations Between Prediction Accuracy, Generality, and Resolution and Species Traits ....	87
System-wide Prediction and Assessment .....	88
DISCUSSION .....	89
Model Complexity and Selection .....	90
Model Complexity and Prediction Accuracy and Resolution .....	90
Relations Between Prediction Accuracy, Generality and Resolution and Species Traits ....	95
System-wide Prediction and Assessment .....	98
CONCLUSIONS .....	101
ACKNOWLEDGMENTS.....	102
LITERATURE CITED.....	103
General Conclusions.....	129

## List of Tables

Table 1.1. Survey region name, USGS hydrologic unit code (HUC), available number of samples (N), estimated sufficient sample size ( $N_{95}$ ), slope of the results of sampling simulations ( $b$ ), observed species richness (SR obs.), estimated species richness (SR est.), percent of species that are rare (%Rare; species in  $\leq 5\%$  of samples), reach diversity (RD; number of reach types), and percent of reach types sampled (%RD) for 84 regions at four spatial scales in Virginia where sampling sufficiency assessments were performed. .... 29

Table 1.2. Spearman correlation coefficients and p-values that are bootstrap-adjusted for multiple comparisons between mean number of samples needed to collect 95% of estimated species richness ( $N_{95}$ ), area, number of reach types (RD), percent of RD sampled (%RD), observed species richness (SR), and percent of species that are rare (% Rare) in 10 drainages (D), 20 drainage-physiographies (D-P), 26 subdrainages (SD) and 28 subdrainage-physiographies (SD-P) in Virginia..... 33

Table 2.1. Spearman’s rho coefficients describing correlation between watershed and stream-reach descriptors used to predict fish occurrence in the Holston-Ridge and Valley (N=972 reaches), Roanoke-Piedmont (N=1471 reaches) and James-Coastal Plain (N=614 reaches) regions in Virginia. Descriptors include stream order (Strahler), mean reach elevation (relev), reach slope (rslope), median watershed aspect (wasp), mean watershed elevation (welev), watershed area (area), watershed shape parameter (shape), drainage density (dens), and average watershed slope (wslope). All correlations are significant ( $< 0.05$ ) due to large sample sizes..... 66

Table 2.2. Principal component analysis (PCA) eigenvalue, proportion of variance explained, cumulative variance explained, and factor loading for the first 3 principal components calculated for stream reaches in the Holston-Ridge and Valley, Roanoke-Piedmont, and James-Coastal Plain regions in Virginia. Predictors include stream order (Strahler), mean reach elevation (relev), reach slope (rslope), median watershed aspect (wasp), mean watershed elevation (welev), watershed area (area), watershed shape parameter (shape), drainage density (dens), and average watershed slope (wslope). Important axes for each PCA and important predictors for each axis are in bold. .... 67

Table 2.3. Classification performance of linear discriminant functions, generalized linear models, classification trees, and artificial neural networks built for 7 stream fish species in each of 3 drainage-physiographic regions of Virginia [Holston-Ridge and Valley (N=138), Roanoke-Piedmont (N=262), and James-Coastal Plain (N=80). Common and scientific name, prevalence (Prev.), sensitivity (SE), specificity (SP) and correct classification (CC) are given for cross-validation errors. Kappa Z-statistic and associated p-values are provided to indicate whether classification success was greater than random. Kappa statistic was not applicable (n/a) when presence was not predicted. .... 68

Table 2.4. Slope estimate ( $b$ ), F-statistic, p-value, and  $R^2$  for linear regressions of the effect of prevalence on prediction sensitivity (SE), specificity (SP), and correct classification (CC) cross-validation error rates for discriminant function analysis (DFA), generalized linear model (GLM), classification tree (TRE) and artificial neural network (ANN) for 7 fish species in each of 3 regions in Virginia. .... 71

Table 2.5. A. Mean and standard deviation (in parentheses) cross-validation sensitivity (SE), specificity (SP) and correct classification (CC) of discriminant function (DFA), generalized linear model (GLM), classification tree (TRE), and artificial neural network models that predict the occurrence of 7 fish species in each of 3 regions in Virginia. B. Bootstrap-adjusted ANOVA p-

values describing significance of differences in SE, SP, and CC among DFA, GLM, TRE, and ANN in each of 3 regions in Virginia. .... 72

Table 3.1. Pearson correlations among Strahler stream order, downstream link magnitude, mean reach elevation, mean reach gradient, reach length and reach sinuosity for 18,487 reaches in 3 physiographic provinces and all physiographic provinces of the upper Tennessee River basin. All correlations were significant ( $p < 0.01$ ). .... 110

Table 3.2. Kendall's  $\tau_c$  rank correlations and probabilities between species prevalence (Prev), maximum recorded length of species (Max), and nesting behavior (Nes), and total length of predicted presence (TPL) and mean length of predicted patches of presence (MPL). Number of species compared are in parentheses and significant correlations ( $p \leq 0.05$ ) are in bold font. Identical nesting behaviors within groups precluded some comparisons. .... 111

Appendix A. Number of predictors (d.f.), likelihood ratio (l.r.) and probability, area under the receiver-operator characteristic curve (AUC), cross-validation sensitivity (SE), Specificity (SP) and total correct classification (CC), Kappa and p-value for logistic regression models that predict the occurrence of fish in the upper Tennessee River basin. For each species the model with the smallest AIC score and /or largest AUC among all models compared is described. Superscripts correspond to analytical groups in Table 3.2. .... 121

Appendix B. Statistics describing the performance of models that predict species occurrence gleaned from peer-reviewed publications. Predictions were made for > 140 species and several taxonomic groups. Some statistics describe mean prediction performance for > 1 species. N is the number of samples reported for that study. Area under the receiver operator characteristic curve (AUC), sensitivity (SE), specificity (SP), correct classification (CC), and Kappa statistics for models assessed using resubstitution, cross-validation, and validation on an independent dataset are presented. Model methods include multiple logistic regression (MLR), generalized regression analysis and spatial prediction (GRASP), multivariate discriminant analysis (MDA), artificial neural networks (ANN), classification tree (TRE), multivariate adaptive regression splines (MARS), Bayesian-MLR model (BYS), and rule-based models (RB). Entries are blank where the appropriate statistics were not published. Authors are identified by superscripts. .... 124



## List of Figures

Figure 1.1. Major drainages (USGS 6-digit hydrologic units, bold outline) and subdrainages (USGS 8-digit hydrologic units) in Virginia (A). Intersections of major drainages with the physiographic provinces in Virginia (B). .....	34
Figure 1.2. Location of the upper Tennessee River basin in the United States and the 5 major tributaries and 3 physiographic provinces of the upper Tennessee River basin. ....	35
Figure 1.3. Locations of the 2,966 stream fish assemblage samples used in sampling simulations relative to drainages, subdrainages, and physiographic provinces in Virginia. ....	36
Figure 1.4. Sampling simulation results for four subdrainages of the James River drainage in Virginia. Values of the ordinate are the mean proportion of the estimated species richness collected in 10,000 sample iterations at the sample size on the abscissa. Error bars are 95% bootstrap confidence bounds. The number of samples, observed species richness, slope of the simulation results ( <i>b</i> ) and $N_{95}$ are provided for comparison. The dashed line is at $N_{95}$ . ....	37
Figure 1.5. Relative prevalence of fish species collected in four subdrainages in the James River drainage in Virginia. Sample sizes, species richness, slope of the sampling simulation curve ( <i>b</i> ) and estimate of the number of samples required to collect 95% of the species are provided for comparison. Three letter codes based on species common names are provided. ....	38
Figure 1.6. Sampling simulation results for A) empirical sample locations and empirical occurrence data ( $N_{95} = 880$ ), B) empirical sample locations and predicted occurrence data ( $N_{95} = 815$ ), C) random samples and predicted occurrence data ( $N_{95} = 142$ ), and D) random balanced samples and predicted occurrence data ( $N_{95} = 138$ ), for $\leq 524$ samples from the upper Tennessee River basin. One trial was conducted for simulations A and B. Ten trials were conducted for simulations C and D and the mean results are shown. For all simulations 95% bootstrap confidence bounds were calculated using 10,000 sampling iterations at each sample size. The confidence bounds overlap for all simulations at all sample sizes. For clarity only confidence bounds for simulation A are shown. Dashed line is at 95% of known species richness and can be used to estimate $N_{95}$ . ....	39
Figure 2.1. Major drainages (i.e., USGS 6-digit hydrologic units; bold outline) and their intersection with 5 physiographic provinces (italics) in Virginia. Study regions are labeled in bold. The York and Rappahannock are considered separately to maintain consistency with Jenkins and Burkhead (1993). ....	73
Figure 2.2. A diagrammatic model of a one-hidden-layer, feed-forward artificial neural network design (following Olden and Jackson 2001). Predictor variables are depicted as large filled disks and are listed in Table 2.2. Number of connections and connection weights varies among models. Number of hidden nodes is determined empirically. Bias nodes act as threshold functions for hidden nodes. Output is the predicted probability of occurrence. ....	74
Figure 2.3. Mean model sensitivity (squares), specificity (circles), and total correct prediction (diamonds) of complex models (filled) built with 9 predictors and parsimonious models (unfilled) built with only statistically significant predictors based on leave-one-out cross-validation for 7 fish species in the James-Coastal Plain region, Virginia. Model types include linear discriminant function (DFA), generalized linear model (GLM), classification tree (TRE), and artificial neural network (ANN). Bars are one standard deviation. ....	75

Figure 3.1. Location of the upper Tennessee River basin and its 5 major sub-basins and 3 physiographic provinces..... 112

Figure 3.2. Mean and 95% confidence intervals for sensitivity (SE), specificity (SP), total correct classification (CC), Kappa statistic and area under the receiver operator characteristic curve (AUC) for logistic regression models built with 1 to 6 predictors to predict the occurrence of 59 stream fish in the upper Tennessee River basin. Statistics describing the prediction accuracy (unfilled) are cross-validation results for models built in the Clinch subbasin whereas those describing generality (filled) are the same models applied to the Holston subbasin. Statistics are calculated for rare (prevalence  $\leq 10\%$ ) and common (prevalence  $> 40\%$ ) species..... 113

Figure 3.3. Relationships between model complexity (number of predictors), mean length of patches (filled symbols; contiguous reaches with predicted presence), and total length of patches (unfilled symbols) for 59 species, 13 rare ( $< 10\%$  prevalence) species and 11 common ( $> 40\%$  prevalence) species of stream fish for which predictions of occurrence were made in the Clinch subbasin..... 114

Figure 3.4. The relations between species prevalence in the Clinch subbasin and prediction accuracy (A; cross-validation AUC), species prevalence in the Holston subbasin and model generality (B; AUC for models built for the Clinch and applied in the Holston), and the relation between species prevalence in the upper Tennessee River basin and mean patch length (C). Dashed lines are the regression lines for the Clinch accuracy assessment and the Holston generality assessment. The model with the largest AUC for each species is used..... 115

Figure 3.5. Relationships between species prevalence and sensitivity, specificity, correct classification, Kappa, and area under the receiver operating characteristic (AUC) prediction accuracy statistics for 120 logistic regression models that predict the occurrence of stream fishes in the upper Tennessee River basin..... 116

Figure 3.6. Relationships between species prevalence and sensitivity, specificity, correct classification, and Kappa prediction accuracy statistics for 104 reference book models that predict the occurrence of stream fishes in the upper Tennessee River basin. .... 117

Figure 3.7. Mean and 95% confidence intervals for sensitivity (SE), specificity (SP), total correct classification (CC), Kappa statistic, and area under the receiver-operator characteristic curve (AUC) for models that predict species occurrence. Multiple logistic regression models (diamonds) and models based on the opinions of experts (circles) predict the occurrence of 120 fish species in the upper Tennessee River drainage. A variety of methods were used in 14 peer reviewed publications (squares) to predict the occurrence of  $>140$  species..... 118

Figure 3.8. Prevalence of 120 fish species in the upper Tennessee River basin. Bars represent individual species. The height of each bar is the percent of 524 samples in which a species was present. Filled bars (N = 59) represent species used to assess the effect of model complexity on prediction accuracy, resolution and generality..... 119

Figure 3.9. Sampling simulation results for 524 fish occurrence samples from the upper Tennessee River basin. Values of the ordinate are the mean proportion of the estimated species richness collected in 10,000 sample iterations at the sample size on the abscissa. Error bars are 95% bootstrap confidence bounds. The dashed line represents 95% of the estimate of species richness (N = 136) that was calculated from the asymptote of the sampling simulation results. 120

## General Introduction

Effective and efficient management and conservation requires knowledge of species locations of occurrence (i.e., presence and absence). Conducting thorough censuses is rarely possible, especially for rare or cryptic species that occur discontinuously (Angermeier et al. 2002) across large spatial extents, necessitating sampling and the use of models that predict their occurrence. Predictions of species occurrence can be used to prioritize sampling and to identify gaps in species protection (e.g., gap analysis, Scott et al. 1993). Limited resources and knowledge of species occurrence make predicting species occurrence at large spatial extents and fine spatial grains desirable. Models that predict species occurrence are often built using survey data. Inadequate survey data can preclude the building of models that accurately predict species occurrence. Unfortunately, protocols for assessing sampling sufficiency that can be used to direct surveys are poorly developed and infrequently applied. The need to make predictions for large extents at fine grains also limits the quantity and quality of data commonly available for use in prediction exercises. Many methods are used to predict species occurrence (Scott et al. 2002), but few comparisons of model predictions are available to direct the selection of appropriate methods (Guisan and Zimmerman 2000) or their appropriate application. Further, the characteristics of model predictions might vary due to species traits, such as their prevalence and size. Managers need protocols to assess sampling sufficiency and to make the best use of survey data. My goals for this work are to 1) develop methods for assessing the sufficiency of sampling at large spatial extents, 2) provide guidance, based on rigorous model assessments, that facilitates the selection of appropriate methods for predicting species occurrence, and 3) and describe relations between species traits and prediction characteristics that facilitate the appropriate application of predictions of occurrence.

Insufficient sampling effort and subsequent species occurrence data contributes to poor performance of predictive models (Fausch et al. 1988, Cao et al. 2002). Sampling simulations have a long history of use in ecology (Gleason 1922, Fisher et al. 1943, Preston 1962), but assessments of sampling sufficiency are infrequently conducted at the large spatial extents at which many management and conservation decisions are made. For example, as part of its “Freshwater Initiative” The Nature Conservancy has described the conservation priority of freshwater systems at several spatial extents in much of North America. However, because available species occurrence data are assumed to insufficiently describe assemblage composition their conservation priorities are usually based on the uniqueness and diversity of environmental

conditions rather than knowledge of species occurrences. I conducted sampling simulations to assess the sufficiency of sampling to describe stream fish species richness at four spatial extents across Virginia. Additional simulations were conducted using predicted species occurrences to describe the relative importance of the spatial allocation of samples on sampling efficiency. The results of this work demonstrate the need for additional stream fish surveys in many regions of Virginia and can be used to develop efficient surveys at large spatial extents.

Descriptions of species occurrence from surveys are rarely sufficient to efficiently direct conservation and management actions, especially at large spatial extents and fine spatial grains. Models are frequently used to predict species occurrence (Scott et al. 2002) and those predictions are used to direct management and conservation actions. For example, in the state of Washington the predicted upstream extent of fish presence is used as a boundary, above which stream buffer width requirements are reduced. Poor prediction accuracy can inordinately constrain forest management or insufficiently protect fish and streams with fish present. Poor model performance (i.e., low prediction accuracy) has led to the use of alternative methods for predicting occurrence (e.g., artificial neural networks [Olden and Jackson 2001]). However, few comparisons of model performance are available to facilitate the selection of the most appropriate prediction method (Guisan and Zimmerman 2000). I compare the predictions of occurrence of five prediction methods for seven species in each of three regions of Virginia to facilitate the selection of appropriate prediction methods.

Characteristics of predictions of species occurrence vary among species and the application of model predictions should account for this variability. The characteristics of predictions of species occurrence (e.g., accuracy, generality and resolution) are expected to vary among species due to differences in their prevalence and other traits. For example, models that predict the occurrence of species that build nests might be expected to be more accurate and general than those for other species because such species may be less susceptible to disturbance or may recover more quickly to disturbance. Assessments of models that predict the occurrence of nest builders might thus have relatively few errors of omission due to absence in suitable habitats due to past disturbances. I describe relations between species traits and predicted occurrence characteristics for several species in the upper Tennessee River basin. I carefully selected the most accurate predictive models and rigorously assessed their accuracy, generality and resolution. Weak and variable relations between species traits and prediction characteristics suggest that my models perform poorly and should be applied cautiously. Comparison of my results with alternative models based on the opinions of experts and model assessments described

in peer-reviewed literature suggests that poor prediction accuracy and perhaps poor resolution are common issues.

Inaccurate and low resolution predictions of species occurrence should guide the appropriate application of predictions of species occurrence. Few predictions of species occurrence are rigorously assessed (Guisan and Zimmerman 2000, Scott et al. 2002) and those that are often appear to predict with similar accuracy. The resolution of predictions of species occurrence is rarely assessed. Inaccurate and coarse predictions of species occurrence should be expected, especially at large spatial extents. Collecting additional data may facilitate the prediction of occurrence for infrequently sampled species, but is unlikely to substantially improve the accuracy and resolution of most predictions. Alternative prediction methods usually provide similar predictions. My predictions of species occurrence provide the best information available to guide management and conservation, but they should be applied conservatively and cautiously. For example, if model predictions of species presence are used to select conservation areas, several areas should be selected to minimize the effect of errors of commission. Further, if model predictions of absence are used to justify management actions that could harm the species if it is present, those actions should occur only in areas of very high probabilities of absence and additional sampling to detect species presence may be warranted. Despite their limitations, these models provide the best information available describing the probability of occurrence of many freshwater fish species and they can be used to make sound management and conservation decisions.

## Chapter 1. Estimating sufficient sampling effort for large-scale stream fish surveys

### INTRODUCTION

Sampling sufficiency is an important concern for understanding species-habitat relations, predicting species occurrence, directing management actions, and biological monitoring of environmental quality. Insufficient sampling effort and inadequate coverage of environmental conditions contribute to failures to detect species-habitat relations and to poor performance of predictive models (Fausch et al. 1988, Leftwich et al. 1997). Inadequate species occurrence information precludes precise management and obscures the detection of changes in species occurrence. Sampling effort exceeding that necessary to meet study objectives may divert funds and efforts from other worthy projects. Sampling that is sufficient describes a region's species richness with desired accuracy and confidence. Unfortunately, protocols for assessing sampling sufficiency and predicting sufficient sampling effort are poorly developed and infrequently applied.

As knowledge of stream fish movement and resource use increases (Gowan et al. 1994, Fausch and Young 1995, Albanese et al. 2004) researchers are increasingly aware of the need to scale survey efforts to the ecological and anthropogenic processes that determine local assemblage composition (McIntyre and Hobbs 1999). As the spatial extent, cost, and conservation importance of surveys increase, so does the importance of sampling sufficiency and efficiency. While questions regarding the sufficiency of stream fish samples have been studied in many locations (e.g., Lyons 1992, Angermeier and Smogor 1995, Paller 1995, Simonson and Lyons 1995, Patton et al. 2000), these analyses have been limited to the local (stream reach) scale. Spatial variability of habitat, fish movement, and population fluctuations may render the reach insufficient as a sampling unit for many studies (Ibbotson et al. 1994, Schlosser 1995, Herger et al. 1996, Dolloff et al. 1997). Discontinuity of stream fish distributions may require many samples to sufficiently describe spatial patterns of assemblages across large regions (Angermeier et al. 2002).

Sampling simulations have a long history of use in ecology (Gleason 1922, Fisher et al. 1943, Preston 1962) to estimate species richness (SR; Connor and McCoy 1979, Coleman 1981, Miller and Wiegert 1989, Mingoti and Meeden 1992) and species diversity (Flather 1996, Ney-Nifle and Mangel 1999), but they are infrequently used to assess the sufficiency of sampling to address specific research questions (but see Cao et al. 2002 and Cook et al. 2004) or to plan surveys. Several researchers have provided guidance for the proper use of sampling simulations. In particular, extensive extrapolation has been cautioned against (Smith and van Belle 1984,

Palmer 1990, Bunge and Fitzpatrick 1993, Hellman and Fowler 1999) and comparing alternative functions to fit simulation results has been recommended (Flather 1996). However, little work has been done to assess the sufficiency of sampling of a given taxonomic group, including stream fishes, in large regions.

Sampling sufficiency assessments for stream fishes have primarily estimated the optimal length of a stream reach to sample (e.g., Lyons 1992, Angermeier and Smogor 1995, Paller 1995, Simonson and Lyons 1995, Patton et al. 2000). The results of such studies guide many surveys. For example, the Oregon Department of Fish and Wildlife used  $\geq 20$  stream widths per site to sample fishes for the Oregon Coastal Coho Assessment (<http://nrimp.dfw.state.or.us/OregonPlan>, Stevens 2002), the US Environmental Protection Agency (EPA) has suggested a minimum of 40 stream widths for sampling stream habitat and fishes for their Environmental Monitoring and Assessment Program (EMAP; <http://www.epa.gov/emap/html/docs.html>), and the Virginia Department of Game and Inland Fisheries (VDGIF) used a minimum of 20 stream widths to sample fishes and habitat for their Statewide Stream Survey (J. Copeland, VDGIF, personal communication). Researchers are increasingly acknowledging the need to survey and manage streams and fish assemblages at large spatial extents (Palmer and Poff 1997, Fausch et al. 2002, Ward et al. 2002, Wiens 2002), but little guidance is available regarding the effort necessary to sufficiently sample stream fishes at large scales.

Survey planning and prioritization would benefit from accurate a priori estimates of sufficient sampling effort. The commonly observed positive correlations between area sampled, environmental heterogeneity, and species richness (e.g., Kerr and Packer 1997, MacNally and Watson 1997, Fraser 1998, Guegan and Oberdorff 1998, Matthews and Robinson 1998) may allow prediction of sufficient sampling effort. Precise and consistent correlations between survey area, environmental heterogeneity, and sufficient sampling effort will allow prediction of sufficient sampling effort that will facilitate survey planning and prioritization if species occurrence data are not available.

Available resources often limit sampling effort, making improvement in survey efficiency (i.e., the rate of species accrual) a priority. Survey efficiency varies among survey strategies (e.g., spatially balanced random sampling; Stevens and Olsen 2004), providing similar information at different sampling efforts. Sampling simulations can be used to assess the efficacy of different sampling designs (Dale et al. 1991, Hirzel and Guisan 2002). Using realistic species distributions in simulations can provide realistic results of different sampling strategies that allow researchers to design, compare, and implement efficient surveys. Unfortunately, managers must

often use “found” data that are collected for many reasons to estimate species distributions and assess sampling sufficiency.

Virginia provides excellent opportunities to refine methods of assessing sampling sufficiency. The nearly 200 species of native freshwater fishes vary widely in prevalence (frequency of occurrence) and occur in a range of assemblages that enable informative comparisons of the effects of assemblage attributes on sufficient sampling effort. The exceptional diversity of stream fishes in Virginia is largely due to the 11 major drainages and 5 physiographic provinces represented in the state (Jenkins and Burkhead 1993, Figure 1.1). These geographic units provide easily identified, ecologically meaningful survey regions. Subsections (constituents) of these units also provide meaningful survey and management units (Angermeier and Winston 1999). Comparisons of sampling-sufficiency simulation results among regions and spatial scales enhance the generality of my work. Virginia’s stream fishes are extensively sampled and the VDGIF maintains the Virginia Fish and Wildlife Information Service database (VAFWIS; <http://128.173.240.58/build02/fwis/default.asp>), which provides access to many fish sample data, thereby allowing sampling sufficiency assessments for several large regions. The high species richness (SR), frequent imperilment, and intensive management of stream fishes in Virginia require large-scale surveys to provide information necessary for precise management. Limited resources promote the development of protocols for assessing the sufficiency and improving the efficiency of large-scale stream fish surveys.

This work was initiated to answer questions the VDGIF posed regarding the sampling sufficiency and efficiency of stream fish surveys. My goal is to use multipurpose stream fish occurrence samples (found data), environmental attributes, and sampling simulations to assess sampling sufficiency and to identify ways to improve survey efficiency. Specific objectives include 1) assessing the usefulness of existing multipurpose data and providing advice on its proper collection and storage, 2) estimating sufficient sampling effort using multipurpose data and sampling simulations, 3) assessing the efficacy of predicting sufficient sampling effort using correlations between survey region environmental conditions and estimates of sufficient sampling effort, and 4) using sampling simulations and spatially realistic simulated assemblages to demonstrate the effect of different survey strategies on achieving sufficient sampling effort.

## METHODS

### *Study Regions and Environmental Attributes*

I conducted sampling sufficiency analyses for regions defined by their hydrography and physiography at four spatial scales to enhance the generality of the results (Figure 1.1). Drainage



regions (D) are 6-digit United States Geological Survey (USGS) hydrologic units (Seaber et al. 1987), drainage-physiography regions (D-P) are intersections of D with physiographic provinces defined by Jenkins and Burkhead (1993). Subdrainage regions (SD) are 8-digit USGS hydrologic units (Seaber et al. 1987) and subdrainage-physiography regions (SD-P) are intersections of SD with physiographic provinces. Very small SD-P were concatenated into their larger conterminous SD-P. There are 10 D, 5 physiographic provinces, 22 D-P, 40 SD, and 53 SD-P in VA. These regions are useful survey units because they have ecologically meaningful boundaries (Holling 1992), distinct fish assemblages (Jenkins and Burkhead 1993, Angermeier and Winston 1999), and different factors regulate species distributions (Angermeier and Winston 1998, 1999, Jackson et al. 2001).

The USGS National Hydrography Dataset (NHD; USGS 2000) 1:100,000-scale stream reach files were used to define streams. Reaches were identified using the unique codes assigned by the NHD. The NHD defined reaches based on the location of confluences, major changes in stream gradient, and changes in channel type (e.g., natural or channelized). Reach and watershed attributes were calculated to describe environmental heterogeneity and to assess sample coverage of environmental conditions. These data were used to assess the efficacy of predicting sufficient sampling effort using regional environmental conditions. The National Elevation Dataset (NED; USGS 2000 <http://gisdata.usgs.gov/ned/default.asp>) was used to delineate the contributing watershed for each reach, and to calculate elevation and slope variables. Mean reach elevation, reach slope, watershed aspect, and stream order (Strahler 1957) were calculated to describe reaches. Calculations were made using Arc/Info 8.0.2 and ArcGIS 9.1 (Environmental Systems Research Institute, Inc., <http://www.esri.com>). The number of reach types in each region was correlated with the estimate of sufficient sampling effort to assess the efficacy of predicting sampling sufficiency using environmental conditions.

To assess sample coverage of environmental conditions, each stream reach was assigned a reach type based on three classes each of reach elevation and reach slope, four classes of watershed aspect and six stream orders. Elevation and slope categories were defined using 33- and 66-percentiles of the values for all reaches in Virginia. Aspect classes were constructed to identify north (N), south (S), east (E) and west (W) flowing watersheds. Strahler orders 1 through 6 were used. The percent of reach types sampled in a region was used to assess coverage of environmental variability. Additionally, the number of reach types in a region (reach diversity; RD) was used to index environmental heterogeneity. The percent of reach types sampled in each region was correlated with the estimate of sampling sufficiency to assess the effect of sample coverage on estimates of sampling sufficiency.

### *Data Collection*

Species occurrence records were procured from the VAFWIS, which includes georeferenced species occurrence records from collection permit reports, scientific reports, and citizen reports. The > 20,000 spatially and temporally extensive collection records in the VAFWIS provide exceptional opportunities to explore patterns of species occurrence and the effects of sampling effort. The VAFWIS includes most of the scientific collection records for one of the most species rich and well-studied fish assemblages in North America (Jenkins and Burkhead 1993). Sample location errors are likely in these data because standard mapping methods and maps are not required. Errors of sampling omission (i.e., false absence) are likely in these data and may be due to failure to sample the appropriate spatial or temporal strata, inadequate sampling effort, or ineffective or inappropriate methods (Angermeier et al. 2002). Errors of sampling commission (false presence) are less likely, but may occur due to species misidentification. Poor sample coverage of the range of environmental variability and repeated samples of the same reaches are also likely in these data because few were collected as part of systematic surveys.

The results of sampling simulations to predict sufficient sampling effort are affected by the quality of the data used in the simulations. Conducting sampling simulations with samples that are not representative of a region's assemblage biases results. Omitting samples with many sampling omission errors minimized the effect of sampling errors on sufficiency assessments and improved the accuracy of subsequent analyses. This was accomplished by using only samples that appeared relatively complete (hereafter, assemblage samples). Assemblage samples were used to realistically simulate surveys. Samples that recorded  $\geq 3$  species, including  $\geq 1$  non-game species, and had reliable location information (i.e., plotted within 50 m of a stream reach) were defined as assemblage samples and assumed to have few omission errors. When possible, the sample collector and collection methods were identified to ensure adequate sampling. When multiple samples of a reach were available the most recent sample was used. Samples were collected from 1965 to 1998. Samples were attributed to individual stream reaches using a 50-m buffer on each side of the reach in a GIS. Samples were omitted from analyses if they could not be confidently attributed to a reach. When identifiable, samples from non-wadeable rivers were omitted from analyses because they were not comparable due to expected differences in species catchability and the gear used (Hubert 1996). These criteria removed many samples with many likely errors of omission, but also may have removed samples that accurately described depauperate assemblages. Despite errors, these databases contain the most comprehensive information available describing fish distributions in Virginia.

### *Sampling Simulations*

I simulated sampling by bootstrap sampling without replacement (Efron and Tibshirani 1994, Manley 1997) at a range of sampling efforts. Bootstrapping randomized samples and smoothed the resulting species accumulation curves. Species accumulation curves estimate SR in each region and the rate of information accrual (via the slope parameter “ $b$ ”), which assess sampling efficiency. Although other simulation methods are available to estimate species richness and to assess sampling sufficiency (Hellman and Fowler 1999), I used bootstrapping because it allows simple calculation of confidence bounds that aid in interpretation of results and because bootstrapping tools are readily available (Gotelli and Colwell 2001). In each region sampling simulations were conducted for 11 to 15 sample sizes, including one sample and the total number of samples available for the region. Consistent simulation results were obtained with 10,000 bootstrap iterations at each sample size.

A minimum number of assemblage samples are required for meaningful sampling simulations, but sample size criteria have not been established. Sampling simulations have been used to assess sampling sufficiency using relatively small datasets. Each sampling simulation used  $\geq 30$  samples to minimize extrapolation of SR beyond that observed. Thirty samples included a mean of 78.4% (s.d. = 6.50) of species known present in the largest, most species-rich regions (i.e., the 10 D), exceeding the 63% to 69% threshold under which Hellman and Fowler (1999) found bootstrapping underestimated SR and the approximately 50% threshold suggested by Smith and van Belle (1984).

Predicting SR from sampling simulations and comparing predictions among regions required fitting a mathematical function to disparate simulation results. Simulation results from every region were fit with a Weibull function:

$$N = a \left[ 1 - e^{-\{b(x-c)\}^d} \right]$$

and a negative exponential function:

$$N = a(1 - e^{-bx}),$$

where  $N$  is the number of species,  $a$  is the asymptote,  $b$  is the slope (rate of species accumulation or information accrual),  $x$  is accumulated samples,  $c$  is a scaling factor that determines the  $x$ -intercept, and  $d$  is a shape parameter (Flather 1996). Akaike’s information criteria (AIC; Akaike 1974, Burnham and Anderson 2002), coefficient of determination ( $R^2$ ), and plots of residuals for all regions were used to assess the goodness-of-fit of the functions. Solving for  $a$  estimated SR. Estimator bias was assessed for D by comparing estimates of SR with known SR from Jenkins and Burkhead (1993).

Many statistics can be used to assess sampling sufficiency, but few are recommended for large-scale surveys. For example, Moreno and Halfpeter (2000) suggest collecting a mean of 90% of the species in a region and Angermeier and Smogor (1995) suggest collecting 90% of the species in a reach, whereas Cao et al. (2001) suggest using a sample similarity statistic that is calculated with sampling simulations. I used the number of samples required to collect a mean of 95% of the species estimated to be present in the region ( $N_{95}$ ) as a measure of sampling sufficiency because it is simple to interpret, particularly with reference to rare species, and because it encourages the collection of many relatively rare species.

#### *Sampling Sufficiency Assessment*

Sampling simulations were conducted for all regions with  $\geq 30$  assemblage samples. Sampling sufficiency was assessed by comparing  $N_{95}$  with the number of assemblage samples available for each region and by visual assessment of bootstrap confidence bounds. Sampling was considered sufficient if the number of available samples was  $\geq N_{95}$ . Bootstrap confidence bounds described likely sampling omissions. The slope of the sampling simulation curve ( $b$ ) was used to index the rate of information accrual, that is, the rate at which additional species were added to a region's list with additional samples.

#### *Predicting Sufficient Sampling Effort*

Strong and consistent relationships between area (ha), RD, proportion of reach types sampled (%RD), SR, and the proportion of species that are rare (%Rare; species present in  $\leq 5\%$  of samples) and  $N_{95}$  may allow prediction of  $N_{95}$  and better understanding of differences in sufficient sampling effort among regions and scales. The strength and consistency of the correlations between area, RD and  $N_{95}$  were assessed with Spearman correlations for all regions and scales with sufficient samples to calculate  $N_{95}$ . Tests of correlation significance were bootstrap-adjusted for multiple comparisons. Sampling simulation results and charts of species prevalence for the Buffalo, Rivanna, Chickahominy, and Appomattox subdrainages of the James River drainage are contrasted to describe variation in the effects of SR, prevalence, and the distribution of species among samples on  $N_{95}$  and  $b$ . These SD provide interesting contrasts because they demonstrate the range of results for  $N_{95}$  and  $b$  that are possible for regions with similar SR, number of samples, and species prevalence.

#### *Assessing Sampling Strategies*

Comparing sampling strategies requires fish distribution data for entire assemblages. Realistic distribution data can provide realistic simulations and comparisons. Complete censuses of stream fish distributions in large regions are rarely possible, but predicted occurrences may approximate real distributions. I used a spatially realistic, simulated stream fish assemblage of

the upper Tennessee River basin (UTRB, Figure 1.2) to assess the effect of four survey strategies on  $N_{95}$ . The simulated fish assemblage of the UTRB was used because it includes one of the most species-rich and frequently sampled SD in VA. The boundary of the UTRB is also more ecologically meaningful than the VA state boundary, making models more likely to predict accurately. Furthermore, predictions of stream fish occurrence in the UTRB were available from a concurrent gap analysis.

The simulated assemblage was constructed by predicting the occurrence of stream fishes in all reaches of the UTRB based on logistic regression models. Assemblage data were procured from existing databases from Virginia, Tennessee, and North Carolina. Assemblage samples were selected identically to those used for the VA sampling sufficiency assessments. Predictor variables included Strahler (1957) stream order, reach length, downstream link (Osborne and Wiley 1992, Smith and Kraft 2005), mean reach elevation, reach gradient, physiography, reach sinuosity (Gordon et al. 1992), state, and path type (i.e., natural or altered via USGS NHD). Dummy variables for state and physiography (Dunham and Vinyard 1997) accounted for differences in sampling methods and intensity. Model predictions were assessed using cross-validation values for correct presence (sensitivity, SE), correct absence (specificity, SP), and total correct (CC) predictions, Kappa statistics (Cohen 1960, Guggenmoos-Holzmann 1996), Akaike's information criteria (AIC; Akaike 1974), and the area under the receiver-operator-characteristic curve (AUC). AUC measures a model's ability to discriminate between correct presence and absence predictions across the range of possible critical values. Model building, application, and assessment details are included in Chapter 3. Species prevalence is often correlated with statistics describing prediction accuracy without accounting for individual occurrences (Hosmer and Lemeshow 2000, Manel et al. 2001). For example, the intercept of logistic regression models partly accounts for species prevalence in the data used to fit the model. I plotted predicted occurrences for the simulated assemblage using species' prevalence as critical values to account for the effect of prevalence on prediction accuracy.

When planning surveys, scientists select a sampling strategy that they believe will effectively and efficiently collect information (e.g., species occurrences and SR). Most sampling strategies differ in how they distribute samples in space. Sampling strategies vary in efficiency (the rate of information accrual) depending on where samples are located and how species are distributed. Selecting an efficient sampling strategy may reduce the effort required to sufficiently sample large regions. Sampling strategies compared included 1) sampling the original data to use as a baseline for comparisons, 2) sampling predicted occurrences at the original sample locations to simulate the effect of decreased sampling errors, 3) randomly sampling predicted occurrences,

and 4) spatially balanced random sampling of predicted occurrences. Many of the samples in the original data were likely located for convenience (e.g., easy access from roads). Such non-random sampling does not provide an unbiased description of the region's assemblage. Random sampling provides an unbiased description of the regions assemblage, but it may be difficult to access some sites and simple random sampling does not ensure the even spatial distribution of samples. Spatially balanced random sampling provides an unbiased description of the regions assemblage and facilitates the even spatial distribution of samples. Ten trials were run for random sampling and spatially balanced random sampling and their mean results were compared. Simulations used 524 samples to facilitate comparisons of  $N_{95}$ . Comparison of  $b$  and  $N_{95}$  from different sampling strategies can demonstrate the limitations of using found data and the relative importance of reducing errors of sampling omission and using well-designed surveys. All GIS analyses were completed in Arc/Info and ArcGIS 9.1 (Environmental Systems Research Institute, [www.esri.com](http://www.esri.com)) and all statistical analyses were completed in SAS 9.1 ([www.sas.com](http://www.sas.com)).

## RESULTS

### *Study Regions and Environmental Attributes*

The number of reach types and the percent of reach types sampled varied among regions within scales and both tended to decline at smaller scales (Table 1.1). The James D is the largest of the 11 D in Virginia, crosses all five PP, and had the largest number of reach types (225) of which 44% were sampled. The Big Sandy D had the lowest number of reach types (16), but the highest percent of reach types sampled (75%). The mean number of reach types was 104 for drainages, 74 for drainage-physiographic provinces, 69 for subdrainages, and 62 for subdrainage-physiographic provinces. The mean percent of reach types sampled was 41 for drainages, 34 for drainage-physiographic provinces, 30 for subdrainages, and 30 for subdrainage-physiographic provinces. At all scales large regions and those in the Piedmont and Ridge and Valley physiographic provinces tended to have more reach types (Tables 1.1 and 1.2).

### *Data Collection*

Only 2,966 of > 20,000 samples in the VAFWIS database were retained for sampling simulations. Inaccurate location data and/or likely sampling errors rendered most samples inappropriate for sampling sufficiency simulations. Imprecise sample coordinates, recording errors, and differences in the location of streams between map and GIS data accounted for about 15% of the omissions. Approximately 80% of omitted samples did not meet the assemblage criteria of having  $\geq 3$  species, including  $\geq 1$  non-game species. Many samples list only game species (e.g., *Micropterus salmoides*) or only species of concern that are often found in species-

rich sites (e.g., *Percina rex*). Few species were present in only those samples not meeting the assemblage sample criteria or were found only in large streams. Commission errors were rarely identified and were likely due to misidentification of species. Commission errors were usually for species with congeners of similar appearance. About 3% of omitted samples were repeat samples of the same stream reach. Metadata describing sampling goals and effort, collection methods, and number of individuals collected were rarely available. Visual assessment of the locations of the remaining 2,966 samples suggests nonrandom distribution of samples at the state extent (Figure 1.3). Many samples were on 3-rd and 4-th order streams, near road crossings and spatially clustered near the cities of Roanoke and Richmond, VA.

#### *Sampling Simulations*

A minimum sampling effort is required to accurately estimate SR and  $N_{95}$ . Using similar methods Smith and Van Bell (1984) found that accurately estimating SR required sampling to collect  $\geq 50\%$  of the species in a region. Approximately 25 samples were needed to collect  $\geq 50\%$  of the species known to be present in each region. I used a conservative minimum sample size of 30 to ensure sufficient data for each simulation. Sampling simulations were conducted for 10 of 10 D, 20 of 27 D-P, 26 of 49 SD and 28 of 73 SD-P (Table 1.1). The Weibull and negative exponential functions fit simulation results well ( $R^2 > 0.95$ ) for all regions and no significant difference was found between AIC values for paired comparisons between the two functions (paired  $t = -0.511$ ,  $df = 83$ ,  $P = 0.6131$ ). Examination of residuals indicated minor additional nonrandom variability in many regions, particularly for the negative exponential function. The Weibull function appeared to better fit the tail of the species accumulation function than the negative exponential function for several regions. Hereafter, results based on the Weibull function are discussed, but conclusions would be similar for analyses based on either function. Sampling simulations consistently underestimated SR within 3% of known SR for each D (Table 1.1).

#### *Sampling Sufficiency Assessment*

Few regions of VA are sufficiently sampled but the frequency of sufficient sampling increases with the spatial extent of sampling units (Table 1.1). The number of samples available was  $> N_{95}$  for 10 of 10 D, 13 of 20 D-P, 15 of 26 SD, and 11 of 28 SD-P. Estimates of  $N_{95}$  are large for most regions, vary among regions, and decrease at smaller spatial extents (Table 1.1). Mean  $N_{95}$  is 150.1 (standard deviation = 70.3) for D, 116.2 (standard deviation = 63.7) for D-P, 106.2 (standard deviation = 81.3) for SD, and 84.3 (standard deviation = 45.2) for SD-P. Bootstrap confidence bounds are generally large and variable and decrease with sample size. The rates of information accrual ( $b$ ) are generally low but increase at smaller spatial extents. Mean  $b$

is 0.09 (standard deviation = 0.03) for D, 0.12 (standard deviation = 0.05) for D-P, 0.15 (standard deviation = 0.07) for SD and 0.16 (standard deviation = 0.10) for SD-P.

#### *Predicting Sufficient Sampling Effort*

Relationships between environmental attributes, assemblage attributes, and  $N_{95}$  are generally weak (Spearman  $< 0.6$ ) and vary among regions and spatial extents (Table 1.2). Area and  $N_{95}$  are positively correlated at all spatial extents, but the strength of the correlation decreases at smaller extents (Table 1.2). Reach diversity and  $N_{95}$  are strongly positively correlated at only the D scale. Interestingly,  $N_{95}$  is not strongly correlated with the percent of reach types sampled at any scale. Percent of species that are rare and SR are correlated with  $N_{95}$  at all scales and area and reach diversity are correlated at all scales.

The rates at which species accrued with sample size varied widely among subdrainages. The Buffalo and Rivanna subdrainages have identical SR and number of samples, but  $N_{95}$  and  $b$  differ substantially between them (Table 1.1, Figures 1.4 and 1.5), likely because the proportion of rare species was higher in the Buffalo SD (27%) than in the Rivanna SD (15%). The Chickahominy SD has been sampled frequently, but the very low proportion of common species and high proportion of rare species result in a slow rate of information accrual ( $b$ ) and a very large  $N_{95}$  (Table 1.1, Figures 1.4 and 1.5). In contrast, the Appomattox SD has been sampled less frequently and has a higher SR and proportion of rare species than the Chickahominy SD, but the Appomattox SD has a relatively high rate of information accrual and small  $N_{95}$ .

#### *Assessing Sampling Strategies*

The UTRB is insufficiently sampled with the available 524 assemblage samples.  $N_{95}$  (880) is  $>$  the number of available samples (524), sampling simulation results are not asymptotic, and bootstrap confidence bounds are large (Figure 1.6). These data were used to predict the occurrence of 120 fish species that were present in  $\geq 5$  assemblage samples. Prediction accuracy was generally low and affected by species prevalence (Chapter 3). Mean SE was 27.3, mean SP was 91.2, mean CC was 90.0 and mean Kappa was 0.22 for all 120 species. Most fishes were rare (mean prevalence = 13.3, range = 0.4 to 88.4). Species prevalence was negatively correlated with SE and positively correlated with SP.

Effects of reducing sampling errors in  $b$  and  $N_{95}$  were assessed by comparing sampling simulation results for empirical occurrences with predicted occurrences from the same locations. Reducing sampling errors increased the rate of information accrual (Figure 1.6). The rate of information accrual was relatively low for the UTRB empirical data ( $b = 0.0281$ ). Sampling the same locations with predicted occurrences increased the rate of information accrual ( $b = 0.0936$ ), but  $N_{95}$  remained large (815). Effects of alternative sampling designs (i.e., different sample



locations) on  $N_{95}$  were assessed by comparing sampling simulation results for predicted occurrences at empirical sample locations with those from random locations. Random sampling increased the rate of information accrual ( $b = 0.1018$ ) and greatly reduced  $N_{95}$  to 142 samples. Random spatially balanced sampling slightly increased the initial rate of information accrual ( $b = 0.1049$ ) from that of simple random sampling, but  $N_{95}$  remained similar (138) to that of simple random sampling. Apparently, carefully distributing samples in space can improve sampling efficiency more than reducing errors of sampling omission at individual sample sites.

## DISCUSSION

Survey data are used to describe species-habitat relationships (Poff and Allen 1995), to predict species occurrence (Scott et al. 2002), to direct management actions, and to monitor environmental quality (Norris 1995). Insufficient sampling can fail to identify ecological patterns (Cao et al. 2002), limit prediction accuracy (Leftwich et al. 1997), misdirect management efforts, and fail to detect environmental changes (Edwards 1998). Sampling sufficiency is increasingly important as the spatial extent of surveys and the application of survey results increase. The importance of estimating sampling efficiency also increases as the scale and associated costs of sampling sufficiently increase. The stream fish assemblages of most regions of VA are insufficiently sampled by the available data. Reliance on found data that were not collected using efficient sampling strategies reduces the rates of information accrual and increases the number of samples required to sufficiently sample the fish assemblages of large regions. Differences in assemblage structure among regions preclude the accurate prediction of sufficient sampling effort.

### *Data Attributes*

Sufficient sampling effort can be estimated by using found data to help plan regional surveys. The process of selecting found data highlights many of their shortcomings (Cairns and Pratt 1986, Bonar and Hubert 2002, Hayes et al. 2003). Sample location uncertainty and sampling errors rendered many found samples unsuitable for estimating sufficient sampling effort. Samples located far from streams or outside of the survey region (e.g., samples located outside VA) are easily identified as erroneous, but it is difficult to identify erroneous samples located near several streams. Many sample location errors are due to faulty map reading, data recording, or data management. Careful collection and recording of location data, including map projections and coordinate systems, can prevent many location errors. Collecting, recording and storing redundant location data from maps, global positioning systems (GPS), and geographic information systems (GIS) will reduce the frequency of location errors and should be encouraged.

Researchers should remember that maps and GIS layers are models that have errors that can make data appear erroneously located. Many samples were located far from mapped streams. Some of these errors may be due to faulty stream delineation. I attributed samples to stream reaches defined in the NHD using a GIS and a 50-m buffer on each side of stream reach centerlines. This was problematic when samples were located near stream confluences or in densely drained regions where a narrow buffer could not distinguish among multiple reaches. Accurately collecting and recording the locations of the start and end of sampling can remedy this problem. Samples should be plotted as lines that coincide with stream reaches rather than points that can be plotted on several stream reaches. Recording the stream name and/or reach code with collection data is also helpful. Efforts to improve the accuracy and completeness of hydrographic data should also be encouraged.

Sampling omission and commission errors should be minimized because they bias estimates of sufficient sampling effort (Patton et al. 2000) and stymie analyses (Cao et al. 2002). Likely errors of sampling omission were the most frequent reason that samples were not used for sampling simulations. Sufficiency of individual samples can be assessed with metadata describing sampling effort and methods (Lyons 1992, Angermeier and Smogor 1995, Paller 1995, Peterson and Rabeni 1995, Simonson and Lyons 1995, Patton et al. 2000), but such metadata were rare. My criteria defining assemblage samples removed many samples with likely omission errors, but the criteria were liberal, likely retaining many samples with omission errors that inflate estimates of  $N_{95}$ . However, some species-poor samples warrant inclusion in analyses because they record the presence of rare species. Additionally, some omitted samples may accurately describe species depauperate assemblages. Their removal would erroneously increase sampling sufficiency estimates. Sampling commission errors were infrequently detected and likely due to species misidentification.

Metadata describing sampling methods, equipment, and effort are vital to assessing the sufficiency of individual samples and may prove more useful than standardized sampling. Sampling using standard methods and equipment can provide commensurate data across large spatial extents (Bonar and Hubert 2002, Hayes et al. 2003), simplifying analyses and ensuring the usefulness of samples. However, a standard method may not be optimal for all species or sample locations (Peterson et al. 2004, Rosenberger and Dunham 2005) and is not necessary for survey planning or assessments of sampling sufficiency. Varying methods and equipment among locations may improve survey efficiency (Longino and Colwell 1997). For example, seining may be preferable to electrofishing if species of concern are likely present (Hubert 1996) or where conductivity is low, but electrofishing may prove preferable in other locations. Additionally,

sufficient sampling effort varies among sample locations (Angermeier and Smogor 1995, Patton et al. 2000) depending on local environmental conditions (Cam et al. 2002, Peterson et al. 2004, Rosenberger and Dunham 2005) and assemblage attributes such as species prevalence and detectability (Boulinier et al. 1998). Selecting the most appropriate sampling and analytical method(s) remain important topics for further research. Regardless of sampling method, metadata should be recorded to assist in estimation of the likelihood of errors of sampling omission.

### *Sampling Simulations*

Sampling simulations can be used to assess sampling sufficiency but the results are limited by the attributes of the data used. Although sampling simulations make no SR or distribution assumptions (Hellman and Fowler 1999), care should be taken to ensure that sufficient data are available to provide reliable results (Smith and van Belle 1984, Keating et al. 1998, Bunge and Fitzpatrick 1993). The minimum sample size of 30 that I used is specific to these assemblages and data. For example, Moreno and Halffter (2000) used 14 samples to assess the completeness of bat (Order Chiroptera) inventories, de Solla et al. (2005) used 50 samples to calculate sufficient sampling effort for frogs (Order Anura), Angermeier and Smogor (1995) used a minimum of 17 contiguous mesohabitat units to estimate sufficient sampling effort to describe local stream fish assemblages, and Patton et al. (2000) used 16 contiguous samples for similar analyses of stream fish assemblages. Many small regions lack sufficient data to conduct reliable or informative sampling simulations. The sufficiency of the data for these analyses should be assessed prior to using the simulation results to direct sampling or to assess sampling sufficiency. Comparing the similarity of sampling simulation results for several analyses that exclude portions of the available data can provide a simple assessment of data sufficiency. If omission of a small portion of the data from an analysis has a large effect on the results, the data are likely insufficient for reliable sampling simulations.

The mathematical function and statistic(s) used in simulations should be selected to meet study objectives. Poor fitting functions bias results (Flather 1996, Diaz-Frances and Soberon 2005) and impede comparisons. Adequately fitting a single function to several distributions is challenging (Flather 1996, Thompson et al. 2003), but necessary for comparisons. With appropriate data, the slope ( $b$ ) of the simulation results indexes species discontinuity (Flather 1996). When calculating and comparing  $b$  among studies the negative binomial function (Efron and Thisted 1976) may be preferred because it is more commonly used. The data used for these analyses were not appropriate for comparisons of  $b$  as a measure of species discontinuity (Flather

1996, Angermeier et al. 2002) because samples were not random. However,  $b$  can be used to compare the rates of information accrual (Table 1.1, Figure 1.6).

The sampling sufficiency statistic should also be selected to meet study objectives. The statistic  $N_{95}$  is the mean number of samples required to collect 95% of the species estimated to be collected in the region with a very large number of additional samples collected using the same methods, equipment, and sampling strategy. The definition of the statistic varies if the sample attributes differ from those described. If samples are randomly distributed and the sampling method is unbiased among species (always an invalid assumption),  $N_{95}$  describes the mean number of samples required to collect 95% of the species present in the region. My selection of  $N_{95}$  was somewhat arbitrary and alternative statistics can be selected if lower or higher accuracy is desired. Moreno and Halffter (2000) used  $N_{90}$  to assess sampling sufficiency for relatively common bats, whereas Cao et al. (2001) recommend further development of sample similarity statistics. Measures of sample similarity may prove less useful for large-scale surveys where species discontinuity is expected to be common (Angermeier et al. 2002).

Acceptable survey variability should be considered when selecting a sampling sufficiency statistic. Bootstrap confidence bounds describe the expected variability in the proportion of SR sampled at particular sampling efforts (Figure 1.4). Statistics such as  $N_{90}$  may prescribe smaller recommended sample sizes on average than  $N_{95}$ , but may have greater variability. A stringent statistic and small bootstrap confidence bounds should be selected when accurate description of species richness (reliable collection of rare species) is desired. Alternatively, if the intended uses of survey results are multivariate analyses that ignore or discount rare species (Gauch 1982), less stringent statistics that require fewer samples may be preferred.

#### *Sampling Sufficiency Assessment*

Additional sampling is needed to adequately describe stream fish assemblages in most large regions of VA. Many fishes are likely present in regions where they have not been recorded and the fish assemblage has not been described at all in many environmental configurations (e.g., stream reach types). This paucity of information obscures ecological patterns, thwarts accurate prediction of occurrence, and precludes management certainty. Sufficient sampling at very large spatial extents (e.g., D and D-P in this study) is common (Conner and McCoy 1979, Palmer 1990), but does not ensure sufficient sampling at smaller extents where most management and conservation efforts occur. Sampling sufficiency must be addressed for the region(s) and scales of interest because species-habitat relationships, the relative importance of environmental predictors of species occurrence (Leftwich et al. 1997, Guisan and Zimmerman 2000), and the distribution and abundance species (Murray et al. 1999) all vary among regions and scales (Weins

1989, 2002, Figure 1.6). For most regions, additional sampling will likely collect species not currently known to be present (Keating et al 1998). This tendency is exacerbated by introductions and invasions of new species. In particular, additional samples will likely collect additional species in most SD and SD-P. For example, the Buffalo SD of the James D is insufficiently sampled and the bootstrap confidence bounds are large (Figure 1.4), indicating that additional samples will likely collect additional species and improve confidence in the sufficiency of sampling effort. Additionally, insufficient sampling exacerbates the problem of confidently asserting species absence. Species richness estimates vary even with large sample sizes, demonstrating that sampling frequently fails to collect species where they are present. Confidently asserting species absence might require additional sampling beyond that sufficient to describe SR.

Prohibitively large  $N_{95}$  may necessitate survey prioritization and coordination at appropriate spatial scales. It is not necessary to sufficiently sample all constituent sub-regions to sufficiently sample the composite region they form. Prioritization of surveys of constituent sub-regions can be coordinated with a survey of the composite region to ensure sufficient sampling of the regions and scales where survey data are most needed. For example, it may be necessary to sufficiently sample those relatively small regions (e.g., SD) where human impacts and associated changes in local assemblage composition are most likely and sufficiently sample the composite region (e.g., D) that provides a context for assessing changes in local assemblages.

#### *Predicting Sufficient Sampling Effort*

Survey planning and prioritization would be facilitated by accurate prediction of  $N_{95}$  based on environmental conditions. However, the effects of environmental conditions on  $N_{95}$  are complex and vary among scales, precluding the accurate prediction of  $N_{95}$ . Regional and local processes regulate SR (Ricklefs 1987, Ricklefs and Schluter 1993, Angermeier and Winston 1998). Large-scale processes such as speciation and dispersal provide the regional species pool from which local assemblages are derived. Within regions most species occur along limited portions of environmental gradients (Schlosser 1982, Moyle and Vondracek 1985). Local habitat conditions, disturbance regimes, and species interactions control the distribution of individual species and determine the local species assemblage (Angermeier and Schlosser 1989, Schlosser 1991). The low proportion of variance in  $N_{95}$  explained by RD precludes precise prediction of  $N_{95}$ , especially at small spatial extents. However, the effects of assemblage structure on  $N_{95}$  may provide guidance for improving sampling efficiency.

Regions with high SR can have many unique local assemblages and require more effort to sample sufficiently than depauperate assemblages simply because more local assemblage

compositions are possible. For example, the higher SR of the Appomattox SD (SR = 75) relative to that of the Buffalo SD (SR = 62) may account for some of the difference in  $N_{95}$  between the two regions (Figures 1.4 and 1.5). However, variation in SR is not necessary for differences in sample composition to affect  $N_{95}$ . The Buffalo SD and Rivanna SD have identical observed SR, but the relatively small proportion of common species in the Buffalo SD relative to the Rivanna SD (Table 1.1, Figure 1.6) requires twice the effort to sufficiently sample the Buffalo SD (Table 1.1, Figures 1.4).

The effect of SR on  $N_{95}$  is usually dominated by species prevalence (i.e., rarity, Table 1.2). If individual samples sufficiently describe the local assemblage, discontinuously distributed species and unique local assemblage compositions will not substantially affect  $N_{95}$  because most samples have a similar species composition, unless many species are rare. For example, although the Chickahominy SD and Appomattox SD have similar SR, the low mean prevalence and high proportion of rare species of the Chickahominy, (mean prevalence = 0.11 and 43% rare) compared to the Appomattox SD (mean prevalence = 0.20 and 25% rare) require much larger  $N_{95}$  (Table 1.2, Figure 1.6). Additionally, the strength of the correlations between environmental and assemblage attributes are expected to vary with the scale of analysis (Levin 1992). The complex interactions between environmental conditions and species distributions make accurate prediction of  $N_{95}$  unlikely.

Prediction of  $N_{95}$  is further complicated by variability in estimates of  $N_{95}$  due to non-standardized sampling. Sampling efficiency varies among sites with different assemblages and environmental conditions (Angermeier and Smogor 1995, Peterson et al. 2004, Rosenberger and Dunham 2005) and when different methods and equipment are used (Hubert 1996). Among regions, differences in environmental conditions that affect sampling efficiency can require different sampling effort, even when the assemblages are similar. Furthermore,  $N_{95}$  estimates are also affected by the spatial distribution of samples. Nonrandom and non-standardized sampling likely increase the variability of estimates of  $N_{95}$  by underestimating  $N_{95}$  in regions where some rare species are not sampled and overestimating  $N_{95}$  in regions where rare species are frequently sampled. For example, while the Buffalo SD, Rivanna SD, Chickahominy SD, and Appomattox SD have similar SR and assemblage compositions (Figure 1.6) they have substantially different  $N_{95}$ . Differences in  $N_{95}$  may be due to sampling errors, how the species are distributed in space, and/or how samples are distributed in space, but the available data do not allow examination of the relative importance of these effects.

Finally, variation among regions in the ecological significance of the attributes used to describe environmental conditions can confound prediction of sampling sufficiency.

Classification systems can be useful for describing the environmental heterogeneity of aquatic systems (Rosgen 1994, Montgomery and Buffington 1998, Higgins et al. 2005). However, because the classes are based on subjective criteria, many classes may be ecologically meaningless (Goodwin 1999, Poole et al. 1997) to one or more of the species in the assemblage. Using ecologically meaningless reach types weakens correlations between environmental heterogeneity and  $N_{95}$ . Well-designed large-scale surveys and careful selection of predictors may improve estimates of  $N_{95}$ , but estimation of  $N_{95}$  using sampling simulations concurrently with sampling are necessary to accurately assess sampling sufficiency.

#### *Assessing Sampling Strategies*

The locations of samples in space and omission errors dramatically affect  $N_{95}$ . Their relative importance for improving sampling efficiency should be explored further. Sampling simulations can identify sampling strategies that improve sampling efficiency without the cost of collecting additional field data (Dale et al. 1991, Hirzel and Guisan 2002). Spatially realistic simulations should provide useful results because sampling simulations are not immune to outliers (Smith 1998) or biased data. Spatially proximate locations usually have similar assemblages within a region (Angermeier and Winston 1999, Angermeier et al. 2002). That is, assemblage composition is usually spatially autocorrelated (Legendre 1993). Thus, spatially clustered samples inefficiently sample assemblages in large region by collecting many similar samples and a few rare species, thereby inflating estimates of  $N_{95}$ . Sampling strategies that more evenly distribute samples in space increase the rate of information accrual and reduce  $N_{95}$ . Researchers should be encouraged to coordinate sampling efforts to improve the distribution of samples in space and maximize the usefulness of all data. Survey planning tools that promote efficient representative sampling are available (Stevens and Olsen 2004) and their use should be encouraged. Spatially representative sampling would dramatically reduce the sampling effort required to describe the fish assemblage for most regions of VA. However, thorough samples with few errors of omission are necessary.

Omission errors also inflate estimates of  $N_{95}$ . The length of stream to sample to reduce errors of omission has been studied in several locations (e.g., Lyons 1992, Angermeier and Smogor 1995, Paller 1995, Simonson and Lyons 1995, Patton et al. 2000). More extensive samples typically have fewer omission errors, but collecting extensive samples may reduce the number of samples that can be collected. The relative importance of reducing errors of omission versus increasing sample size will vary with the study objectives. However, local sampling sufficiency should be encouraged to reduce omission errors. Metadata describing sampling effort

should also be collected to allow subsequent assessment of local sampling sufficiency and omission errors to facilitate survey planning (Peterson and Dunham 2002).

## CONCLUSIONS

Survey data are vital to management and research. Insufficient sampling is common and provides weak inferences and predictions, whereas excessive sampling diverts resources from other worthy projects. The development and application of methods to predict sufficient sampling effort and to assess the sufficiency of sampling are increasingly important as the scale of survey efforts and the importance of results increase. Sampling simulations using found data are useful for assessing sampling sufficiency and estimating required sampling effort. However, inappropriate sampling methods and failure to accurately report data and metadata limit the usefulness of most found samples. These shortcomings are easily rectified in future surveys, but common reporting standards are necessary.

The effort required to sufficiently sample a region varies with the survey objectives and analytical methods used, but generally acceptable and easily interpretable descriptive statistics can be used to guide survey efforts. The stream fish assemblages of most large regions of VA have not been sufficiently sampled. Many samples have been collected in most regions, but additional samples will likely produce additional species. Estimates of sufficient sample effort are large for most regions at all spatial extents of analysis. However, using found data likely overestimates sufficient sampling effort due to errors of sampling omission and spatially clustered samples. Reducing errors of sampling omission and improving sample coverage of the range of environmental variability can reduce the number of samples required to sufficiently sample large regions.

Required sampling effort usually increases with a region's spatial extent and environmental heterogeneity. However, precise estimation of required sampling effort using environmental descriptors is precluded by variability in the relationships between environmental attributes and required sampling effort among spatial extents and regions. Surveys should be designed and sampling sufficiency should be assessed at spatial extents that are ecologically meaningful and relevant to management and conservation. Hydrologic unit and physiographic boundaries can be used to identify such regions, among which important assemblage attributes such as species occurrence patterns and prevalence differ. The proportion of species that are rare primarily drives the required sampling effort. However, the prevalence of most species varies among regions and scales. Many species that are rare in some regions are common in others. Additionally, because many samples are not representative, it is difficult to determine whether



rare species preferentially use rare environmental conditions. Prediction of sampling errors will likely remain elusive, but well planned and coordinated sampling can improve survey efficiency.

Limited resources typically necessitate survey prioritization. Prioritization can be facilitated by using estimates of sufficient sampling calculated from found data and modeled assemblages. Representative sampling and reduced errors of sampling omission can improve survey efficiency, but it is not possible to attribute the additional needed effort to each type of error.

#### ACKNOWLEDGMENTS

This work greatly benefited from the assistance of many people. In particular I thank L. Delarche Roghair and S. Klopfer of the Conservation Management Institute at Virginia Tech for providing outstanding GIS assistance and E. P. Smith, T. Robinson, and D. P. Larsen provided helpful statistical advice. J. Copeland and R. Greenlee of the Virginia Department of Game and Inland Fisheries provided valuable advice and encouragement. The Virginia Department of Game and Inland Fisheries funded this work.

## LITERATURE CITED

- Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19: 716–723.
- Albanese, B., P. L. Angermeier, and S. Dorai-Raj. 2004. Ecological correlates of fish movement in a network of Virginia streams *Canadian Journal of Fisheries and Aquatic Sciences* 61: 857-869.
- Angermeier, P. L., and I. J. Schlosser. 1989. Species-area relationships for stream fishes. *Ecology* 70: 1450-1462.
- Angermeier, P. L., and M. R. Winston. 1998. Local vs. Regional influences on local diversity in stream fish communities of Virginia. *Ecology* 79: 911-927.
- Angermeier, P. L., and M. R. Winston. 1999. Characterizing fish community diversity across Virginia landscapes: prerequisite for conservation. *Ecological Applications* 9: 335-349.
- Angermeier, P. L., and R. A. Smogor. 1995. Estimating number of species and relative abundances in stream-fish communities: effects of sampling effort and discontinuous spatial distributions. *Canadian Journal of Fisheries and Aquatic Sciences* 52: 936-949.
- Angermeier, P. L., K. L. Krueger, and C. A. Dolloff. 2002. Discontinuity in stream-fish distributions: Implications for assessing and predicting species occurrence. In Scott, J. M., P. J. Heglund, M. Morrison, M. Raphael, J. Haufler, and B. Wall, Editors. *Predicting species occurrences: issues of accuracy and scale*. Island Press, Covello, CA. 840 p.
- Bonar, S. A., and W. A. Hubert. 2002. Standard sampling of inland fish: benefits, challenges, and a call for action. *Fisheries* 27: 10-16.
- Boulinier, T., J. D. Nichols, J. R. Sauer, J. E. Hines, and K. H. Pollock. 1998. Estimating species richness: the importance of heterogeneity in species detectability. *Ecology* 79: 1018-1028.
- Bunge, J., and M. Fitzpatrick. 1993. Estimating the number of species: a review. *Journal of the American Statistical Association* 88: 364-373.
- Burnham, K. P., and D. R. Anderson. 2002. *Model selection and multimodel inference: a practical-theoretic approach*, 2nd Edition. Springer, New York.
- Cairns, J., Jr., and J. R. Pratt. 1986. Developing a sampling strategy. *American Society for Testing and Materials* 894: 168-186.
- Cam, E., J. D. Nichols, J. E. Hines, J. R. Sauer, R. Alpizar-Jara, and C. H. Flather. 2002. Disentangling sampling and ecological explanations underlying species-area relationships. *Ecology* 83: 1118-1130.
- Cao, Y., D. P. Larsen, and R. M. Hughes. 2001. Evaluating sampling sufficiency in fish assemblage surveys - a similarity-based approach. *Canadian Journal of Fisheries and Aquatic Sciences* 58: 1782-1793.
- Cao, Y., D. P. Larsen, R. M. Hughes, P. L. Angermeier, and T. M. Patton. 2002. Sampling effort affects multivariate comparisons of stream assemblages. *Journal of the North American Benthological Society* 21: 701-714.
- Cohen J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20: 37-46.
- Coleman, B. D. 1981. On random placement and species-area relations. *Mathematical Biosciences* 54: 191-215.
- Connor, E. F., and E. D. McCoy. 1979. The statistics and biology of the species-area relationship. *The American Naturalist* 113: 791-833.
- Dale, V. H., R. L. A. Franklin, W. M. Post, and R. H. Gardner. 1991. Sampling ecological information: choice of sampling size. *Ecological Modeling* 57: 1-10.
- de Solla, S. R., L. J. Shirose, K. J. Fernie, G. C. Barrett, C. S. Brousseau, and C. A. Bishop. 2005. Effect of sampling effort and species detectability on volunteer based anuran monitoring programs. *Biological Conservation* 121: 585-594.

- Diaz-Frances, E., and J. Soberon. 2005. Statistical estimation and model selection of species accumulation functions. *Conservation Biology* 19: 569-573.
- Dolloff, C. A., H. E. Jennings, and M. D. Owen. 1997. A comparison of basinwide and representative reach habitat survey techniques in three southern Appalachian watersheds. *North American Journal of Fisheries Management* 17: 339-347.
- Dunham, J. B., and G. L. Vinyard. 1997. Incorporating stream level variability into analyses of site level fish habitat relationships: some cautionary examples. *Transactions of the American Fisheries Society* 126: 323-329.
- Edwards, D. 1998. Issues and themes for natural resources trend and change detection. *Ecological Applications* 8: 323-325.
- Efron, B. and R. Tibshirani. 1976. Estimating the number of unseen species: how many words did Shakespeare know? *Biometrika* 63: 435-447.
- Efron, B., and R. J. Tibshirani. 1994. *An introduction to the bootstrap*. Chapman & Hall, Norwell, Massachusetts, 456 pp.
- Fausch, K. D., and M. K. Young. 1995. Evolutionary significant units and movement of resident stream fishes: a cautionary tale. Pages 360-370 *in* Nielson, editor. *Evolution and the Aquatic Ecosystem*, American Fisheries Society Symposium 17.
- Fausch, K. D., C. E. Torgersen, C. V. Baxter, and H. W. Li. 2002. Landscapes to riverscapes: bridging the gap between research and conservation of stream fishes. *BioScience* 52: 483-498.
- Fausch, K. D., C. L. Hawkes, and M. G. Parsons. 1988. Models that predict standing crop of stream fish from habitat variables: 1950-85. General technical Report, USDA Forest Service, PNW-GTR-213, Portland, Oregon. 52 p.
- Fisher, R. A., A. S. Corbet, and C. B. Williams. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology* 12: 42-58.
- Flather, C. H. 1996. Fitting species-accumulation functions and assessing regional land use impacts on avian diversity. *Journal of Biogeography* 23: 155-168.
- Fraser, R. H. 1998. Vertebrate species richness at the mesoscale: relative roles of energy and heterogeneity. *Global Ecology and Biogeography* 7: 215-220.
- Gauch, H. G., Jr. 1982. *Multivariate analysis in community ecology*. Cambridge University Press, Cambridge, 314 p.
- Gleason, H. A. 1922. On the relation between species and area. *Ecology* 3: 158-162.
- Goodwin, C. N. 1999. Fluvial classification: Neanderthal necessity or needless normalcy. Pages 229-236 *in* D. S. Olsen and J. P. Potyondy, editors. *Wildland Hydrology*, American Water Resources Association, TPS-99-3, Herndon, Va.
- Gordon, N. D., T. A. McMahon, and B. L. Finlayson. 1992. *Stream hydrology: an introduction for ecologists*. John Wiley & Sons. New York.
- Gotelli, N., and R. K. Colwell. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters* 4: 379-391.
- Gowan, C., M. K. Young, K. D. Fausch, and S. C. Riley. 1994. Restricted movement of resident stream salmonids: a paradigm lost. *Canadian Journal of Fisheries and Aquatic Sciences* 51: 2626-2637.
- Guegan, J., S. Lek, and T. Oberdorff. 1998. Energy availability and habitat heterogeneity predict global riverine fish diversity. *Nature* 391: 382-384.
- Guggenmoose-Holzmann, I. 1996. The meaning of kappa: probabilistic concepts of reliability and validity revisited. *Journal of Clinical Epidemiology* 49: 775-82.
- Guisan, A., and N. E. Zimmerman. 2000. Predictive habitat distribution models in ecology. *Ecological Modelling* 135: 147-186.

- Hayes, D., E. Baker, R. Bednarz, D. Borgeson, Jr., J. Braunscheidel, J. Breck, M. Bremigan, A. Harrington, R. Hay, R. Lockwood, A. Nuhfer, J. Schneider, P. Seelbach, J. Waybrant, and T. Zorn. 2003. Developing a standardized sampling program: the Michigan experience. *Fisheries* 28: 18-25.
- Hellman, J. J., and G. W. Fowler. 1999. Bias, precision, and accuracy of four measures of species richness. *Ecological Applications* 9: 824-834.
- Herger, L. G., W. A. Hubert, and M. K. Young. 1996. Comparison of habitat composition and cutthroat trout abundance at two flows in small mountain streams. *North American Journal of Fisheries Management* 16: 294-301.
- Higgins, J. V., M. T. Bryer, M. L. Khoury, and T. W. Fitzhugh. 2005. A freshwater classification approach for biodiversity conservation planning. *Conservation Biology* 19: 432-445.
- Hirzel, A., and A. Guisan. 2002. Which is the optimal sampling strategy for habitat suitability modeling. *Ecological Modelling* 157: 331-341.
- Holling, C. S. 1992. Cross-scale morphology, geometry, and dynamics of ecosystems. *Ecological Monographs* 62: 447-502.
- Hosmer D. W., Jr., and S. Lemeshow. 2000. *Applied Logistic Regression*, 2nd Edition. John Wiley and Sons, New York.
- Hubert, W. A. 1996. Passive capture techniques. Pages 157-192 *in* B. R. Murphy and D. W. Willis, editors. *Fisheries techniques*, 2nd edition. American Fisheries Society, Bethesda, Maryland.
- Ibbotson, A., P. Armitage, W. Beaumont, M. Ladle, and S. Welton. 1994. Spatial and temporal distribution of fish in a small lowland stream. *Fisheries Management and Ecology* 1: 143-156.
- Jackson, D. A., Peres-Neto, P. R., and J. D. Olden. 2001. What controls who is where in freshwater fish communities – the roles of biotic, abiotic and spatial factors? *Canadian Journal of Fisheries and Aquatic Science* 58: 157-170.
- Jenkins, R. E., and N. M. Burkhead. 1993. *Freshwater Fishes of Virginia*. American Fisheries Society, Bethesda, Maryland.
- Keating, K. A., J. F. Quinn, M. A. Ivie, and L. L. Ivie. 1998. Estimating the effectiveness of further sampling in species inventories. *Ecological Applications* 8: 1239-1249.
- Kerr, J. T. and L. Packer. 1997. Habitat heterogeneity as a determinant of mammal species richness in high-energy regions. *Nature* 385: 252-254.
- Leftwich, K. N., P. L. Angermeier, and C. A. Dolloff. 1997. Factors influencing behavior and transferability of habitat models for a benthic stream fish. *Transactions of the American Fisheries Society* 126: 725-734.
- Legendre, P. 1993. Spatial autocorrelation: trouble or a new paradigm? *Ecology* 74: 1659-1673.
- Levin, S. A. 1992. The problem of pattern and scale in ecology. *Ecology* 73: 1943-1967.
- Longino, J. T., and R. K. Colwell. 1997. Biodiversity assessment using structured inventories: capturing the ant fauna of a tropical rain forest. *Ecological Applications* 7: 1263-1277.
- Lyons, J. 1992. The length of stream to sample with a towed electrofishing unit when fish species richness is estimated. *North American Journal of Fisheries Management* 12: 198-203.
- MacNally, R., and D. M. Watson. 1997. Distinguishing area and habitat heterogeneity effects on species richness: Birds in Victorian buloke remnants. *Australian Journal of Ecology* 22: 227-232.
- Manel, S., H. C. Williams, and S. J. Ormerod. 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology* 38: 921-931.
- Manly, B. F. J. 1997. *Randomization, bootstrap and Monte Carlo methods in biology*. CRC

- Press, Boca Raton, Florida. 399 pp.
- Matthews, W. J., and H. W. Robison. 1998. Influence of Drainage Connectivity, Drainage Area and Regional Species Richness on Fishes of the Interior Highlands in Arkansas. *The American Midland Naturalist* 139: 1-19.
- McIntyre, S., and R. Hobbs. 1999. A framework for conceptualizing human effects on landscapes and its relevance to management and research methods. *Conservation Biology* 13: 1282-1292.
- Miller, R. I., and R. G. Wiegert. 1989. Documenting completeness, species-area relations, and the species-abundance distribution of a regional flora. *Ecology* 70: 16-22.
- Mingoti, S. A., and G. Meeden. 1992. Estimating the total number of distinct species using presence and absence data. *Biometrics* 48: 863-875.
- Montgomery, D. R. and J. M. Buffington. 1998. Channel processes, classification, and response. Pages 13-42 *in* R. Naiman and R. Bilby, Editors, *River Ecology and Management*. Springer-Verlag Inc., New York, New York.
- Moreno, C. E., and G. Halffter. 2000. Assessing the completeness of bat biodiversity inventories using species accumulation curves. *Journal of Applied Ecology* 37: 149-158.
- Moyle, P. B., and B. Vondracek. 1985. Persistence and structure of the fish assemblage in a small California stream. *Ecology* 66: 1-13.
- Murray, B. R., and 7 co-authors. 1999. Species in the tail of rank-abundance curves. *Ecology* 80: 1806-1816.
- Ney-Nifle, M., and M. Mangel. 1999. Species-area curves based on geographic range and occupancy. *Journal of Theoretical Biology* 196: 327-342.
- Norris, R. H. 1995. Biological monitoring: the dilemma of data analysis. *Journal of the North American Benthological Society* 14: 440-450.
- Osborne, L. L., and M. J. Wiley. 1992. Influence of tributary spatial position on the structure of warmwater fish assemblages. *Canadian Journal of Fisheries and Aquatic Sciences* 49:671-681.
- Paller, M. H. 1995. Interreplicate variance and statistical power of electrofishing data from low-gradient streams in the southeastern United States. *North American Fisheries Management* 15: 542-550.
- Palmer, M. A., and N. L. Poff. 1997. Heterogeneity in streams: the influence of environmental heterogeneity on patterns and processes in streams. *Journal of the North American Benthological Society* 16: 169-173.
- Palmer, M. W. 1990. The estimation of species richness by extrapolation. *Ecology* 71: 1195-1198.
- Patton, T. M., W. A. Hubert, F. J. Rahel, and K. G. Gerow. 2000. Effort needed to estimate species richness in small streams of the Great Plains in Wyoming. *North American Journal of Fisheries Management* 20: 394-398.
- Peterson, J. T., and C. F. Rabeni. 1995. Optimizing sampling effort for sampling warmwater stream fish communities. *North American Journal of Fisheries Management* 15: 528-541.
- Peterson, J. T., and J. Dunham. 2002. Combining inferences from models of capture efficiency, detectability, and suitable habitat to classify landscapes for conservation of threatened bull trout. *Conservation Biology* 17: 1070-1077.
- Peterson, J. T., R. F. Thurow, and J. W. Guzevich. 2004. An evaluation of multipass electrofishing for estimating the abundance of stream-dwelling salmonids. *Transactions of the American Fisheries Society* 133: 462-475.
- Poff, N. L., and J. D. Allen. 1995. Functional organization of stream fish assemblages in relation to hydrological variability. *Ecology* 76: 606-627.
- Poole, G. C., C. A. Frissell, and S. C. Ralph. 1997. In-stream habitat unit classification: inadequacies for monitoring and some consequences for management. *Journal of the American Water Resources Association* 33: 879-896.

- Preston, F. W. 1962. The canonical distribution of commonness and rarity. *Ecology* 43: 185-215 and 410-432.
- Ricklefs, R. E. 1987. Community diversity: relative roles of local and regional processes. *Science* 235: 167-171.
- Ricklefs, R. E. and D. Schluter, editors. 1993. *Species diversity in ecological communities: historical and geographical perspectives*. University of Chicago Press, Chicago, Illinois.
- Rosenberger, A. E. and J. B. Dunham. 2005. Validation of abundance estimates from mark-recapture and removal techniques for rainbow trout captured by electrofishing in small streams. *North American Journal of Fisheries Management* 25: 1395-1410.
- Rosgen, D. L. 1994. A classification of natural rivers. *Catena* 22: 169-199.
- Schlosser, I. J. 1982. Fish community structure and function along two habitat gradients in a headwater stream. *Ecological Monographs* 52: 395-414.
- Schlosser, I. J. 1991. Stream fish ecology: a landscape perspective. *BioScience* 41: 704-712.
- Schlosser, I. J. 1995. Critical landscape attributes that influence fish population dynamics in headwater streams. *Hydrobiologia* 303: 71-81.
- Scott, J. M., P. A. Heglund, M. L. Morrison, J. B. Haufler, M. B. Raphael, W. A. Wall, and F. B. Samson. 2002. *Predicting species occurrences: issues of accuracy and scale*. Island Press, Washington, D.C. 868 p.
- Seaber, P.R., Kapinos, F.P., and Knapp, G.L. 1987. Hydrologic unit maps: U.S. Geological Survey water-supply paper 2294.
- Simonson, T. D., and J. Lyons. 1995. Comparison of catch per effort and removal procedures for sampling stream fish assemblages. *North American Journal of Fisheries Management* 15: 419-427.
- Smith, E. P. 1998. Randomization methods and the analysis of multivariate ecological data. *Environmetrics* 9: 37-51.
- Smith, E. P., and G. van Belle. 1984. Nonparametric estimation of species richness. *Biometrics* 40: 119-129.
- Smith, T. A., and C. E. Kraft. 2005. Stream fish assemblages in relation to landscape position and local habitat variables. *Transactions of the American Fisheries Society* 134: 430-440.
- Stevens, D. L. 2002. Sampling design and statistical analysis methods for the integrated biological and physical monitoring of Oregon streams. The Oregon Plan for Salmon and Watersheds, EPA National Health and Environmental Effects Research Laboratory, Western Ecology Division, Corvallis, Oregon. Report Number OPSW-ODFW-2002-07.
- Stevens, D. L., and A. R. Olsen. 2004. Spatially balanced sampling of natural resources. *Journal of the American Statistical Association* 99: 262-278.
- Strahler, A. N. 1957. Quantitative analysis of watershed geomorphology. *Transactions of the American Geographical Union* 38: 913-920.
- Thompson, G. G., P. C. Withers, E. R. Pianka, and S. A. Thompson. 2003. Assessing biodiversity with species accumulation curves; inventories of small reptiles by pit-trapping in Western Australia. *Austral Ecology* 28: 361-383.
- Ward, J. V., K. Tockner, D. B. Arscott, and C. Claret. 2002. Riverine landscape diversity. *Freshwater Biology* 47: 517-539.
- Weins, J. A. 1989. Spatial scaling in ecology. *Functional Ecology* 3: 385-397.
- Weins, J. A. 2002. Riverine landscapes: taking landscape ecology into the water. *Freshwater Biology* 47: 501-515.

Table 1.1. Survey region name, USGS hydrologic unit code (HUC), available number of samples (N), estimated sufficient sample size (N<sub>95</sub>), slope of the results of sampling simulations (*b*), observed species richness (SR obs.), estimated species richness (SR est.), percent of species that are rare (%Rare; species in ≤ 5% of samples), reach diversity (RD; number of reach types), and percent of reach types sampled (%RD) for 84 regions at four spatial scales in Virginia where sampling sufficiency assessments were performed.

Survey Region Name	HUC	N	N <sub>95</sub>	<i>b</i>	SR obs.	SR est.	%Rare	RD	%RD
<b>Drainage</b>									
Potomac	020700	282	160	0.071	79	80	41	151	34
Rappahannock	020801	184	95	0.107	72	71	32	102	29
York	020801	150	75	0.130	66	65	29	74	22
James	020802	820	225	0.080	99	97	36	225	44
Roanoke	030101	580	275	0.044	91	90	51	183	42
Chowan	030102	303	124	0.095	85	83	34	100	35
New	050500	177	165	0.089	69	68	35	74	35
Big Sandy	050702	68	48	0.139	42	43	24	16	75
Holston	060101	176	131	0.079	82	83	44	50	42
Clinch	060102	226	203	0.070	91	89	37	66	52
<b>Drainage-Physiography</b>									
Potomac-Piedmont	020700	131	105	0.117	63	62	33	96	34
Potomac-Ridge and Valley	020700	101	102	0.130	57	59	33	78	24
Rappahannock-Coastal Plain	020801	79	69	0.099	49	49	33	43	16
Rappahannock-Piedmont	020801	94	70	0.141	60	60	32	87	30
York-Coastal Plain	020801	80	91	0.093	52	50	35	48	17
York-Piedmont	020801	70	56	0.171	59	58	19	63	21
James-Coastal Plain	020802	218	248	0.080	67	66	39	57	25
James-Piedmont	020802	383	227	0.091	88	85	36	134	40
James-Blue Ridge	020802	29	29	0.291	46	45	9	83	17

Table 1.1. Continued.

Survey Region Name	HUC	N	N <sub>95</sub>	<i>b</i>	SR obs.	SR est.	%Rare	RD	%RD
<b>Drainage-Physiography</b>									
James-Ridge and Valley	020802	190	105	0.120	62	62	31	104	37
Roanoke-Piedmont	030101	397	245	0.052	87	85	48	163	41
Roanoke-Blue Ridge	030101	94	85	0.138	58	57	31	79	37
Roanoke-Ridge and Valley	030101	89	96	0.139	52	53	29	61	38
Chowan-Coastal Plain	030102	158	146	0.071	75	73	41	55	29
Chowan-Piedmont	030103	145	136	0.104	75	74	33	86	28
New-Blue Ridge	050500	70	58	0.209	52	51	21	49	18
New-Ridge and Valley	050500	107	171	0.102	62	60	31	70	33
Big Sandy-Cumberland Plateau	050702	65	48	0.140	42	43	24	16	75
Clinch-Ridge and Valley	060102	154	135	0.080	87	86	37	57	53
Clinch-Cumberland Plateau	060102	72	101	0.121	76	73	28	48	60
<b>Subdrainage</b>									
South Fork Shenandoah	02070005	64	59	0.186	50	50	28	8	25
North Fork Shenandoah	02070006	30	19	0.165	38	50	0	54	22
Middle Potomac-Catoctin	02070008	54	43	0.356	45	45	16	100	26
Middle Potomac-Anacostia-Occoquan	02070010	79	84	0.158	52	51	27	75	32
Rapidan-Upper Rappahannock	02080103	83	43	0.221	49	48	18	88	28
Lower Rappahannock	02080104	101	138	0.072	66	64	39	45	24
Mattoponi	02080105	56	56	0.164	57	55	21	55	13
Pamunkey	02080106	89	57	0.145	64	63	22	67	21
Cowpasture	02080201	190	97	0.132	62	61	26	106	38
Buffalo	02080203	90	107	0.124	62	61	27	114	23



Table 1.1. Continued.

Survey Region Name	HUC	N	N <sub>95</sub>	<i>b</i>	SR obs.	SR est.	%Rare	RD	%RD
<b>Subdrainage</b>									
Rivanna	02080204	90	53	0.268	62	62	15	76	22
Tuckahoe	02080205	53	133	0.182	63	59	22	65	15
Chickahominy	02080206	291	406	0.064	70	68	43	47	43
Appomattox	02080207	122	95	0.142	75	74	25	87	38
Upper Roanoke	03010101	313	230	0.066	69	69	45	118	44
Upper Dan	03010103	142	224	0.074	58	56	41	93	42
Lower Dan	03010104	34	33	0.182	57	56	5	78	22
Nottoway	03010201	125	92	0.096	77	75	30	83	27
Blackwater	03010202	81	104	0.097	53	54	30	36	25
Meherrin	03010204	81	65	0.165	69	69	28	71	27
Upper New	05050001	113	84	0.150	60	58	25	63	32
Middle New	05050002	64	119	0.140	54	52	26	58	26
North Fork Holston	06010101	94	118	0.158	67	66	27	45	36
South Fork Holston	06010102	82	81	0.108	65	65	32	41	39
Upper Clinch	06010205	194	179	0.069	91	90	40	59	51
Powell	06010206	32	43	0.246	61	59	13	50	54
<b>Subdrainage-Physiography</b>									
South Fork Shenandoah-Ridge and Valley	02070005	51	66	0.205	48	47	23	6	33
North Fork Shenandoah-Ridge and Valley	02070006	30	18	0.560	38	38	0	54	22
Middle Potomac-Catoctin-Piedmont	02070008	48	39	0.293	45	44	16	86	27
Middle Potomac-Anacostia-Occoquan-Piedmont	02070010	65	77	0.190	44	43	23	74	31
Rapidan-Upper Rappahannock-Piedmont	02080103	72	41	0.252	49	48	16	81	27

Table 1.1. Continued.

Survey Region Name	HUC	N	N <sub>95</sub>	<i>b</i>	SR obs.	SR est.	%Rare	RD	%RD
<b>Subdrainage-Physiography</b>									
Mattoconi-Coastal Plain	02080105	35	46	0.105	44	44	11	43	5
Pamunkey-Coastal Plain	02080106	40	50	0.101	45	45	33	35	14
Pamunkey-Piedmont	02080106	49	45	0.206	55	54	18	59	20
Cowpasture-Ridge and Valley	02080201	174	95	0.123	62	61	29	92	39
Buffalo-Piedmont	02080203	85	116	0.130	62	60	29	96	24
Rivana-Piedmont	02080204	48	48	0.082	61	60	13	68	25
Tuckahoe-Piedmont	02080205	53	133	0.180	63	60	22	65	15
Chickahominy-Coastal Plain	02080206	212	190	0.087	65	64	37	43	33
Chickahominy-Piedmont	02080206	80	111	0.082	51	51	41	30	37
Appamattox-Piedmont	02080207	117	89	0.150	72	71	26	85	39
Upper Roanoke-Piedmont	03010101	138	135	0.069	59	59	42	95	36
Upper Roanoke-Blue Ridge	03010101	86	73	0.151	56	56	21	77	38
Middle Roanoke-Piedmont	03010102	57	57	0.147	72	71	26	95	23
Lower Rappahannock-Piedmont	03010103	134	194	0.078	57	55	40	89	43
Lower Dan-Piedmont	03010104	34	31	0.182	57	56	5	78	22
Nottoway-Coastal Plain	03010201	51	67	0.124	66	64	21	46	20
Blackwater-Piedmont	03010201	74	98	0.109	68	66	31	71	20
Meherrin-Piedmont	3010204	71	71	0.152	65	66	28	64	30
Upper New-Ridge and Valley	05050001	43	68	0.173	50	48	24	54	28
Middle New-Ridge and Valley	05050002	64	102	0.140	54	52	26	58	26
Upper Levisa-Cumberland Plateau	05070202	60	54	0.139	41	41	24	16	75
Upper Clinch-Ridge and Valley	06010205	138	155	0.078	87	86	38	49	59
Upper Clinch-Cumberland Plateau	06010205	56	91	0.135	74	71	31	42	38

Table 1.2. Spearman correlation coefficients and p-values that are bootstrap-adjusted for multiple comparisons between mean number of samples needed to collect 95% of estimated species richness ( $N_{95}$ ), area, number of reach types (RD), percent of RD sampled (%RD), observed species richness (SR), and percent of species that are rare (% Rare) in 10 drainages (D), 20 drainage-physiographies (D-P), 26 subdrainages (SD) and 28 subdrainage-physiographies (SD-P) in Virginia.

	$N_{95}$	Area	RD	% RD	SR	$N_{95}$	Area	RD	% RD	SR
	<b>D</b>					<b>SD</b>				
Area	0.842					0.328				
	0.002					0.101				
RD	0.608	0.681				0.049	0.422			
	0.062	0.030				0.814	0.032			
% RD	0.321	0.321	-0.261			0.339	0.151	0.080		
	0.366	0.366	0.466			0.089	0.460	0.698		
SR	0.821	0.736	0.549	0.334		0.510	0.284	0.048	0.387	
	0.004	0.015	0.100	0.345		0.008	0.159	0.816	0.051	
% Rare	0.733	0.624	0.340	0.188	0.681	0.775	0.344	-0.115	0.486	0.514
	0.016	0.054	0.336	0.603	0.030	<0.001	0.086	0.575	0.012	0.007
	<b>D-P</b>					<b>SD-P</b>				
Area	0.631					0.267				
	0.003					0.169				
RD	0.381	0.525				0.110	0.569			
	0.098	0.017				0.576	0.002			
% RD	0.317	0.066	0.175			0.446	0.062	0.026		
	0.174	0.782	0.462			0.017	0.754	0.893		
SR	0.820	0.538	0.436	0.470		0.517	0.348	0.256	0.141	
	<0.001	0.014	0.055	0.036		0.005	0.069	0.188	0.475	
% Rare	0.781	0.482	0.213	0.104	0.625	0.802	0.178	-0.005	0.491	0.438
	<0.001	0.031	0.368	0.663	0.003	<0.001	0.365	0.981	0.008	0.020

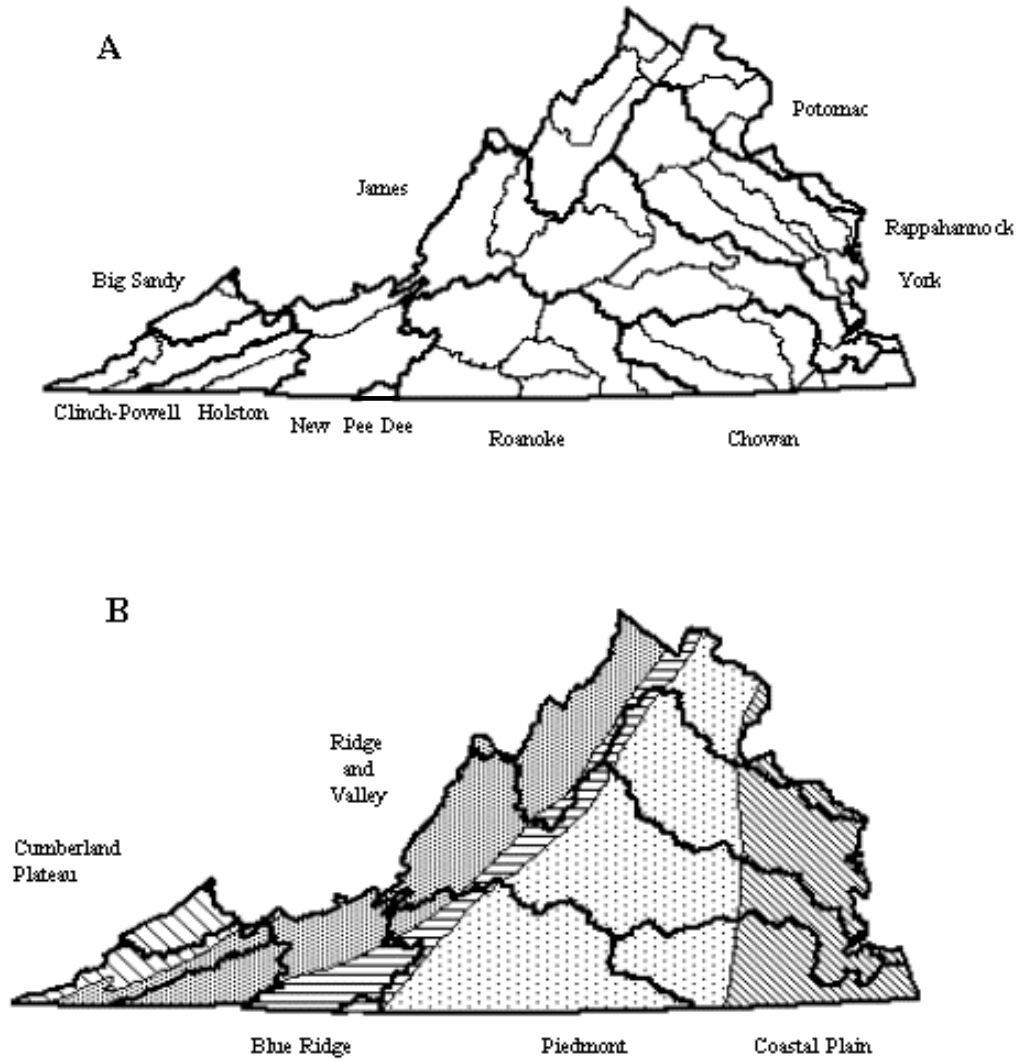


Figure 1.1. Major drainages (USGS 6-digit hydrologic units, bold outline) and subdrainages (USGS 8-digit hydrologic units) in Virginia (A). Intersections of major drainages with the physiographic provinces in Virginia (B).

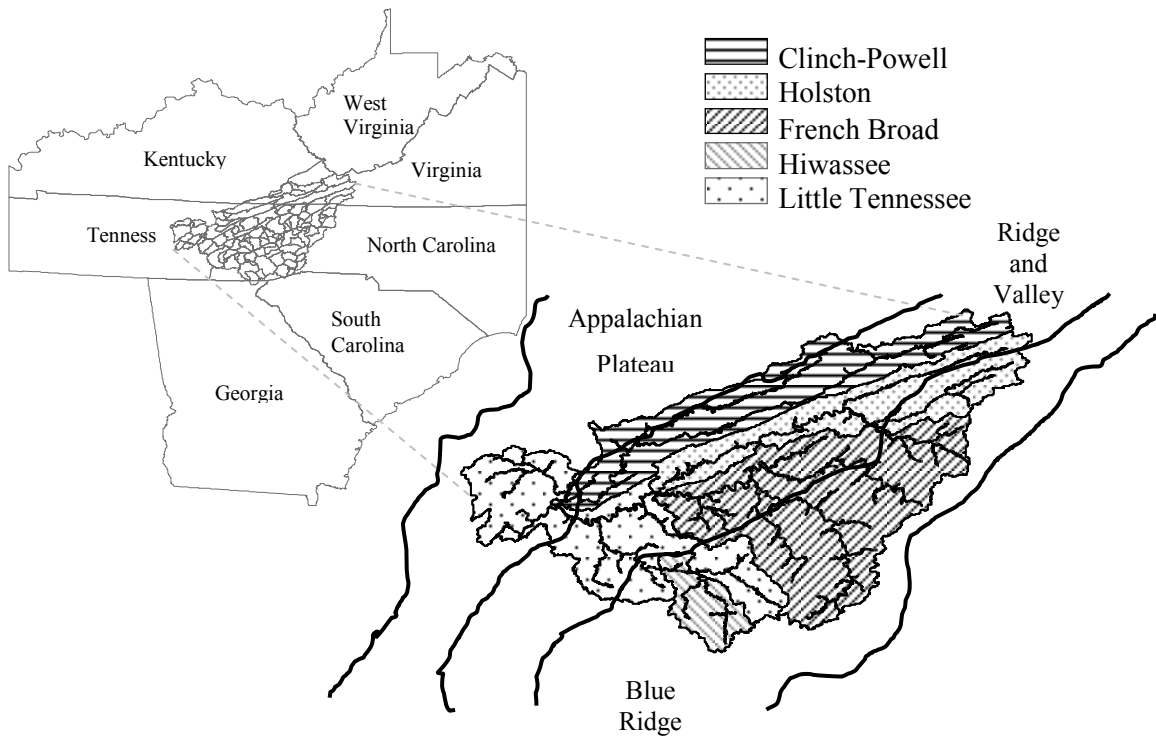


Figure 1.2. Location of the upper Tennessee River basin in the United States and the 5 major tributaries and 3 physiographic provinces of the upper Tennessee River basin.



Figure 1.3. Locations of the 2,966 stream fish assemblage samples used in sampling simulations relative to drainages, subdrainages, and physiographic provinces in Virginia.

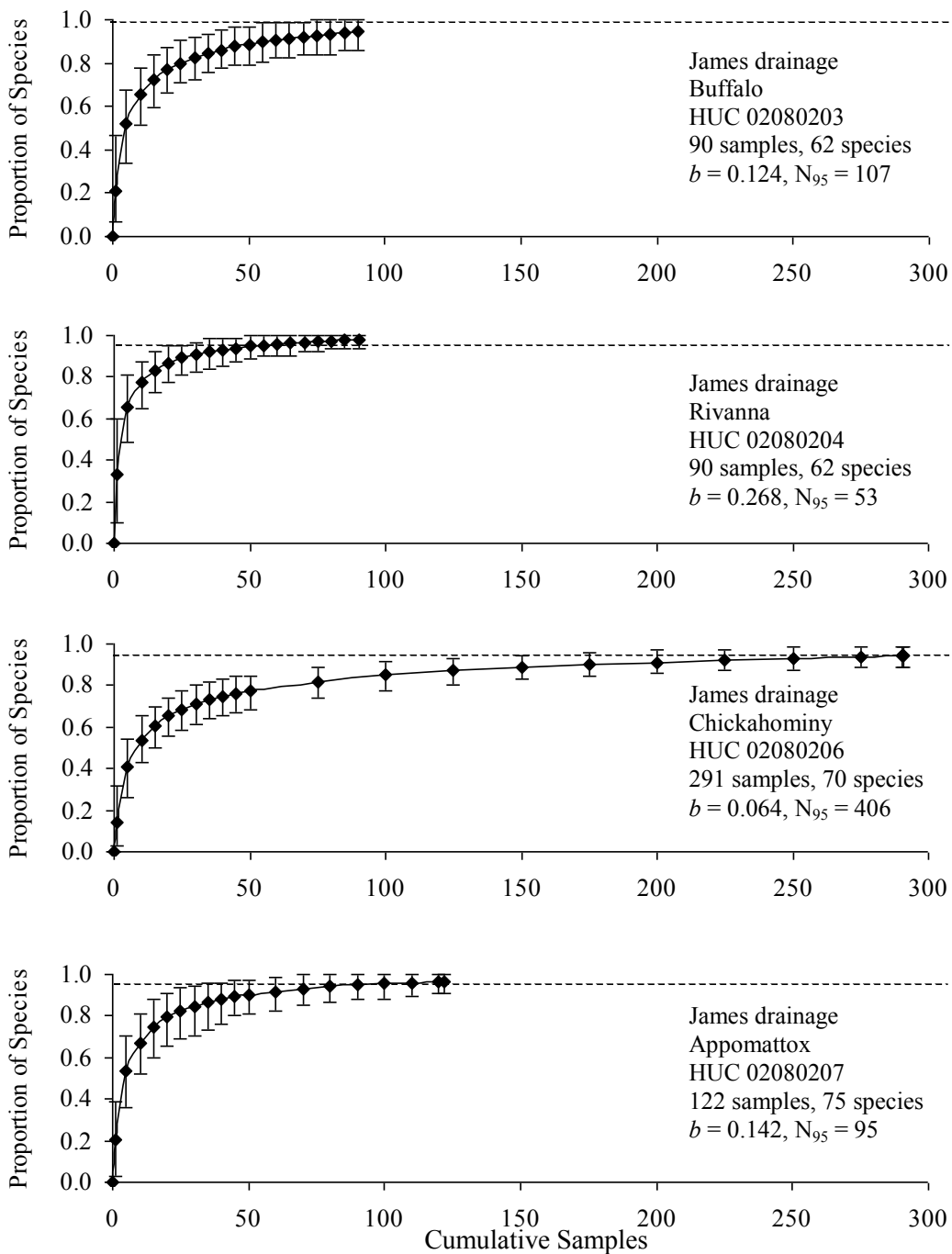


Figure 1.4. Sampling simulation results for four subdrainages of the James River drainage in Virginia. Values of the ordinate are the mean proportion of the estimated species richness collected in 10,000 sample iterations at the sample size on the abscissa. Error bars are 95% bootstrap confidence bounds. The number of samples, observed species richness, slope of the simulation results ( $b$ ) and  $N_{95}$  are provided for comparison. The dashed line is at  $N_{95}$ .

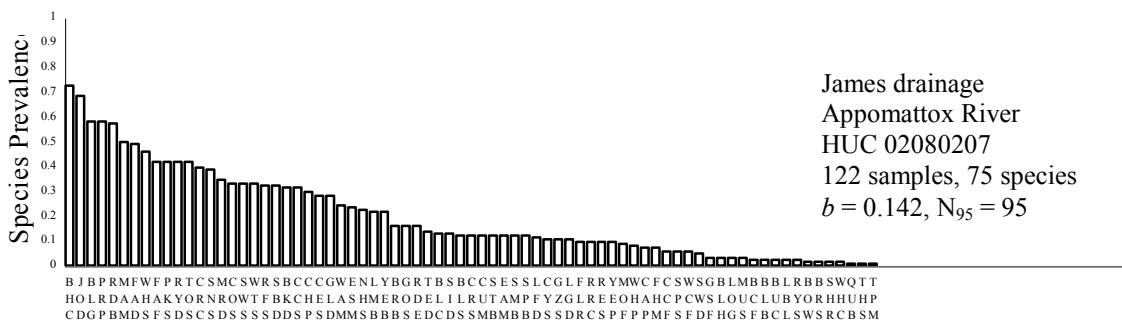
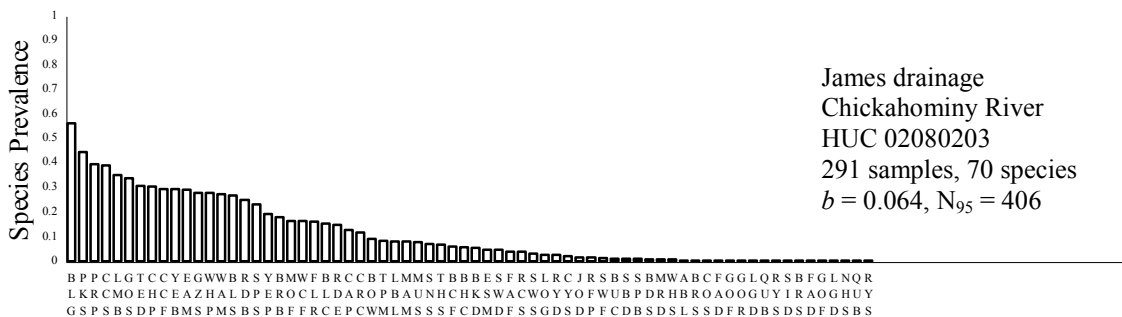
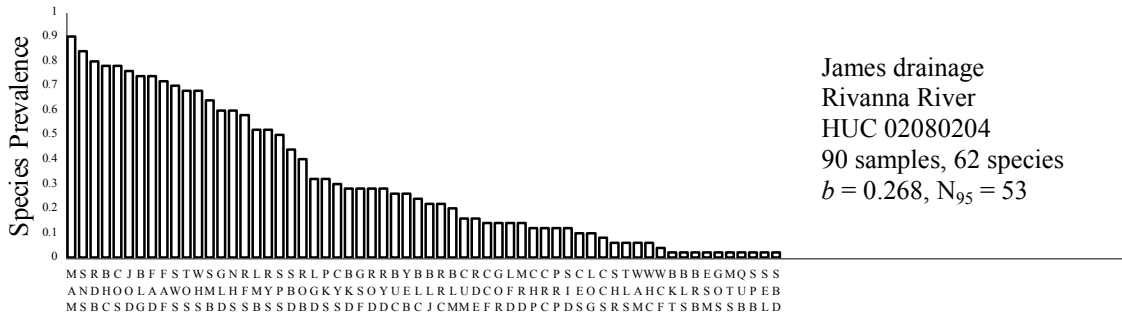
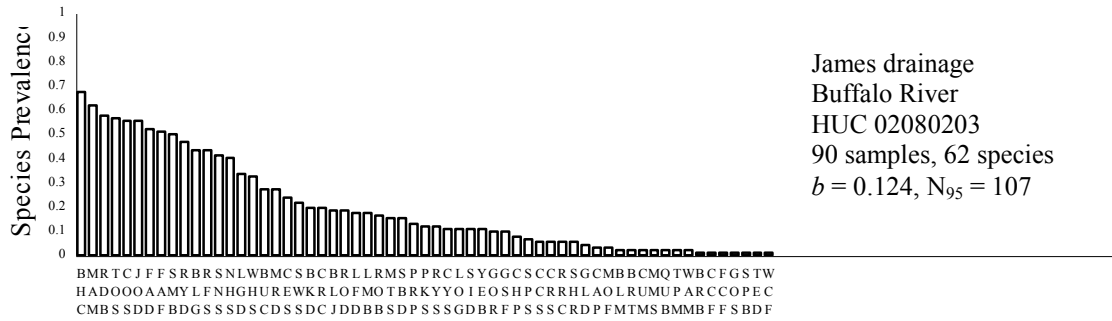


Figure 1.5. Relative prevalence of fish species collected in four subdrainages in the James River drainage in Virginia. Sample sizes, species richness, slope of the sampling simulation curve ( $b$ ) and estimate of the number of samples required to collect 95% of the species are provided for comparison. Three letter codes based on species common names are provided.



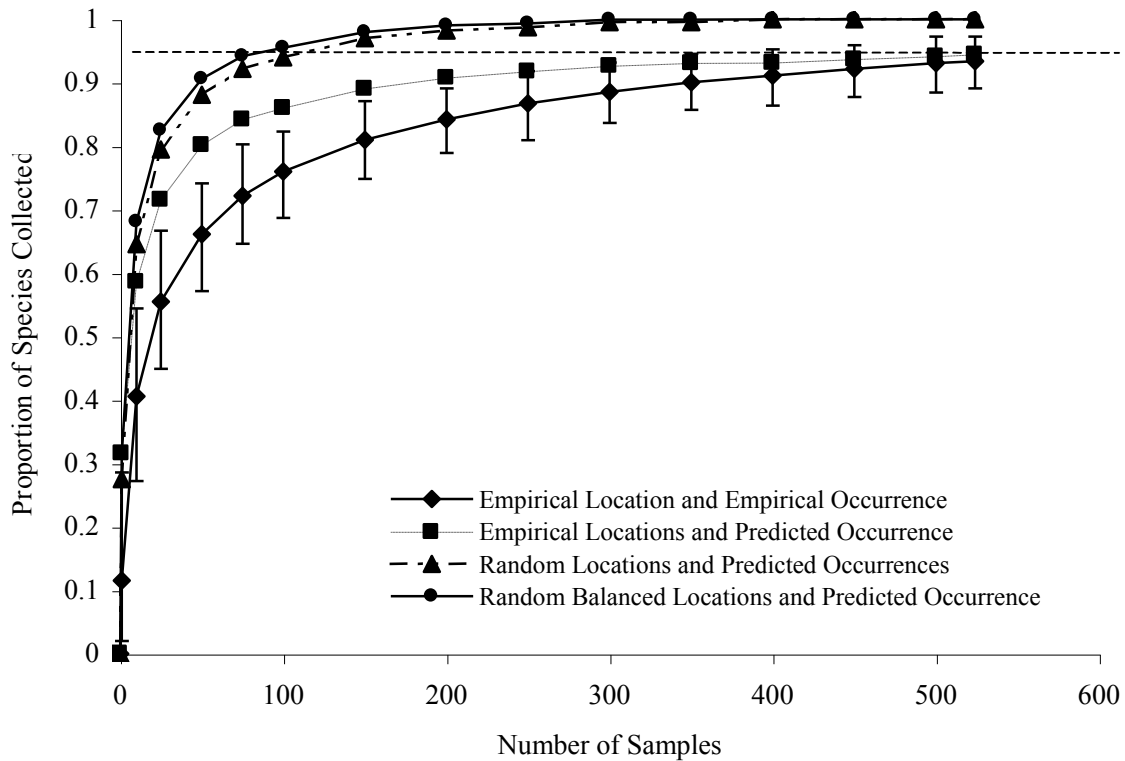


Figure 1.6. Sampling simulation results for A) empirical sample locations and empirical occurrence data ( $N_{95} = 880$ ), B) empirical sample locations and predicted occurrence data ( $N_{95} = 815$ ), C) random samples and predicted occurrence data ( $N_{95} = 142$ ), and D) random balanced samples and predicted occurrence data ( $N_{95} = 138$ ), for  $\leq 524$  samples from the upper Tennessee River basin. One trial was conducted for simulations A and B. Ten trials were conducted for simulations C and D and the mean results are shown. For all simulations 95% bootstrap confidence bounds were calculated using 10,000 sampling iterations at each sample size. The confidence bounds overlap for all simulations at all sample sizes. For clarity only confidence bounds for simulation A are shown. Dashed line is at 95% of known species richness and can be used to estimate  $N_{95}$ .

## **Chapter 2. Using remotely sensed data to predict occurrence of fish species in stream reaches of Virginia: a comparison of modeling methods.**

### INTRODUCTION

Effective management and conservation of biota is often constrained by inadequate knowledge of species locations. Conducting complete censuses to describe species locations is rarely possible, particularly for rare or cryptic species that occur over a large spatial extent, which necessitates sampling and using models that predict their occurrence. Limited knowledge of species distributions and the need to make management and conservation decisions at fine spatial grains and across large spatial extents make predicting species occurrence at fine spatial grains and large spatial extents desirable. Predictions of stream fish occurrence can be used to identify sites to study species' habitats (e.g., instream flow incremental methodology, Bovee 1982), prioritize sampling, and identify spatial gaps in species protection (e.g., gap analyses, Scott et al. 1993). Unfortunately, protocols for predicting the occurrence of stream fish are poorly developed and comparisons of the results of alternative methods are infrequently described, especially at the large spatial extents and fine spatial grains that facilitate management and conservation.

To be useful for guiding other work, comparison of modeling methods should provide information that describes the effects of common challenges to predicting species occurrence and suggest means of addressing those challenges. Common challenges include the need to make predictions at large spatial extents and fine spatial grains to make the predictions useful for management and conservation, and the subsequent necessities of using remotely sensed data to predict species occurrence because they are often the only data available that describe environmental conditions across large spatial extents, and using multipurpose species occurrence data (i.e., found data, Cairns and Pratt 1986) because sampling across large spatial extents is often logistically infeasible. Also, many projects need to predict the occurrence of several species that have different prevalence in the region. Predicting the occurrence of many species in several regions can be logistically taxing and scientists need to account for the effect of species prevalence when predicting species occurrence (Manel et al. 2001).

The spatial extent and grain at which predictions of species occurrences are made affect the usefulness of the predictions. Predictions of species occurrence should be made at ecologically meaningful spatial extents and grains. Prediction extent may coincide with ecological boundaries [e.g., ecoregions (Omernik 1987) or drainages (Angermeier and Winston 1999)] that affect species occurrence patterns. Fortunately, ecologically meaningful spatial extents for predicting fish occurrence are identifiable. Angermeier and Winston (1999) found

statistically significant differences in fish community composition and the identity of environmental factors that regulate stream fish occurrence among major river drainages and regions defined by the intersection of a drainage and a physiographic province (D-P) in Virginia. Leftwich et al. (1997) and Rosenberger (2002) found that the identity of significant predictors of fish occurrence did not transfer across such regional boundaries and Osborne and Suarez-Seoane (2002) found that geographic partitioning of modeling efforts produced better predictions than global models. Such findings indicate that the spatial extent of predictions of occurrence should be limited to regions that share structure and function. That is, model predictions are most accurate where the correlation matrix of predictors and species occurrence is homogeneous, and D-P provide useful, easily identifiable approximations of such regions.

Predictions of species occurrence should also be made at a spatial grain that allows meaningful interpretation of differences in predicted probability of occurrence among prediction units. Prediction grain is frequently based on the grain of sampling of prediction data or chosen arbitrarily. For example, remotely sensed data are often available at large extents and fine grains (e.g., 10-m elevation pixels) allowing precise predictions of plant occurrence (Miller 1986). However, it is unlikely that fishes respond to many predictors described at such fine grains. For example, while it is possible to predict occurrence based on the elevation of a 10-m-long portion of stream, it is likely that other habitat conditions at that location (e.g., whether it is a pool or riffle) are more important than elevation relative to the surrounding 10-m sections of stream.

Many stream fish modeling efforts predict occurrence at the spatial grain of stream reaches and watersheds. These are the two largest units described by Frissell et al. (1986), who refer to them as stream segments and basins, respectively. A reach ( $10^1$  to  $10^3$  m) is a contiguous array of mesohabitat units (e.g., pools and riffles) bounded by geomorphic or hydrologic discontinuities. A watershed ( $10^6$  to  $10^8$  m<sup>2</sup>) is a network of reaches including the contributing land with a single outflow (Angermeier et al. 2002). Fish species occurrence is commonly predicted at the reach and watershed grains using descriptors of these geomorphic units (e.g., Reiman and McIntyre 1995, Kruse et al. 1997, Mastririllo et al. 1997, Reyjol et al. 2001, Porter et al. 2000). Many stream fishes also complete much of their life cycles within a single reach or watershed, making them meaningful units for prediction.

The need to make predictions for large extents at fine grains limits the quantity and quality of data commonly available for use in prediction exercises. Managers need protocols to make the best use of such data and examples of prediction performance to guide expectations. Predictor variables (i.e., independent variables) are commonly limited to those that can be procured via remote sensing or derived from maps. Such predictors allow prediction in regions

and at sites where little or no sampling has been conducted. However, remotely sensed data are currently limited primarily to geological, geomorphological, and land cover descriptors that are frequently correlated because of functional interactions and shared descriptive units (Legendre 1993). For example, watershed shape, drainage density and stream size are commonly used to describe streams, and all can be estimated using watershed area (Knighton 1993). Another common predictor variable, land cover, can be estimated from measures of slope, elevation and geology (Newson 1992). Multipurpose species occurrence data (Cairns and Pratt 1986, see Ponder et al. 2001) are often the only occurrence data available for model building and evaluation. Because such data are collected for a variety of reasons, samples may cover only a portion of the range of environmental variability of the region and be non-randomly distributed through space. Furthermore, not all samples may have the same probability of collecting all species in the sampled region due to differences in sampling effort (Angermeier and Smogor 1995) or sampling efficiency (Peterson and Rabeni 1995, Bayley and Peterson 2001). Greater confidence should be given to records of presence than to absences (Angermeier et al. 2002). I describe variation in the performance of prediction procedures due to the limitations of available data.

Many non-statistical and statistical methods are used to predict species occurrence (see Scott et al. 2002 for examples). Each procedure has assumptions and attributes that affect its suitability for a specific project, given the qualities of the available data. Unfortunately, protocols for predicting occurrence are poorly developed and few comparisons of the utility of available procedures are available (Guisan and Zimmerman 2000). I compare the utility of some commonly used approaches, each with different assumptions and limitations, among regions and species with different data qualities, to help elucidate effective protocols for predicting species occurrence.

In lieu of statistical procedures, descriptions of environmental conditions at sites of known species presence are often used to predict occurrence. Stream classification procedures (SC) are non-statistical abstract summary representations (Goodwin 1999) that are an alternative to statistical prediction methods when species are rare or sampling is inadequate. Descriptors (i.e., predictors) are binned and combinations of descriptors are used to produce stream reach classes. The presence of a species in a reach class is used to predict its occurrence in other locations of that reach class. For example, the South Dakota aquatic gap analysis (<http://wfs.sdstate.edu/sdgap/aquaticgap.htm>) uses SC to predict fish occurrences in 11-digit United States Geological Survey (USGS) Hydrologic Units (HU, <http://nhd.usgs.gov>) and individual reaches at the state extent. Stream classification can describe stream reaches using

remotely sensed data and taxonomic reference books (e.g., Freshwater Fishes of Virginia, Jenkins and Burkhead 1993) often provide predictor information. Stream classifications are free of statistical assumptions, but they can be especially sensitive to the interaction of sampling sufficiency and classification complexity. Also, the necessarily arbitrary construction of SC makes assessing their prediction accuracy difficult and potentially tautological. Therefore, SC are incommensurate with the other procedures, but are discussed below because they are commonly used.

The failure of data to meet the statistical assumptions of prediction methods is often referenced as a cause of poor model performance and is an impetus for the development of alternative methods for prediction. Many (parametric) methods used to predict species occurrence assume identical and independently distributed errors and linear relationships between the predictors and the dependent variable. For example some models based on linear discriminant function analysis (DFA) have these assumptions. Linear discriminant analyses are also constrained to using continuous predictors. Generalized linear models (GLM, e.g., logistic regression) can use continuous or categorical predictors, but are susceptible to overfitting (see below). Methods less constrained by such assumptions, such as classification trees (TRE, Breiman et al. 1984) and artificial neural networks (ANN, Lek and Guegan 1999), appear to be very promising alternatives to parametric approaches and have been used to predict fish occurrence in rivers (Mastririllo et al. 1997, Reyjol et al. 2001) and lakes (Brosse et al. 1999, Olden and Jackson 2001). Classification trees and ANN compare favorably to parametric methods in the few comparisons that have been conducted (e.g., Mannel et al. 1999a, 1999b, Olden and Jackson 2002a). However, these methods may have attributes that limit their utility in some instances and further comparisons are warranted.

How a prediction method is built and applied can affect prediction performance. The prediction performance of some methods, such as GLM, can be affected by using critical values that account for species prevalence. Also, relations between species occurrence and habitat descriptions (i.e., independent variables) can be complex (Jackson et al. 2001). Poor model performance often promotes the use of additional predictors to better describe the suitability of habitat. Few assessments of the effects of using alternative critical values or increasing model complexity on prediction performance are available.

The high rates of imperilment of aquatic species, including fishes (Etnier 1997), have prompted efforts to predict fish occurrence (e.g., aquatic gap analyses), and provide incentive to establish protocols that will promote efficiency, accuracy and communication of results. Predicting fish occurrence provides an excellent opportunity to compare predictive methods and

develop such protocols because stream fish sampling methods are well-developed (Angermeier and Smogor 1995, Peterson and Rabeni 1995) promoting confidence in the data used to build and assess models. Additionally, understanding the sources and severity of sampling errors allows better interpretation and implementation of model results.

Virginia provides exceptional opportunities to compare methods of predicting species occurrence. With nearly 200 species of native freshwater fishes (Jenkins and Burkhead 1993) there is ample opportunity to compare procedures among species with a variety of ecological attributes and distribution patterns. The exceptional diversity of stream fishes in Virginia is largely due to the existence of 11 major drainages and 5 physiographic provinces within the state (Figure 2.1). These geographic units provide easily identified, ecologically meaningful prediction extents that provide a basis for comparing modeling procedure prediction accuracy through space. Virginia's stream fishes have been extensively sampled and the Virginia Department of Game and Inland Fisheries maintains an online database, the Virginia Fish and Wildlife Information Service (VAFWIS, <http://128.173.240.58/build02/fwis/default.asp>), providing access to many georeferenced fish samples so that sample size alone should not limit the predictive utility of procedures. High species richness and high environmental heterogeneity likely requires the building of many models to predict the occurrence of all stream fishes throughout the state, promoting the elucidation of protocols for efficiently and accurately predicting fish occurrence.

My goal is to compare the utility of 5 methods of predicting stream fish occurrence at large spatial extents and fine spatial grains using remotely sensed predictors describing streams and their watersheds and multipurpose species occurrence data. Specific objectives include 1) comparing the effort (time) required to build predictive models, 2) assessing the effect of sample location on the usefulness of stream classification models 3) assessing the effect of critical value selection on prediction accuracy, 4) comparing the accuracy of the statistical prediction methods, and 5) assessing the effect of model complexity on prediction accuracy.

Comparison of methods for predicting species occurrence for a number of disparate species, in environmentally distinct regions, using data of varying qualities should provide a robust comparison. Previous comparisons, while laudable, may be of limited generality due to the use of a single prediction region or species. Comparisons among multiple species and regions should make these analyses useful to those working on a range of prediction problems.

Additionally, the comparison of methods for predicting stream fish occurrence at a fine spatial grain allows investigation of the constraints of model complexity and arbitrary critical value selection on prediction success and model resolution. These issues are commonly confronted in

prediction problems, but they are infrequently investigated or discussed. Doing so will aid in the development of generally applicable protocols for predicting species occurrence.

## METHODS

### *Region and Species Selection*

To promote the generality of results, 7 species from each of 3 drainage-physiography combinations in Virginia were selected as focal species. Eighteen species were selected from 10 genera in 5 families to ensure coverage of a range of ecological attributes and prevalence (i.e., percent of samples in which a species occurs). Drainages are 6-digit USGS hydrologic units (Seaber et al. 1987) and physiographic provinces follow those of Jenkins and Burkhead (1993) and are similar to the ecoregions of Omernik (1987, Figure 2.1). Drainage-physiography combinations were selected as regions of analysis because community composition and factors regulating species distributions vary among them (Angermeier and Winston 1998, 1999) and model accuracy can be improved by building unique models for a species among regions with different characteristics (Osborne and Suarez-Seoane 2002, Stockwell and Peterson 2002). The Holston-Ridge and Valley (HN-RV), Roanoke-Piedmont (RN-PD) and James-Coastal Plain (JM-CP) were selected because they differ in species composition and habitat types, and each region has many usable assemblage samples (N = 138, 262 and 80 respectively).

Species occurrence records were procured from the VAFWIS. This database includes georeferenced collection records and species occurrence records from collection permit reports, scientific reports, and citizen reports. Errors of sampling omission (i.e., false absence) are likely in these data and may be due to failure to sample the appropriate spatial or temporal strata, inadequate sampling effort, or ineffective or inappropriate methods (Angermeier et al. 2002). Errors of sampling commission (false presence) are less likely, but may occur due to species misidentification. Measures of prediction performance include proportion of correct occurrence predictions [sensitivity (SE)], proportion of correct absence predictions [specificity (SP)], and proportion of total correct classifications (CC). Errors of sampling omission affect model performance by erroneously deflating SE and inflating SP. The effect of sampling omission errors is most pronounced for rare species (low prevalence).

Omitting from model building those samples that likely had high rates of sampling omission minimized the effect of sampling errors on model performance. This was accomplished by selecting samples that appeared to provide relatively complete species lists of the local assemblage. Such assemblage samples were selected for these analyses because building DFA, GLM, TRE and ANN require information on both presence and absence of species. Samples

were included if they recorded  $> 2$  species, recorded  $\geq 1$  non-game species, and had reliable location information (i.e., plotted within 50 m of a stream reach). When possible, the sample collector and collection methods used were identified to ensure adequate sampling. When multiple samples of a reach were available the most recent sample was used. Samples were collected from 1965 to 1998.

The USGS National Hydrology Dataset (NHD, USGS 2000) 1:100,000-scale stream reach files were used to define streams. Reaches were identified using the unique code assigned by the NHD. The NHD defined reaches based on the location of confluences, major changes in stream gradient and changes in channel type (e.g., natural or channelized). When identifiable, samples from non-wadeable rivers were omitted from analyses because they were not comparable due to expected differences in species catchability and the sampling gear used (Hubert 1996).

#### *Variable Selection*

Easily procurable data and easily calculable predictors were selected to promote generality of model predictions. Elevation data from the National Elevation Dataset (NED, USGS 2000) were used to delineate the contributing watershed for each reach, and to calculate elevation and slope variables. Median watershed aspect, mean watershed elevation, drainage area, drainage density (cumulative stream length / watershed area), mean stream slope, and shape parameter (elongation ratio, Schumm 1956) were calculated to describe the contributing watershed for each reach. Strahler stream order (Strahler 1957), mean elevation, and stream slope were used to describe reaches. Calculations were made using ARC/INFO 8.0.2 (Environmental Systems Research Institute, Inc.).

Because species samples were collected over a long time period (33 y), temporally variable predictors (e.g., land cover) may have changed between the time of species sampling and the time of predictor measurement, confounding prediction. Use of temporally stable predictors may promote accurate prediction and limit unwarranted inferences. Predictor data from each of the 3 aforementioned regions were assessed for outliers, collinearity, and normality. Because remotely sensed predictors were used, data for the entire region, including sites with and without fish samples were assessed. Predictor characteristics and interactions were assessed with scatterplots, Spearman correlation matrices, and principal components analysis. These analyses were used to determine which variable(s) should be removed to reduce multicollinearity for DFA. Outliers were identified and suitable transformations were applied to reduce heteroscedasticity and improve normality for DFA and GLM.



### *Modeling Methods*

Five methods were selected for comparison; stream classification (SC), linear discriminant function analysis (DFA), generalized linear model (i.e, logistic regression, GLM), classification tree (TRE), and artificial neural network (ANN). Stream classification was selected because it is analogous to the methods commonly used for terrestrial gap analyses (Scott et al. 1993). The SC was developed by using combinations of 3 classes each of reach elevation and reach slope, 4 classes of watershed aspect, and 7 Strahler stream orders. Elevation and slope categories were defined using 33- and 66-percentiles of the values for all reaches in Virginia. This provided high (H), medium (M) and low (L) classes for these variables. Aspect classes reflected north (N), south (S), east (E) and west (W) flowing watersheds. Reaches of Strahler orders 1 through 7 were used, although predictions were made only for stream order < 6 (i.e., wadeable streams).

Discriminant function analysis, GLM, and TRE were selected because they are commonly used to predict species occurrence (Manel et al. 1999a, 1999b, Olden and Jackson 2002a). Artificial neural networks were selected because they are capable of modeling complex nonlinear interactions that may be present (Judd 1990). The DFA models were built using Proc Discrim with the cross-validate option in SAS 8.0.2 (SAS Institute). The GLM and ANN were built using NevProp 4.0, an unsupported freeware version of a feed-forward back-propagation algorithm capable of multiple architectures. Model structure varied substantially among modeling methods. GLM were built with no hidden layers and no interaction or nonlinear terms. GLM were also built with SAS 8.0.2 and compared to those trained using NevProp 4.0 to ensure results similar to those obtained with common statistical packages. The two forms of GLM produced nearly identical results.

Artificial neural networks are statistical methods of exploring a response (error) surface, patterned after biological neural networks. The architecture of ANN commonly includes 3 layers, an input layer of predictor nodes, a layer of hidden nodes that are activated or inactivated depending on the weight they acquire during training, and an output layer of predictions (Figure 2.2). In feed-forward networks the signal is transmitted only from the input layer, through the hidden layer to the output layer. Back-propagation is an algorithm used to train the model by iteratively, through a number of training epochs, identifying gradients along the error surface. The error surface is explored by the ANN in a series of steps, the size of which are referred to as the learning rate. The optimal number of hidden nodes, learning rate, and number of epochs were determined empirically as the values that maximize prediction accuracy with minimal complexity.

ANN were built with a fully connected single hidden layer. I compared models built with 1 to 9 hidden nodes to determine empirically the optimal number of nodes in the hidden layer (i.e., the number after which additional nodes do not improve prediction performance). A maximum of 5000 and a minimum of 100 training epochs were used to train each model. Bootstrapping with a 50% holdout was used to promote generality of the final model and guard against overfitting (Caruana et al. 2000). Learning rate was adjusted and early stopping based on preliminary cross-validation was used to minimize overtraining (Caruana et al. 2000). Detailed descriptions of feed-forward single-hidden-layer ANN in similar applications are provided in Mastorillo et al. (1997), Brosse et al. (1999), Lek and Guegan (1999), Manel et al. (1999b), Reyjol et al. (2001), and Olden and Jackson (2002b).

Models were built following standard heuristic model building practices, including building and comparing multiple models where necessary and appropriate. Initially, a complex model was built using all possible predictors. The statistical significance of individual predictors was assessed. Statistically significant predictors were used to build parsimonious models and their predictions were compared to the complex models. The ability of each method to accurately predict the presence of rare species was a primary concern for this comparison. Therefore, critical values for classification of predicted occurrence were set at the species' prevalence to weight results and reduce errors of omission. Confusion matrices represents instances in a predicted class by columns and instances in the actual class by rows, allowing the easy identification of misclassification (Fielding and Bell 1997). I created confusion matrices for each model and used them to select the final model for each method, species, and region for further comparisons among methods. Final models were the complex or parsimonious model that best predicted occurrence for each method, species and region. Model predictions were mapped using a geographic information system (GIS, ArcView 8.0, Environmental Systems Research Institute, Inc.), and visually compared to available records of species' presence.

#### *Model Evaluation*

##### Modeling Effort

Model building effort was measured by the cumulative processing time required to determine and fit the final model used for prediction. Because building multiple models was frequently required prior to the final model fitting, presented results may be overestimates, but the relative effort required by each method is informative. Effort required by the model building process to create 105 final models (5 model types for each of 7 species in each of 3 regions) was compared.

### Sample Location

Stream classification accuracy was assessed using proportion of stream classes sampled by the available species occurrence data and predictive omission error rate. The proportions of stream classes sampled in each region were recorded. A low proportion of stream classes sampled indicates insufficient sampling and/or an overly complex classification system. The presence of many unsampled stream classes also promotes the use of methods to extrapolate predictions of occurrence among unsampled classes. SC assumes that classes occupied by a species describe the species' habitat and that other sampled instances of those classes should include the species. Creating a confusion matrix for a SC is problematic because the relative value of presence and absence information is uncertain. Sampled classes where a species is not found may be assumed to predict absence, but not all classes may have been sampled and extrapolation of SC results to unsampled classes is not reliable. Therefore, the utility of SC was assessed using the total predictive omission error for all stream classes occupied by a species. The spatial allocation of samples was also visually assessed by plotting them in a GIS (ArcView 8.0, Environmental Systems Research Institute, Inc.).

### Critical Value Selection

Statistical models used to predict species occurrence often provide a predicted probability of occurrence. Predicted probabilities of occurrence  $> 0.5$  usually indicate presence, whereas those  $< 0.5$  usually indicate absence. The value used, commonly 0.5, is referred to as a critical value. Because SE, SP and CC are systematically affected by species prevalence (Guisan and Zimmermann 2000, Manel et al. 2001, Olden et al. 2002), an alternative (to 0.5) critical value for indicating presence is needed to allow for meaningful model comparison. I used the prevalence of each species in the collection data for a region as the critical value, assigning presence when the model predicted the probability of presence  $\geq$  the species prevalence in the region. The effect of using prevalence as the critical value on prediction performance was examined using linear regression and plots of SE, SP and CC at observed values of prevalence for each modeling method.

### Prediction Performance

Olden et al. (2002) recommend using cross-validation methods to assess model performance. Cross-validation was used to calculate SE, SP, and CC for DFA, GLM, TRE, and ANN models. Prediction performance of model types was assessed using the Kappa statistic (Abramowitz and Stegun 1972, Titus et al. 1984). The number and identity of species for which a method produced statistically significant predictions were recorded. Within each region differences among statistical methods in prediction performance measured by SE, SP, and CC

were assessed using analysis of variance (ANOVA). When appropriate, a leave-one-out bootstrap test was used to adjust critical values for multiple ANOVA tests. The effect of species prevalence on prediction performance was assessed using linear regression. Predicted patterns of species presence were also assessed by comparing them to known presences that were not used to build models. Species samples used for this assessment were not suitable for model building because they likely had many errors of sampling omission. However, the additional information they provide on species presence warrants their use.

### Model Complexity

Model overfitting occurs when predictors included in a model are allowed to fit (account for) statistical noise in the dependent variable; in other words, too many predictor variables are used. The effects of overfitting are infrequently explored, but potentially important. Model overfitting can inflate measures of prediction performance unless prediction performance is evaluated with an independent dataset. Model overtraining with ANN is similar to overfitting in that extraneous interaction and nonlinearity terms are allowed to fit statistical noise in the dependent variable. Hereafter overfitting and overtraining are referred to as overfitting. Model overfitting is assumed not to be a problem with DFA or TRE, but it is commonly avoided when using GLM (e.g., Burnham and Anderson 1998). Overfitting is infrequently considered a problem with ANN because measures are commonly taken to prevent it (e.g., early stopping, adjusting learning rates, and/or bootstrapping or cross-validation, Bishop 1995), thereby ensuring statistically valid models. The effect of model overfitting was assessed by building “parsimonious” DFA, GLM, TRE and ANN using only predictors found to be significant (odds ratio > 1.0) in “complex” models based on either GLM or ANN. Prediction performance was compared between complex and parsimonious models for 7 species in the JM-CP by comparing mean SE, SP, and CC values.

## RESULTS

### *Variable Selection*

As expected, all predictor variables were substantially correlated with at least one other predictor, portending difficulties with collinearity (Table 2.1). The strength of correlations varied among regions. In the HN-RV and JM-CP Strahler stream order is strongly positively correlated with watershed area. In the RN-PD and JM-CP reach elevation was strongly positively correlated with reach slope, mean watershed elevation and mean watershed slope; mean watershed elevation was strongly positively correlated with mean watershed slope. In the RN-PD mean watershed

area was strongly negatively correlated with drainage density. The HN-RV had fewer correlations with absolute values  $> 0.7$  than the RN-PD and JM-CP.

PCA was used to identify sources of variation among stream reaches and regions. PCA results show differences in the correlative structure of predictors among regions (Table 2.2). PCA axes 1, 2 and 3 had eigenvalues  $> 1$  for the HN-RV and RN-PD, whereas only PCA axes 1 and 2 were important in describing the JM-CP. The proportion of variability explained by PCA axes and the identity of important components (i.e., predictors) varied among regions (Table 2.2), indicating that the usefulness of predictors may vary among regions. In the HN-RV the first principal component primarily describes reach elevation, watershed area and watershed shape. In the RN-PD the first principal component describes reach elevation, watershed elevation and watershed slope. In the JM-CP, Strahler stream order, watershed elevation, watershed area and watershed slope dominate the first principal component (Table 2.2). These descriptions of the correlation of predictors were used to interpret the results of model predictions. Predictors were not removed from further analyses because there was no information that described which predictors would prove most useful for predicting species occurrence.

#### *Model Evaluation and Comparison*

##### Modeling Effort

There were substantial differences in the required processing time among methods. Processing time was trivial for individual SC, DFA, GLM and TRE. All four methods took only seconds of processing time. Mean processing time for ANN was 18.5 h (range 1.5 – 42.0 h), substantially longer than any other method. Building multiple models for each species and region was required as part of the modeling process for each method. Cumulative processing time to produce final models was also similar among most methods. Multiple models were not built using SC, because little a priori habitat preference information was available for the modeled species. Adjusting SC to minimize error for each species would have been more time consuming than any other method. Cumulative processing time was relatively short for DFA and TRE, rarely requiring more than 4 models to identify a best model for a species. Cumulative processing time was relatively high for GLM, frequently requiring building and comparison of more than 6 models to identify a best model. Cumulative processing time was considerably higher for ANN than for any other method, always requiring more than 10 models to select a best model.

##### Sample Location

SC was inadequate for most species and regions due to inadequate sampling and uninformative classes. Only, 28 of 50, 67 of 163 and 48 of 57 stream classes were sampled in the

HN-RV, RN-PD and JM-CP, respectively. The mean percentage of stream classes sampled was low (mean 55.8; standard deviation 24.6) but higher than the average for 11 similar regions in Virginia (mean 38.0; standard deviation 15.3), indicating that additional samples will frequently be required to produce a useful SC. Omission error of SC, the percentage of sampled classes in which a species was predicted to be present but was not found, was high and variable (mean 51.7%; standard deviation 24.2) for the 3 regions. Omission error also varied among regions. Mean omission error was 64.2% (standard deviation 20.8) for the HN-RV, 63.0% (standard deviation 63.0) for the RN-PD, and 27.8% (standard deviation 11.9) for the JM-CP.

#### Critical Value Selection

Species' prevalence (the percent of samples in which the species was present) ranged from 2.3 to 76.7 (mean = 22.9, standard deviation = 20.3, Table 2.3). Therefore, selection of critical values had a large effect on measures of model accuracy. Use of 0.5 as a critical value frequently produced SE values of zero for rare species for all but the ANN results. The use of prevalence as critical values allowed comparison of SE, SP and CC, primarily by increasing SE at the expense of a small decrease in CC. For example, longhead darter (*Percina macrocephala*) occurred in only 2.2% of samples from the HN-RV. Use of 0.5 as a critical value resulted in SE of 0.0, 3.3, 0.0 and 3.3 and SP of 88.4, 91.5, 100.0, and 100.0 for DFA, GLM, TRE and ANN, respectively. Use of its prevalence (0.02) as a critical value changed SE to 66.7, 100.0, 0.0, and 100.0, while changing SP to 82.2, 88.1, 68.5, 97.8 and 98.5 (Table 2.3).

Adjustment of critical values did not remove all effects of prevalence on prediction performance. Linear regressions of prediction performance versus prevalence indicate differences in the effect of prevalence among methods and regions (Table 2.4). Regression analyses within regions found 7 significant negative relations and 1 significant positive relation between prediction performance and prevalence. Only TRE had a significant positive relation between SE and prevalence. Across all regions CC was significantly negatively related to prevalence for all methods, but significant relations between SE (positive) and SP (negative) with prevalence were found only for TRE. A positive relation between prediction performance and prevalence indicates that higher prevalence improves prediction accuracy measured by the statistic. The effect of prevalence on prediction performance, while frequently statistically significant, was usually small subsequent to adjustment of critical values. The exception to this pattern was for TRE, because TRE had difficulty predicting the occurrence of most rare species (Table 2.3). The use of prevalence as critical values allowed for the prediction of presence of rare species. However, the presence and absence of common species was still predicted more accurately than that of rare species for all methods (Table 2.3).

### Prediction Performance

Prediction performance was relatively high for all methods (Table 2.5). All methods (DFA, GLM, TRE and ANN) correctly predicted occurrence greater than expected by chance (Kappa  $p < 0.10$ ) for most species (Table 2.3). Calculation of Kappa for TRE models was not possible for 11 of 21 species-region combinations because the models did not correctly predict any presences. A pattern of prediction error was common among regions for all measures (Table 2.5). Although statistical significance varied among regions, SE was always lowest for TRE, followed by DFA, GLM and ANN. Similarly, SP and CC were always lowest for DFA followed by GLM, TRE and ANN. Additionally, the accuracy of presence predictions was also assessed using ancillary presence data. Species presence data were from single-species collections (e.g., museum collections, Shaffer et al. 1998 and Ponder et al. 2001) not suitable for model fitting. Species presence locations were plotted on maps and compared with predictions. Similarly, SE of the best model for each species was further assessed by using such collections for the modeled species. Overall, 78.4% of presences were correctly predicted. Additionally, presence of rare species, herein those in  $< 10\%$  of samples, were correctly predicted in 63.4% of samples.

### Model Complexity

Many models were likely overfit. However, paired t-tests of the difference in prediction accuracy between complex and parsimonious models for the JM-CP indicate that the complex GLM and ANN models provide significantly higher mean SE, SP and CC than the respective parsimonious models (Figure 2.3). The only exception to this pattern was for TRE models, where parsimonious models produced higher SP than complex models, at the cost of greatly reduced SE. This consistent pattern of tradeoff between error types suggests that relations between prediction accuracy and model complexity may be due to differences in prediction resolution (how well the model can differentiate among prediction units) of model methods and the resulting differences in how the methods allocate errors. For example, methods might differ in how well they can differentiate among prediction units (e.g., stream reaches) that have many similar attributes. A method that has low resolution will predict similar probabilities of occurrence for reaches that have several similar attributes (predictors), but that may differ in one or more attributes that affect a species occurrence.

Because predictors are spatially correlated, predictions may also be spatially correlated and fail to differentiate among spatially proximate reaches. That is, prediction resolution may vary among modeling methods and this may affect prediction accuracy. To assess this variation, maps were generated in ArcView 8.0 using results from each modeling method. Examining the discontinuity of presence predictions visually assessed prediction resolution. That is, the lengths

of contiguous reaches with predictions of species presence were visually assessed. Increasing model complexity usually increased prediction resolution for all methods, but often did not substantially improve prediction accuracy.

## DISCUSSION

The management of species and assemblages requires dependable knowledge of their locations, and large-scale surveys can rarely provide sufficient sampling to delimit species' occurrences. The effect of insufficient sampling is particularly problematic when species occurrence is discontinuous, because managers may need to delimit presence and absence rather than only the extent of occurrence (Angermeier et al. 2002). Fortunately, sample sizes required to build accurate models can be surprisingly small (Stockwell and Peterson 2002) and dependable knowledge of species locations can be obtained from predictive models that can be built with multipurpose data (e.g., Bozek and Hubert 1992, Mastrorillo et al. 1997, Rahel and Nibbelink 1999, Filipe et al. 2002). A scientific and management conundrum is the need to predict the occurrence of rare species (Rabinowitz 1981, Bevill and Louda 1999), whose low prevalence make model building and model assessment difficult (Karl et al. 2000). Protocols for selecting modeling methods and applying modeling results are not well developed, especially for rare species.

### *Variable Selection and Assessment*

Data quality (e.g., location and abundance of samples and number and quality of predictors) can constrain the utility of prediction methods and the predictions themselves (Peterson and Cohoon 1999). Using multipurpose species occurrence data (Cairns and Pratt 1986, Ponder et al. 2001) is often a prerequisite for obtaining the required spatial coverage of predictor variables for studies conducted at large spatial extents, which makes data quality problematic. My data are from large heterogeneous regions with recognized boundaries, distinct faunas (Angermeier and Winston 1999), and substantial differences in prevalence of shared species. These circumstances suggest model interpretation will likely prove unfruitful and model transfer ineffective. Individual models need to be built for each species in each region to accurately predict occurrence. This is a likely scenario for most large-scale studies where biologically meaningful boundaries are crossed (Allen et al. 1984, Wiens et al. 1985, Kotliar and Wiens 1990, Townsend 1996). At large spatial extents only indirect distal gradients (Austin 2002), with remote linkages to the mechanisms affecting occurrence, are available as predictor variables (Naiman et al. 1992). Interpretations of predictor variable importance should be made cautiously and expectations for prediction performance should be reasonable. Fortunately, such



predictors often prove adequate (Bozek and Hubert 1992, Rahel and Nibbelink 1999, Filipe et al. 2002) for building static models of species occurrence (Guisan and Zimmermann 2000). Higher resolution predictor variable with more-direct linkages to the mechanisms affecting occurrence might improve prediction performance, but should not be assumed.

### *Model Evaluation and Comparison*

#### Modeling Effort

The effort (time) required to build a model becomes increasingly important as the number of models to be built mounts. For example, more than 4,000 models are required to predict the occurrence of each native stream fish species in each region (i.e., drainage-physiography combination) in which it is known to occur in Virginia. Building more models may be required since the modeling process often includes building and evaluating several models prior to selecting a final model (Loehle 1983, Hilborn and Mangel 1997, and Starfield 1997, Anderson et al. 2000). SC, DFA, GLM, and TRE required very little computing time (< 10 min), whereas ANN required a mean of approximately 18.5 h of computing time. The apparently excessive time required to build ANN may preclude their frequent use in favor of faster methods. The use of bootstrapping to estimate variable importance in the ANN (Olden and Jackson 2001, 2002b) is optional and its omission can substantially reduce computing time, but at the cost of reduced interpretability. Selecting among DFA, GLM, and TRE as a preferred prediction method is best accomplished by comparing prediction performance rather than modeling effort.

#### Sample Location

Sample coverage of the full range of environmental variability is desirable for all modeling methods (Nicholls 1989, Leftwich et al. 1997) and is particularly important for SC. The only data quality assumptions of SC are correct occurrence data (common to all methods) and sufficient sampling to cover the range of important (i.e., useful) variable combinations. The low proportions of stream classes sampled in the data used here indicate that SC is unsuitable for predicting fish occurrence. Although frequently applied in terrestrial studies (Scott et al. 1993), SC does not allow prediction of occurrence in unsampled classes. Given the number and location of samples, greater sample coverage of classes may be obtained by decreasing the number of stream classes, which decreases prediction resolution and increases trite predictions (Scott et al 1996, Karl et al. 2000, Stockwell and Peterson 2002). For example, given that all examples of a suitable class within a region will probably not be occupied, it is more informative if a species is sampled in 3 of 30 classes than if that species is found in 1 of 10 classes. Although the proportion of classes occupied is identical in both classification schemes, the prediction resolution and the ability to differentiate among reaches is greater in the former, which allows

inferences regarding the importance of individual variables, and a more precise view of the species' habitat and predicted distribution.

Frequent omission errors confirm the poor performance of SC. While there are many uses for SC systems (Naiman, et al. 1992, Goodwin 1999, Smith et al. 2002), methods that convert continuous data to categorical data to predict species occurrence share the problem of limited prediction and they should be used only when better options are not tractable.

#### Critical Value Selection

Low species prevalence is common (Gaston 1994, Gotelli and Graves 1996, Karl et al. 2000) and methods of making better predictions of occurrence for rare species are needed. Among Virginia drainage-physiography regions a mean of 31.1% (range 8.7 – 48.3%) of native stream fishes are in  $\leq 5\%$  of samples. Such a high proportion of species at low prevalence portend difficulty in accurate prediction of occurrence (Stockwell and Peterson 2002) and meaningful assessment of results. Rare species are often excluded from analyses (e.g., Gauch 1982, Karl et al. 2000, Oberdorff et al. 2000, Porter et al. 2000). Exclusion is not an option for conservation-oriented studies where rare species are the focus (see Boone and Krohn 2000 and citations therein). Varying the prediction critical value (i.e., cutoff value or decision threshold) from the standard 0.5 (e.g., Carroll et al. 1999, Porter et al. 2000, Olden and Jackson 2001) can adjust for the effect of prevalence on prediction accuracy by increasing sensitivity and decreasing specificity, thereby allowing comparison of results. Because many of the species modeled here had low prevalence and sampling errors of omission were probably frequent, I used the species' sample prevalence in the region as the critical value. This adequately adjusted sensitivity, specificity and correct classification rates to allow comparison of methods (Fielding and Bell 1997, Filipe et al. 2002).

Adjustment of critical values for specific applications should be encouraged to enhance model utility. For example, if a model is being applied to predict the occurrence of a common species, the locations of absence may be more interesting than the locations of presence. A large critical value will decrease SE and increase SP, allowing more confident prediction of absence. Alternatively, misclassifying even a few presences may be a critical error in the management of rare species. Using a small critical value to predict the occurrence of a rare species will increase SE at the cost of lower SP, which allows more confident identification of presence. Critical values should be selected to make the best use of model predictions.

#### Prediction Performance

Empirical assessments (e.g., Lek et al. 1996) are adding to our understanding of model performance, given data of various qualities, but results are variable and additional assessments

are warranted (Guisan and Zimmerman 2000). Greater robustness to departures from assumptions and greater flexibility commonly allow ANN and TRE to outperform GLM and GLM to outperform DFA (Prager and Fabrizio 1990). Generally, TRE perform better than GLM and DFA for common species, but TRE perform poorly for rare species. For example, Olden and Jackson (2002a) found that TRE and ANN outperformed GLM and DFA at predicting fish species occurrence in lakes. Similarly, Oakes et al. (2005) found that ANN generally performed better than TRE and DFA at predicting stream fish occurrence and Mastrorillo et al. (1997) found that ANN outperformed DFA in cases with nonlinearly related variables. However, Manel et al. (1999a and 1999b) found conflicting results from similar comparisons and Moisen and Frescino (2002) found only small improvement over simple linear models by using complex, nonlinear models for predicting forest characteristics. Multicollinearity and nonlinear predictor effects may be responsible for some of the reported differences in prediction accuracy among methods (Carroll et al. 1999), but how methods are applied and assessed may be as important as which method is used.

Differences in findings among such comparisons may be due, in part, to differences in species prevalence and how it is addressed during model building and assessment. Methods for assessing prediction performance should be improved, particularly for models that predict the occurrence of rare species. Assessment of model prediction performance can be improved by using presence data not used (or suitable) for modeling to assess the accuracy of predictions of presence (i.e., SE), which is the most variable component of prediction error and arguably the most important accuracy metric for rare species. Overall, my models correctly predicted 78.4% of presences from data not used for building these models. Additionally, presence of rare species, herein those in < 10% of samples, were correctly predicted in 63.4% of samples. For comparison, Olden and Jackson (2002) report mean cross-validation errors of SE = 63.5, SP = 56.6, and CC = 63.4 for predicting fish occurrence in lakes, suggesting that my models performed similarly to other models and that ancillary data can be used to assess model predictions.

#### Model Complexity

Most of my models are likely overfit. Overfitting is problematic when inferences are made regarding variable importance and the model is used in a region where it was not fit (i.e., transferred). However, overfit models might provide useful predictions in the region for which it was fit. Model parsimony is promoted to ease interpretation and enhance generality (i.e., transfer to other regions). Interpretation of model parameters and model transfer are frequently unsuccessful and are not goals of this work, which reduces the imperative for parsimony. Overfitting may also hinder model comparisons by inflating measures of SE, SP and CC, but not

all methods are equally susceptible to overfitting. Because discriminating between presence and absence sites was the use for these models, all predictors were used in all methods. However, unusually high SE, SP, and CC values for GLM and ANN call into question comparisons of results. This is a problem not directly addressed in other studies, where well fit models were assumed and possible due to high species prevalence and relatively large sample sizes (e.g., Manel et al. 1999a, 1999b). Olden and Jackson (2002b) recognized similar issues of prevalence, but did not address them in detail.

Effects of model overfitting are usually minimized by variable elimination. My analysis of the effect of model complexity usually found 3 or fewer statistically significant predictors (odds ratio > 1.0) in the GLM and ANN models and 4 or fewer variables in the TRE models. The identity of useful predictors varies among region and among species, inhibiting a priori predictor selection. This is a common phenomenon and likely results from statistical artifacts related to insufficient sampling and species prevalence, multicollinearity of predictors, as well as from real differences in predictor importance among regions and species. Exploratory analyses of the effects of model simplification to include only significant predictors indicate that simplification can substantially reduce SE, SP and CC (Figure 2.3). Additionally, the prediction resolution is reduced. That is, the less complex models do not differentiate among reaches that are environmentally similar and often spatially proximate because predictors are spatially autocorrelated and imprecise. Statistically sound (i.e., well-fit) models may allow for simple assessment, but the utility of the model should not be sacrificed (Hakanson 1995). Model complexity is not best chosen based only on currently measurable model performance (Dunbar 1980, but see Anderson et al. 2000). Rather, parsimony should be balanced with the ability of the model to differentiate among sites at the desired resolution.

These analyses suggest that it may be useful to apply reasonably overfit models and adjusted critical values to specific prediction problems to make more useful predictions. Overfit models may be able to differentiate among sites more readily than simple, statistically valid models. Adjusted critical values can then be used to classify occurrence with the desired level of SE or SP. However, care should be taken to avoid unwarranted inference from such statistically unsupported models. I suggest that this is the default approach for rare species when non-statistical models such as SC are used. Managers and scientists often record copious information on the environmental attributes where rare species are present and then predict that similar habitat elsewhere may be suitable for the species, although such a complex model is not statistically supported. Ensuring model utility at the expense of parsimony may be a preferred strategy

regardless of the method of prediction. How the model is built and its predictions are applied may be more important to model utility than the selection of the method used.

## CONCLUSIONS

This study was initiated to compare 5 common methods of predicting species occurrence. The purpose of the comparison was to select an efficient method of predicting stream fish occurrence using multipurpose species collections and easily procured, temporally stable stream descriptors. Comparisons were based on the importance of the time required to build the models and model prediction accuracy. Particular consideration was given to the ability of the methods to predict the occurrence of rare species, an often-avoided issue. Predicting the occurrence of rare species is important because rarity is a common species attribute and models that predict the occurrence of rare species may be particularly valuable, but difficult to construct.

Stream classification methods made poor use of available data and required excessively large numbers of samples to make predictions at a useful resolution. Simplifying the SC by using fewer predictors or classes of predictors could reduce the sampling required, but the resulting SC might provide little discrimination ability and high prediction error rates. Linear discriminant function analyses, generalized linear models, classification trees and artificial neural networks all provide useful predictions of occurrence and DFA, GLM and TRE required relatively little time to build compared to SC and ANN. However, shortcomings were identified for each method. Linear discriminant function analyses cannot make use of all available (i.e., categorical) predictors and generally performed poorly compared to the other methods. Generalized linear models performed well, but require sufficient sampling to avoid partial separation of species occurrences among predictors and were prone to overfitting. Classification trees provided some accurate predictions, but were unable to predict presence accurately for rare species. Excessive computing time, high sample size requirements, and the high degree of technical knowledge required makes the frequent use of ANN unlikely at present. While rare or ubiquitous species are extremely susceptible to chance predictions (Olden and Jackson 2002a), my results indicate that it is possible to accurately predict the occurrence of rare species by using easily procured predictors in DFA, GLM and ANN models.

Failure to find substantial and consistent differences in prediction success among methods using limited predictor data (i.e., simple models) suggests that species occurrence data and predictor data quantity and quality, rather than the statistical method used are the factors most responsible for the limited success of predicting species occurrence. Differences in prediction success among methods with complex models suggest that it is possible to improve prediction

success with complex modeling methods if sufficient data are available. However, the quantity and quality of data likely to be available may preclude the building of complex models, without encountering statistical difficulties such as overfitting. Since effort may be negligible for many methods or situations, it is feasible to try different methods and use the one most effective in a given situation.

While improper use of statistical methods is abhorrent in most situations, prediction at a useful spatial resolution may necessitate overfitting models and adjusting critical values to maximize sensitivity for rare species. Model evaluation may also prove difficult due to low species prevalence. The dearth of occurrence data for many species will likely be the limiting factor in predicting their occurrence. Assessment of model predictions is problematic for many species, but better use can be made of available technology and data by invoking a little creativity. When necessary, statistical assumptions and standard methodologies may be cautiously set aside in favor of greater utility, provided the model is not misapplied and interpretations of findings are cautious. When species are rare or the study spatial extent is large, model complexity and concomitant overfitting may be a preferred alternative to statistically correct but uninformative models. Selection of critical values should be based on the desired balance of sensitivity versus specificity with the species' prevalence and the model application in mind. Predictions from complex models that make predictions at a fine grain with high resolution can be adjusted by altering critical values to more accurately depict what is known about species occurrence patterns. The resulting predictions can be very useful and may serve as the basis for more accurate and statistically sound future models or for interim management actions.

#### ACKNOWLEDGMENTS

This work greatly benefited from the assistance of many people. In particular I thank Laura Delarche Roghair and Scott Klopfer of the Conservation Management Institute at Virginia Tech for outstanding GIS assistance and James Peterson, Julian Olden, and Eric Smith for helpful statistical advice. The Virginia Department of Game and Inland Fisheries funded this work.

## LITERATURE CITED

- Abramowitz, M. and I. A. Stegun. 1972. Handbook of mathematical functions. Dover Publications, Inc. New York.
- Allen, T. F. H., R. V. O'Neill, and T. W. Hoekstra. 1984. Interlevel relations in ecological research and management: some working principles from hierarchy theory. USDA Forest Service General Technical Report RM-110.
- Anderson, D. A., K. P. Burnham, W. L. Thompson. 2000. Null hypothesis testing: problems, prevalence, and an alternative. *Journal of Wildlife Management* 64(4): 912-923.
- Angermeier, P. L., K. L. Krueger, and C. A. Dolloff. 2002. Discontinuity in stream-fish distributions: implications for assessing and predicting species occurrence. Pages 519-527 in J. M. Scott et al., editors. Predicting species occurrences: issues of accuracy and scale. Island Press, Washington D.C.
- Angermeier, P. L., and M. R. Winston. 1998. Local vs. regional influences on local diversity in stream fish communities of Virginia. *Ecology* 79: 911-927.
- Angermeier, P. L., and M. R. Winston. 1999. Characterizing fish community diversity across Virginia landscapes: prerequisite for conservation. *Ecological Applications* 9: 335-349.
- Angermeier, P. L., and R. A. Smogor. 1995. Estimating number of species and relative abundances in stream-fish communities: effects of sampling effort and discontinuous spatial distributions. *Canadian Journal of Fisheries and Aquatic Sciences* 52: 936-949.
- Austin, M. P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling* 157: 101-118.
- Bayley, P. B., and J. T. Peterson. 2001. An approach to estimating probability of presence and richness of fish species. *Transactions of the American Fisheries Society* 130: 620-633.
- Bevill R. L., and S. M. Louda. 1999. Comparison of related rare and common species in the study of rare plants. *Conservation Biology* 13(3): 493-498.
- Bishop, C. M. 1995. Neural networks for pattern recognition. Oxford University Press, Oxford.
- Boone, R. B. and W. B. Krohn. 2000. Predicting broad-scale occurrences of vertebrates in patchy landscapes. *Landscape Ecology* 15: 63-74.
- Bovee, K. D. 1982. A guide to stream habitat analysis using the instream flow incremental methodology. U.S. Fish and Wildlife Service, Instream Flow Information Paper 12 FWS/OBS/82/86, Washington, D.C.
- Bozek, M. A., and W. A. Hubert. 1992. Segregation of resident trout in streams as predicted by three habitat dimensions. *Canadian Journal of Zoology* 70: 886-890.
- Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. Classification and regression trees. Wadsworth, Belmont, California, USA.
- Brosse, S., J. Guegan, J. Tourenq, and S. Lek. 1999. The use of artificial neural networks to assess fish abundance and spatial occupancy in the littoral zone of a mesotrophic lake. *Ecological Modelling* 120: 299-311.
- Burnham, K. P., and D. R. Anderson. 1998. Model selection and inference: a practical information-theoretic approach. Springer-Verlag, New York. 353 p.
- Cairns, J., Jr., and J. R. Pratt. 1986. Developing a sampling strategy. *American Society for Testing and Materials* 894: 168-186.
- Carroll, C., W. J. Zielinski, and R. F. Noss. 1999. Using presence-absence data to build and test habitat models for the fisher in the Klamath region, U.S.A. *Conservation Biology* 13: 1344-1359.
- Caruana, R., S. Lawrence, and L. Giles. 2000. Overfitting in neural nets: backpropagation, conjugate gradient, and early stopping. *Neural Information Processing Systems*, Denver, Colorado.

- Dunbar, M. J. 1980. The blunting of Occam's razor, or to hell with parsimony. *Canadian Journal of Zoology* 58: 123-128.
- Etnier, D. A. 1997. Jeopardized southeastern freshwater fishes: a search for causes. *In* G. W. Benz and D. E. Collins, editors. *Aquatic fauna in peril: the southeastern perspective*. Special Publication 1, Southeastern Aquatic Research Institute, Lenz Design & Communications, Decatur, Georgia.
- Fielding, A. H. and J. F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* 24: 38-49.
- Filipe, A. F., I. G. Cowx, and M. J. Collares-Pereira. 2002. Spatial modeling of freshwater fish in semi-arid river systems: a tool for conservation. *River Research and Applications* 18: 123-136.
- Frissell, C. A., W. J. Liss, C. E. Warren, and M. D. Hurley. 1986. A hierarchical framework for stream habitat classification: viewing streams in a watershed context. *Environmental Management* 10: 199-214.
- Gaston, K. J. 1994. *Rarity*. Chapman and Hall, London, U.K.
- Gauch, H. G. 1982. *Multivariate analysis in community ecology*. Cambridge University Press, Cambridge.
- Goodwin, C. N. 1999. Fluvial classification: Neanderthal necessity or needless normalcy. *Wildland Hydrology* 229-236.
- Gotelli, N. J., and G. R. Graves. 1996. *Null models in ecology*. Smithsonian Institution Press, Washington, D.C.
- Guisan, A., and N. E. Zimmerman. 2000. Predictive habitat distribution models in ecology. *Ecological Modelling* 135: 147-186.
- Hakanson, L. 1995. Optimal size of predictive models. *Ecological Modelling* 78(1995): 195-204.
- Hilborn, R. and M. Mangel. 1997. *The ecological detective: confronting models with data*. Princeton University Press, Princeton, New Jersey.
- Hubert, W. A. 1996. Passive capture techniques. Pages 157-192 *in* B. R. Murphy and D. W. Willis, editors. *Fisheries techniques*, 2nd edition. American Fisheries Society, Bethesda, Maryland.
- Jenkins, R. E., and N. M. Burkhead. 1993. *Freshwater fishes of Virginia*. American Fisheries Society, Bethesda, Maryland.
- Judd, S. J. 1990. *Neural network design and the complexity of learning*. MIT Press, Cambridge, Massachusetts.
- Karl, J. W., P. J. Heglund, E. O. Garton, J. M. Scott, N. M. Wright, and R. L. Hutto. 2000. Sensitivity of species-habitat relationship model performance to factors of scale. *Ecological Applications* 10(6): 1690-1705.
- Knighton, D. 1993. *Fluvial forms and processes*. Routedledge, Chapman and Hall, New York.
- Kotliar, N. B., and J. A. Wiens. 1990. Multiple scales of patchiness and patch structure: a hierarchical framework for the study of heterogeneity. *Oikos* 59: 253-260.
- Kruse, C. G., W. A. Hubert, and F. J. Rahel. 1997. Geomorphic influences on the distribution of Yellowstone cutthroat trout in Absaroka Mountains, Wyoming. *Transactions of the American Fisheries Society* 126: 418-427.
- Leftwich, K. N., P. L. Angermeier, and C. A. Dolloff. 1997. Factors influencing behavior and transferability of habitat models for a benthic stream fish. *Transactions of the American Fisheries Society* 126: 725-734.
- Legendre, P. 1993. Spatial autocorrelation: trouble or a new paradigm? *Ecology* 74: 1659-1673.



- Lek, S., and J. F. Guegan. 1999. Artificial neural networks as a tool in ecological modeling, an introduction. *Ecological Modelling* 120: 65-73.
- Lek, S., M. Delacoste, P. Baran, I. Dimopoulos, J. Lauga, S. Aulagnier. 1996. Application of neural networks to modeling nonlinear relationships in ecology. *Ecological Modelling* 90: 39-52.
- Loehle, C. 1983. Evaluation of theories and calculation tools in ecology. *Ecological Modelling* 19: 239-247.
- Manel, S., H. C. Williams, and S. J. Ormerod. 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology* 38: 921-931.
- Manel, S., J. M. Dias, S. T. Buckson, and S. J. Ormerod. 1999b. Alternative methods for predicting species distribution: an illustration with Himalayan river birds. *Journal of Applied Ecology* 36: 734-747.
- Manel, S., J-M. Dias, and S. J. Omerod. 1999a. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modeling* 120: 337-347.
- Mastrorillo, S., S. Lek, F. Dauba, and A. Belaud. 1997. The use of artificial neural networks to predict the presence of small-bodied fish in a river. *Freshwater Biology* 38: 237-246.
- Miller, R. I. 1986. Predicting rare plant distribution patterns in the southern Appalachians of the south-eastern U.S.A. *Journal of Biogeography* 13: 293-311.
- Moisen, G. G., and T. S. Frescino. 2002. Comparing five modeling techniques for predicting forest characteristics. *Ecological Modelling* 157: 209-225.
- Naiman, R. J., D. G. Lonzarich, T. J. Beechie, and S. C. Ralph. 1992. General principles of classification and the assessment of conservation potential of rivers. Pages 93-123 in P. J. Boon, P. Calow, and G. E. Petts, editors. *River Conservation and Management*, John Wiley and Sons Ltd.
- Newson, M. D. 1992. *Land, water, and development: river basin systems and their sustainable management*. Routedledge, Chapman and Hall, New York.
- Nicholls, A. O. 1989. How to make biological surveys go further with generalised linear models. *Biological Conservation* 50: 51-75.
- Oberdorff, T., D. Pont, B. Hugueny, and D. Chessel. 2001. A probabilistic model characterizing fish assemblages of French rivers: a framework for environmental assessment. *Freshwater Biology* 46: 399-417.
- Olden, J. D., and D. A. Jackson. 2001. Fish-habitat relationships in lakes: gaining predictive and explanatory insight using artificial neural networks. *Transactions of the American Fisheries Society* 130: 878-897.
- Olden, J. D., and D. A. Jackson. 2002a. A comparison of statistical approaches for modelling fish species distributions. *Freshwater Biology* 47: 1976-1995.
- Olden, J. D., and D. A. Jackson. 2002b. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling* 154: 135-150.
- Olden, J. D., D. A. Jackson, and P. R. Peres-Neto. 2002. Predictive models of fish species distributions: a note on proper validation and chance predictions. *Transactions of the American Fisheries Society* 131: 329-336.
- Omernik, J. M. 1987. Ecoregions of the conterminous United States. *Annals of the Association of American Geographers* 77: 118-125.
- Osborne, P. E., and S. Suarez-Seoane. 2002. Should data be partitioned spatially before building large-scale distribution models. *Ecological Modelling* 157: 249-259.
- Peterson, A. T., and K. P. Cohoon. 1999. Sensitivity of distributional prediction algorithms to geographic data completeness. *Ecological Modelling* 117(1999): 159-164.

- Peterson, J. T., and C. F. Rabeni. 1995. Optimizing sampling effort for sampling warmwater stream fish communities. *North American Journal of Fisheries Management* 15: 528-541.
- Ponder, W. F., G. A. Carter, P. Flemons, and R. R. Chapman. 2001. Evaluation of museum collections data for use in biodiversity assessment. *Conservation Biology* 15(3): 648-657.
- Porter, M. S., J. Rosenfeld, and E. A. Parkinson. 2000. Predictive models of fish species distribution in the Blackwater drainage, British Columbia. *North American Journal of Fisheries Management* 20: 349-359.
- Prager, M. H., and M. C. Fabrizio. 1990. Comparison of logistic regression and discriminant analysis for stock identification of anadromous fish, with applications to striped bass and American shad. *Canadian Journal of Fisheries and Aquatic Sciences* 47: 1570-1577.
- Rabinowitz, D. 1981. Seven forms of rarity. Pages 205-217 in H. Synge, editor. *The Biological Aspects of Rare Plant Conservation*. John Wiley & Sons. New York.
- Rahel, F. J., and N. P. Nibbelink. 1999. Spatial patterns in relations among brown trout (*Salmo trutta*) distribution, summer air temperature, and stream size in Rocky Mountain streams. *Canadian Journal of Fisheries and Aquatic Science* 56(Supplement 1): 43-51.
- Reiman, B. E., and J. D. McIntyre. 1995. Occurrence of bull trout in naturally fragmented habitat patches of varied size. *Transactions of the American Fisheries Society* 124: 285-297.
- Resetarits, W. J., Jr., and J. Bernardo. (editors) 1998. *Experimental ecology: issues and perspectives*. Oxford University Press, Inc. New York.
- Reyjol, Y., P. Lim, A. Belaud, and S. Lek. 2001. Modelling of microhabitat used by fish in natural and regulated flows in the river Garonne (France). *Ecological Modelling* 146: 131-142.
- Rosenberger, A. 2002. Multi-scale patterns of habitat use by Roanoke logperch (*Percina rex*) in Virginia rivers: a comparison among populations and life stages. PhD dissertation. Virginia Tech.
- Schumm, S. A. 1956. Evolution of drainage systems and slopes in badlands at Perth Amboy, New Jersey. *Geological Society of America Bulletin* 67: 597-646.
- Scott, J. M., P. J. Heglund, M. J. Morrison, J. B. Haufler, M. G. Raphael, W. A. Wall, and F. B. Samson, Editors. 2002. *Predicting species occurrences: issues of scale and accuracy*. Island Press, Washington, D. C.
- Scott, J. M., and eleven co-authors. 1993. Gap analysis: a geographic approach to protection of biological diversity. *Wildlife Monographs* 123: 1-41.
- Scott, J. M., M. Jennings, R. G. Wright, and B. Csuti. 1996. Landscape approaches to mapping biodiversity. *BioScience* 46: 77-78.
- Seaber, P.R., Kapinos, F.P., and Knapp, G.L. 1987. Hydrologic unit maps: U.S. Geological Survey water-supply paper 2294.
- Shaffer, H. B., R. N. Fisher, and C. Davidson. 1998. The role of natural history collection in documenting species declines. *Trends in Ecology and Evolution* 13: 27-30.
- Smith, R. K., P. L. Freeman, J. V. Higgins, K. S. Wheaton, T. W. FitzHugh, A. A. Das, and K. J. Ernstrom. 2002. Priority areas for freshwater conservation action: a biodiversity assessment of the southeastern United States. *The Nature Conservancy*.
- Starfield, A. M. 1997. A pragmatic approach to modeling for wildlife management. *Journal of Wildlife Management* 61: 261-270.
- Stockwel, D. R. B., and A. T. Peterson. 2002. Effects of sample size on accuracy of species distribution models. *Ecological Modelling* 148(2002): 1-13.
- Strahler, A. N. 1957. Quantitative analysis of watershed geomorphology. *Transactions of the American Geographical Union* 38: 913-920.
- Titus, K., J. A. Mosher, and B. K. Williams. 1984. Chance-corrected classification for use in discriminant analysis. *American Midland Naturalist* 136: 345-375.

- Townsend, C. R. 1996. Concepts in river ecology: pattern and process in a river hierarchy. *Archiv fur Hydrobiologie Supplement* 113: 3-21
- Wiens, J. A., C. A. Crawford, and J. R. Gosz. 1985. Boundary dynamics: a conceptual framework for studying landscape ecosystems. *Oikos* 45: 421-427.

Table 2.1. Spearman's rho coefficients describing correlation between watershed and stream-reach descriptors used to predict fish occurrence in the Holston-Ridge and Valley (N=972 reaches), Roanoke-Piedmont (N=1471 reaches) and James-Coastal Plain (N=614 reaches) regions in Virginia. Descriptors include stream order (Strahler), mean reach elevation (relev), reach slope (rslope), median watershed aspect (wasp), mean watershed elevation (welev), watershed area (area), watershed shape parameter (shape), drainage density (dens), and average watershed slope (wslope). All correlations are significant ( $< 0.05$ ) due to large sample sizes.

<b>Holston-Ridge and Valley</b>								
	<u>Strahler</u>	<u>relev</u>	<u>rslope</u>	<u>wasp</u>	<u>welev</u>	<u>area</u>	<u>shape</u>	<u>dens</u>
relev	-0.313							
rslope	-0.487	0.138						
wasp	0.108	0.327	-0.054					
welev	0.281	0.689	-0.086	0.348				
area	0.882	-0.287	-0.480	0.134	0.273			
shape	-0.516	0.309	0.125	-0.025	-0.077	-0.441		
dens	-0.214	-0.079	0.223	-0.163	-0.192	-0.514	-0.308	
wslope	-0.081	-0.095	0.443	-0.095	0.147	-0.112	-0.040	0.102
<b>Roanoke-Piedmont</b>								
	<u>Strahler</u>	<u>relev</u>	<u>rslope</u>	<u>wasp</u>	<u>welev</u>	<u>area</u>	<u>shape</u>	<u>dens</u>
relev	-0.194							
rslope	-0.254	0.818						
wasp	-0.061	0.065	0.080					
welev	-0.200	0.996	0.824	0.066				
area	0.037	0.039	-0.075	0.053	0.052			
shape	0.027	-0.227	-0.240	0.087	-0.211	0.196		
dens	-0.056	0.096	0.196	-0.087	0.075	-0.812	-0.673	
wslope	-0.241	0.881	0.891	0.077	0.901	0.040	-0.207	0.086
<b>James-Coastal Plain</b>								
	<u>Strahler</u>	<u>relev</u>	<u>rslope</u>	<u>wasp</u>	<u>welev</u>	<u>area</u>	<u>shape</u>	<u>dens</u>
relev	-0.224							
rslope	-0.412	0.748						
wasp	0.026	0.408	0.302					
welev	0.035	0.926	0.646	0.440				
area	0.858	-0.211	-0.390	0.021	0.046			
shape	-0.415	-0.036	0.042	-0.105	-0.199	-0.295		
dens	-0.242	0.198	0.245	0.094	0.117	-0.558	-0.374	
wslope	0.024	0.847	0.669	0.435	0.929	0.032	-0.185	0.127

Table 2.2. Principal component analysis (PCA) eigenvalue, proportion of variance explained, cumulative variance explained, and factor loading for the first 3 principal components calculated for stream reaches in the Holston-Ridge and Valley, Roanoke-Piedmont, and James-Coastal Plain regions in Virginia. Predictors include stream order (Strahler), mean reach elevation (relev), reach slope (rslope), median watershed aspect (wasp), mean watershed elevation (welev), watershed area (area), watershed shape parameter (shape), drainage density (dens), and average watershed slope (wslope). Important axes for each PCA and important predictors for each axis are in bold.

Variable	Holston-Ridge and Valley			Roanoke-Piedmont			James-Coastal Plain		
	PC1	PC2	PC3	PC1	PC2	PC3	PC1	PC2	PC3
Eigenvalue	<b>2.728</b>	<b>2.193</b>	<b>1.480</b>	<b>3.098</b>	<b>1.848</b>	<b>1.534</b>	<b>5.472</b>	<b>1.576</b>	0.881
Proportion	0.303	0.244	0.164	0.344	0.205	0.170	0.608	0.175	0.098
Cumulative	0.303	0.547	0.711	0.344	0.550	0.720	0.608	0.783	0.881
Strahler	0.336	<b>-0.480</b>	-0.033	-0.129	<b>0.611</b>	-0.179	<b>-0.401</b>	-0.019	-0.142
relev	<b>-0.541</b>	-0.178	0.019	<b>0.522</b>	0.004	0.052	0.271	-0.343	<b>0.430</b>
rslope	-0.122	0.225	<b>0.579</b>	0.394	-0.096	0.216	0.247	0.001	<b>0.703</b>
wasp	-0.225	<b>-0.411</b>	-0.065	-0.067	0.303	<b>0.626</b>	-0.347	-0.083	<b>0.409</b>
welev	-0.310	<b>-0.504</b>	0.260	<b>0.526</b>	0.214	-0.073	<b>-0.388</b>	-0.300	0.090
area	<b>0.458</b>	-0.268	0.140	-0.103	<b>0.465</b>	-0.265	<b>-0.390</b>	-0.275	0.038
shape	<b>-0.464</b>	0.125	-0.096	0.044	<b>-0.402</b>	0.150	0.270	<b>-0.516</b>	-0.303
dens	0.070	<b>0.411</b>	0.179	-0.094	0.232	<b>0.652</b>	-0.238	<b>0.598</b>	0.149
wslope	0.051	-0.085	<b>0.729</b>	<b>0.502</b>	0.218	-0.055	<b>-0.389</b>	-0.293	0.099

Table 2.3. Classification performance of linear discriminant functions, generalized linear models, classification trees, and artificial neural networks built for 7 stream fish species in each of 3 drainage-physiographic regions of Virginia [Holston-Ridge and Valley (N=138), Roanoke-Piedmont (N=262), and James-Coastal Plain (N=80). Common and scientific name, prevalence (Prev.), sensitivity (SE), specificity (SP) and correct classification (CC) are given for cross-validation errors. Kappa Z-statistic and associated p-values are provided to indicate whether classification success was greater than random. Kappa statistic was not applicable (n/a) when presence was not predicted.

Common name	Scientific name	Prev	SE	SP	CC	Kappa	P-value
<b>Linear Discriminant Function</b>							
<b>Holston-Ridge and Valley</b>							
Longhead darter	<i>Percina macrocephala</i>	2.2	66.7	82.2	81.9	3.916	<0.001
Margined madtom	<i>Noturus insignis</i>	10.1	85.7	75.0	76.1	5.557	<0.001
Mimic shiner	<i>Notropis volucellus</i>	3.6	0.0	65.4	63.0	1.374	0.085
Northern hog sucker	<i>Hypentelium nigricans</i>	73.9	66.7	75.0	56.5	5.054	<0.001
Logperch	<i>Percina caprodes</i>	6.5	22.2	82.2	78.3	2.107	0.018
Rosyface shiner	<i>Notropis rubellus</i>	25.4	71.4	84.5	81.2	6.897	<0.001
Spotted bass	<i>Micropterus punctulatus</i>	3.6	20.0	69.9	68.1	3.695	<0.001
<b>Roanoke-Piedmont</b>							
Margined madtom	<i>Noturus insignis</i>	64.8	66.7	42.0	51.2	3.458	<0.001
Roanoke hog sucker	<i>Hypentelium roanokense</i>	30.3	47.1	69.8	65.0	4.039	<0.001
Roanoke logperch	<i>Percina rex</i>	4.2	79.3	56.9	65.0	4.231	<0.001
Roanoke darter	<i>Percina roanoka</i>	28.7	33.3	71.6	68.7	1.375	0.085
Smallmouth bass	<i>Micropterus dolomieu</i>	7.3	67.9	84.6	78.7	6.266	<0.001
Spottail shiner	<i>Notropis hudsonius</i>	5.4	0.0	71.1	67.5	1.034	0.051
White shiner	<i>Luxilus albeolus</i>	27.3	50.0	68.0	61.2	2.462	0.007
<b>James-Coastal Plain</b>							
Creek chubsucker	<i>Erimyzon oblongus</i>	37.5	62.1	63.0	62.5	4.248	<0.001
Flier	<i>Centrarchus macropterus</i>	21.3	81.8	74.8	75.1	4.902	<0.001
Golden shiner	<i>Notemigonus crysoleucas</i>	36.3	54.4	58.2	57.1	3.410	<0.001
Margined madtom	<i>Noturus insignis</i>	7.5	72.0	73.7	73.2	7.389	<0.001
Spottail shiner	<i>Notropis hudsonius</i>	35.0	78.9	81.8	81.6	6.749	<0.001
Swamp darter	<i>Etheostoma fusiforme</i>	5.0	35.7	88.3	85.4	5.525	<0.001
Tessellated darter	<i>Etheostoma olmstedi</i>	37.5	66.2	67.4	67.0	5.909	<0.001
<b>Generalized Linear Model</b>							
<b>Holston-Ridge and Valley</b>							
Longhead darter	<i>Percina macrocephala</i>	2.2	100.0	88.1	88.4	4.383	<0.001
Margined madtom	<i>Noturus insignis</i>	10.1	100.0	91.1	92.0	8.392	<0.001
Mimic shiner	<i>Notropis volucellus</i>	3.6	100.0	63.9	65.2	2.885	<0.001
Northern hog sucker	<i>Hypentelium nigricans</i>	73.9	71.6	77.8	73.2	5.170	<0.001
Logperch	<i>Percina caprodes</i>	6.5	100.0	72.9	74.6	4.535	<0.001
Rosyface shiner	<i>Notropis rubellus</i>	25.4	80.0	80.6	80.4	6.501	<0.001
Spotted bass	<i>Micropterus punctulatus</i>	3.6	80.0	76.7	76.8	2.860	<0.001

Table 2.3.-Continued.

Common name	Scientific name	Prev	SE	SP	CC	Kappa	P-value
Generalized Linear Model							
<b>Roanoke-Piedmont</b>							
Margined madtom	<i>Noturus insignis</i>	64.8	61.5	68.5	64.0	4.634	<0.001
Roanoke hog sucker	<i>Hypentelium roanokense</i>	30.3	62.0	60.4	60.9	3.340	<0.001
Roanoke logperch	<i>Percina rex</i>	4.2	90.9	73.6	74.3	4.609	<0.001
Roanoke darter	<i>Percina roanoka</i>	28.7	74.7	75.3	75.1	7.482	<0.001
Smallmouth bass	<i>Micropterus dolomieu</i>	7.3	89.5	81.4	82.0	6.990	<0.001
Spottail shiner	<i>Notropis hudsonius</i>	5.4	92.9	78.9	79.7	1.351	0.088
White shiner	<i>Luxilus albeolus</i>	27.2	70.4	69.5	69.7	5.824	<0.001
<b>James-Coastal Plain</b>							
Creek chubsucker	<i>Erimyzon oblongus</i>	37.5	83.3	74.0	77.5	4.971	<0.001
Flier	<i>Centrarchus macropterus</i>	21.3	94.1	76.2	80.0	5.280	<0.001
Golden shiner	<i>Notemigonus crysoleucas</i>	36.3	80.0	89.6	66.7	4.857	<0.001
Margined madtom	<i>Noturus insignis</i>	7.5	16.7	98.6	92.5	2.311	0.010
Spottail shiner	<i>Notropis hudsonius</i>	35.0	85.7	84.6	85.0	6.124	<0.001
Swamp darter	<i>Etheostoma fusiforme</i>	5.0	75.0	77.6	77.5	2.370	<0.001
Tessellated darter	<i>Etheostoma olmstedi</i>	37.5	63.3	68.0	66.3	2.735	<0.001
Classification Tree							
<b>Holston-Ridge and Valley</b>							
Longhead darter	<i>Percina macrocephala</i>	2.2	0.0	97.8	97.8	0.000	n/a
Margined madtom	<i>Noturus insignis</i>	10.1	80.0	95.3	94.2	0.636	<0.001
Mimic shiner	<i>Notropis volucellus</i>	3.6	80.0	95.3	94.2	0.636	<0.001
Northern hog sucker	<i>Hypentelium nigricans</i>	73.9	81.0	76.5	80.4	0.417	<0.001
Logperch	<i>Percina caprodes</i>	6.5	0.0	93.5	93.5	0.000	n/a
Rosyface shiner	<i>Notropis rubellus</i>	25.4	81.8	85.3	84.8	0.666	<0.001
Spotted bass	<i>Micropterus punctulatus</i>	3.6	0.0	96.4	96.4	0.000	n/a
<b>Roanoke-Piedmont</b>							
Margined madtom	<i>Noturus insignis</i>	64.8	70.5	57.4	67.8	0.322	<0.001
Roanoke hog sucker	<i>Hypentelium roanokense</i>	30.3	0.0	95.8	95.8	0.000	n/a
Roanoke logperch	<i>Percina rex</i>	4.2	0.0	69.7	69.7	0.000	n/a
Roanoke darter	<i>Percina roanoka</i>	28.7	60.3	84.7	77.4	0.585	<0.001
Smallmouth bass	<i>Micropterus dolomieu</i>	7.3	0.0	92.7	92.7	0.000	n/a
Spottail shiner	<i>Notropis hudsonius</i>	5.4	0.0	94.6	94.6	0.000	n/a
White shiner	<i>Luxilus albeolus</i>	27.2	62.8	79.8	77.0	0.422	<0.001
<b>James-Coastal Plain</b>							
Creek chubsucker	<i>Erimyzon oblongus</i>	37.5	67.6	84.8	77.5	0.920	<0.001
Flier	<i>Centrarchus macropterus</i>	21.3	0.0	78.8	78.8	0.000	n/a
Golden shiner	<i>Notemigonus crysoleucas</i>	36.3	0.0	63.8	63.8	0.000	n/a
Margined madtom	<i>Noturus insignis</i>	7.5	0.0	92.5	92.5	0.000	n/a
Spottail shiner	<i>Notropis hudsonius</i>	35.0	63.6	85.1	76.3	0.719	0.001
Swamp darter	<i>Etheostoma fusiforme</i>	5.0	0.0	95.0	95.0	0.000	n/a
Tessellated darter	<i>Etheostoma olmstedi</i>	37.5	81.8	69.6	71.3	0.298	<0.001

Table 2.3. -Continued.

Common name	Scientific name	Prev	SE	SP	CC	Kappa	P-value
Artificial Neural Network							
<b>Holston-Ridge and Valley</b>							
Longhead darter	<i>Percina macrocephala</i>	2.2	100.0	98.5	98.6	9.032	<0.001
Margined madtom	<i>Noturus insignis</i>	10.1	92.8	99.2	98.6	15.414	<0.001
Mimic shiner	<i>Notropis volucellus</i>	3.6	100.0	98.5	98.6	9.853	<0.001
Northern hog sucker	<i>Hypentelium nigricans</i>	73.9	87.2	87.3	87.2	90.600	<0.001
Logperch	<i>Percina caprodes</i>	6.5	66.7	99.2	97.1	9.622	<0.001
Rosyface shiner	<i>Notropis rubellus</i>	25.4	100.0	95.1	96.4	10.718	<0.001
Spotted bass	<i>Micropterus punctulatus</i>	3.6	100.0	97.7	97.8	9.182	<0.001
<b>Roanoke-Piedmont</b>							
Margined madtom	<i>Noturus insignis</i>	64.8	94.7	100.0	96.6	15.003	<0.001
Roanoke hog sucker	<i>Hypentelium roanokense</i>	30.3	100.0	91.2	93.6	14.548	<0.001
Roanoke logperch	<i>Percina rex</i>	4.2	100.0	98.4	98.5	13.724	<0.001
Roanoke darter	<i>Percina roanoka</i>	28.7	98.7	97.3	97.7	15.273	<0.001
Smallmouth bass	<i>Micropterus dolomieu</i>	7.3	94.7	100.0	99.6	14.562	<0.001
Spottail shiner	<i>Notropis hudsonius</i>	5.4	100.0	98.4	98.5	14.132	<0.001
White shiner	<i>Luxilus albeolus</i>	27.2	95.8	93.7	94.2	13.950	<0.001
<b>James-Coastal Plain</b>							
Creek chubsucker	<i>Erimyzon oblongus</i>	37.5	100.0	96.0	97.5	8.844	<0.001
Flier	<i>Centrarchus macropterus</i>	21.3	100.0	95.2	96.2	8.749	<0.001
Golden shiner	<i>Notemigonus crysoleucas</i>	36.3	96.6	98.0	97.5	8.479	<0.001
Margined madtom	<i>Noturus insignis</i>	7.5	100.0	100.0	100.0	12.649	<0.001
Spottail shiner	<i>Notropis hudsonius</i>	35.0	100.0	94.2	96.3	8.252	<0.001
Swamp darter	<i>Etheostoma fusiforme</i>	5.0	100.0	100.0	100.0	8.944	<0.001
Tessellated darter	<i>Etheostoma olmstedi</i>	37.5	93.3	100.0	97.5	8.467	<0.001



Table 2.4. Slope estimate ( $b$ ), F-statistic, p-value, and  $R^2$  for linear regressions of the effect of prevalence on prediction sensitivity (SE), specificity (SP), and correct classification (CC) cross-validation error rates for discriminant function analysis (DFA), generalized linear model (GLM), classification tree (TRE) and artificial neural network (ANN) for 7 fish species in each of 3 regions in Virginia.

Method	SE				SP				CC			
	$b$	F	p	$R^2$	$b$	F	p	$R^2$	$b$	F	p	$R^2$
<b>Holston-Ridge and Valley</b>												
DFA	0.49	0.97	0.370	0.16	0.03	0.06	0.818	0.01	-0.20	2.16	0.202	0.30
GLM	-0.35	6.59	0.050	0.57	0.01	0.00	0.960	0.00	-0.07	0.19	0.681	0.04
TRE	0.80	1.65	0.256	0.25	-0.28	83.42	<0.001	0.94	-0.23	31.50	0.003	0.86
ANN	-0.08	0.15	0.714	0.03	-0.16	94.46	<0.001	0.95	-0.15	98.87	<0.001	0.95
<b>Roanoke-Piedmont</b>												
DFA	0.21	0.14	0.719	0.03	-0.40	3.45	0.123	0.41	-0.29	6.53	0.051	0.57
GLM	-0.54	13.63	0.014	0.73	-0.17	1.79	0.238	0.26	-0.25	4.34	0.092	0.46
TRE	1.32	10.81	0.021	0.68	-0.40	2.78	0.156	0.36	-0.28	1.73	0.246	0.26
ANN	-0.05	1.36	0.296	0.21	-0.01	0.00	0.950	0.00	-0.05	1.26	0.313	0.20
<b>James-Coastal Plain</b>												
DFA	0.23	0.20	0.674	0.04	-0.56	5.49	0.066	0.52	-0.53	4.99	0.076	0.50
GLM	0.74	0.87	0.393	0.15	-0.24	0.53	0.498	0.10	-0.40	2.46	0.177	0.33
TRE	1.98	4.69	0.083	0.48	-0.60	5.04	0.075	0.50	-0.72	17.44	0.009	0.78
ANN	-0.10	1.83	0.234	0.27	-0.06	0.66	0.454	0.12	-0.07	3.02	0.143	0.38
<b>All Regions</b>												
DFA	4.16	2.28	0.147	0.11	-0.22	3.93	0.062	0.17	-0.28	10.92	0.004	0.36
GLM	-0.29	2.08	0.165	0.10	-0.10	1.00	0.330	0.05	-0.19	4.44	0.049	0.19
TRE	1.03	8.51	0.009	0.31	-0.40	16.29	<0.001	0.46	-0.35	13.06	0.002	0.41
ANN	-0.04	0.25	0.624	0.01	-0.09	3.93	0.062	0.17	-0.10	22.07	<0.001	0.54

Table 2.5. A. Mean and standard deviation (in parentheses) cross-validation sensitivity (SE), specificity (SP) and correct classification (CC) of discriminant function (DFA), generalized linear model (GLM), classification tree (TRE), and artificial neural network models that predict the occurrence of 7 fish species in each of 3 regions in Virginia. B. Bootstrap-adjusted ANOVA p-values describing significance of differences in SE, SP, and CC among DFA, GLM, TRE, and ANN in each of 3 regions in Virginia.

Holston-Ridge and Valley					Roanoke-Piedmont					James-Coastal Plain				
<b>A.</b>														
	DFA	GLM	TRE	ANN		DFA	GLM	TRE	ANN		DFA	GLM	TRE	ANN
SE	47.53 (32.71)	90.23 (12.50)	46.11 (43.14)	92.39 (12.37)	SE	49.19 (26.58)	77.41 (13.63)	27.66 (34.63)	97.70 (2.53)	SE	64.44 (15.83)	71.16 (25.84)	30.43 (38.35)	98.56 (2.64)
SP	76.31 (7.07)	78.73 (9.16)	91.44 (7.75)	96.50 (4.29)	SP	66.29 (13.42)	72.51 (7.08)	82.10 (14.33)	97.00 (3.33)	SP	72.46 (10.52)	81.23 (10.40)	81.37 (11.47)	98.89 (2.20)
CC	72.16 (9.79)	78.66 (9.19)	91.61 (6.46)	96.33 (4.11)	CC	65.33 (8.28)	72.24 (7.81)	82.14 (11.99)	96.96 (2.28)	CC	71.70 (10.16)	77.93 (9.38)	79.31 (11.09)	98.76 (1.60)
<b>B.</b>														
	DFA	GLM	TRE		DFA	GLM	TRE		DFA	GLM	TRE			
GLM	SE	0.9910			GLM	SE	0.9924			GLM	SE	0.9997		
	SP	0.0005				SP	0.0017				SP	0.0530		
	CC	0.0003				CC	<0.0001				CC	0.0143		
TRE	SE	0.0124	0.9909		TRE	SE	<0.0001	0.0462		TRE	SE	0.0015	0.0107	
	SP	0.9999	1.0000			SP	0.8731	0.9901			SP	0.1522	0.8837	
	CC	1.0000	0.9636			CC	0.7017	0.9841			CC	0.1072	0.4967	
ANN	SE	0.9151	0.0338	0.4491	ANN	SE	0.9994	0.0144	0.0121	ANN	SE	0.4816	0.3619	0.0173
	SP	0.5446	0.7056	0.0142		SP	0.8035	1.0000	0.3157		SP	0.9127	0.9538	0.9977
	CC	0.1102	0.9672	0.0068		CC	0.4574	0.9999	0.0849		CC	0.9956	0.9979	0.9980

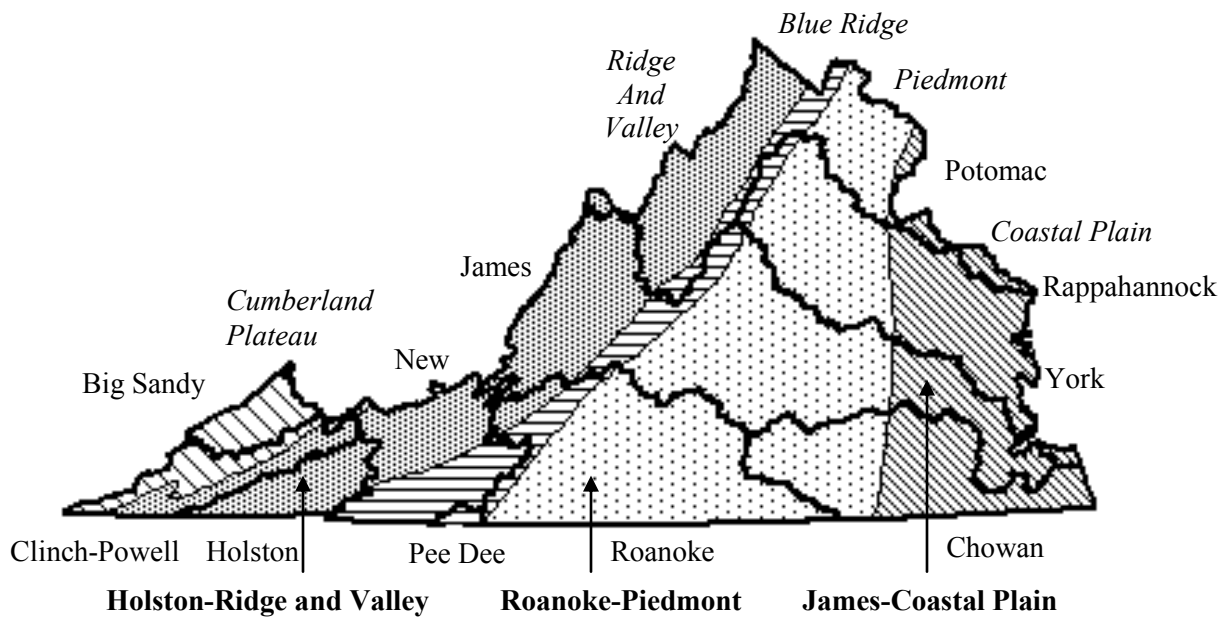


Figure 2.1. Major drainages (i.e., USGS 6-digit hydrologic units; bold outline) and their intersection with 5 physiographic provinces (italics) in Virginia. Study regions are labeled in bold. The York and Rappahannock are considered separately to maintain consistency with Jenkins and Burkhead (1993).

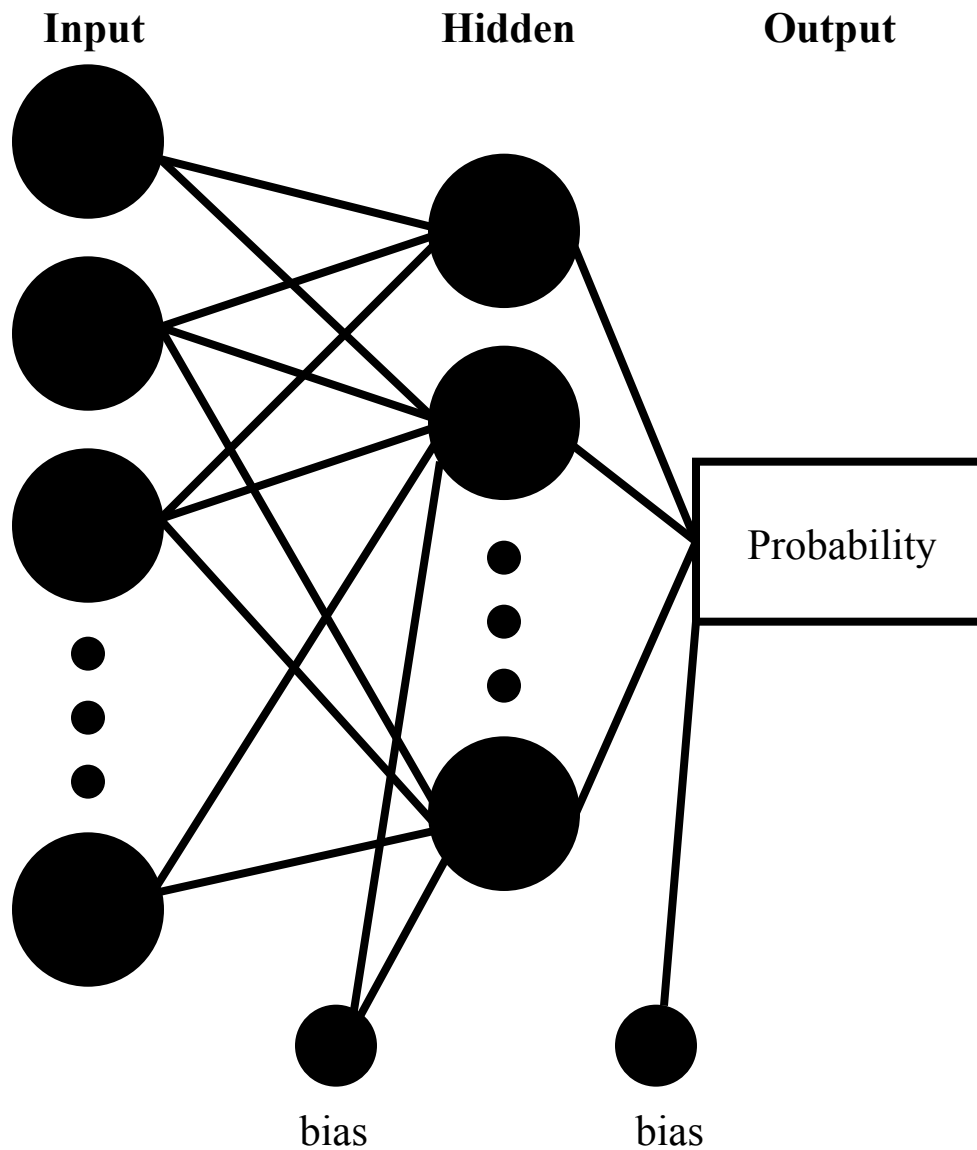


Figure 2.2. A diagrammatic model of a one-hidden-layer, feed-forward artificial neural network design (following Olden and Jackson 2001). Predictor variables are depicted as large filled disks and are listed in Table 2.2. Number of connections and connection weights varies among models. Number of hidden nodes is determined empirically. Bias nodes act as threshold functions for hidden nodes. Output is the predicted probability of occurrence.

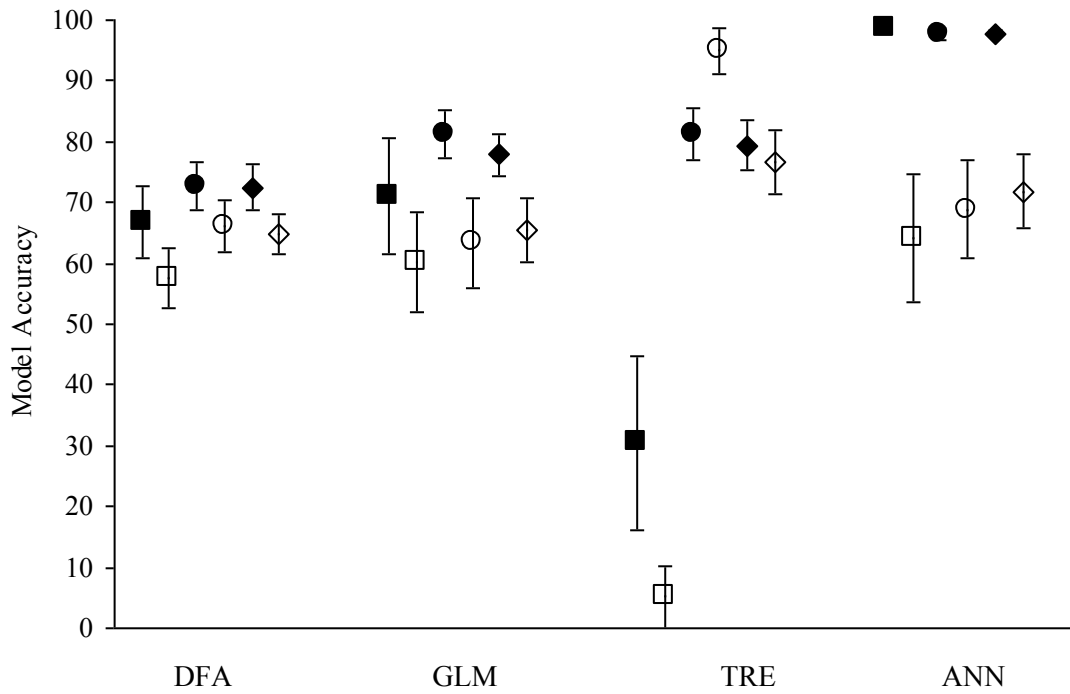


Figure 2.3. Mean model sensitivity (squares), specificity (circles), and total correct prediction (diamonds) of complex models (filled) built with 9 predictors and parsimonious models (unfilled) built with only statistically significant predictors based on leave-one-out cross-validation for 7 fish species in the James-Coastal Plain region, Virginia. Model types include linear discriminant function (DFA), generalized linear model (GLM), classification tree (TRE), and artificial neural network (ANN). Bars are one standard deviation.

### **Chapter 3. Conservation implications of relations among the accuracy, generality and resolution of predictions of occurrence and species traits for fishes in the upper Tennessee River drainage.**

#### INTRODUCTION

Limited resources for conservation require efficient conservation – directing actions to where they are most needed and most likely to succeed. Knowledge of species occurrence (presence and absence), including their spatial patterns, at spatial extents and grains that are ecologically meaningful and relevant to conservation is a prerequisite for efficient conservation. Unfortunately, insufficient knowledge of species occurrence is a pervasive constraint on efficient conservation of stream fishes. We know little about the spatial distribution of most stream fishes, especially at the large spatial extents and fine grains at which many conservation decisions are made. In particular, we have few descriptions of the patterns of discontinuity of occurrence of most stream fishes. To facilitate conservation, models are often used to predict species occurrence in lieu of empirical observations (Scott et al. 1993, 2002, Sowa 1998, McCleary and Hassan 2008). Predictions of species occurrence are mapped and conservation efforts are directed by assuming that the predictions of occurrence have sufficient accuracy and resolution to improve conservation decisions. However, predictions of species occurrence are often incorrect and may have insufficient resolution to direct some conservation decisions. Prediction accuracy and resolution differ among species, possibly due to differences in their ecological traits, but little information describing the likely accuracy and resolution of predictions is available. The ability to anticipate prediction error rates and the size of patches of predicted presence will facilitate the appropriate application of model predictions to conservation decisions.

Accurate high-resolution descriptions of occurrence that facilitate the conservation of stream fishes are needed. Anthropogenic degradation of aquatic ecosystems has contributed to pervasive imperilment and extinction of stream fish (Williams et al. 1989, Etnier 1997). At least 19% of the approximately 490 freshwater fish species that are native to the southeastern United States are imperiled (Etnier 1997). Limited resources and frequent imperilment make their efficient conservation imperative. This work was initiated to provide information that supports the appropriate application of predictions of species occurrence in aquatic gap analyses.

Models that predict fish occurrence with high accuracy and resolution are necessary to direct conservation efforts and provide meaningful descriptions of prediction characteristics to identify their relations with species traits. Predicting the occurrence of stream fishes is a challenging conservation imperative. The physical mechanisms that define lotic system structure

and generate patterns of habitat and species occurrence are vectors of water, sediment, energy and nutrients that form stream networks (Schumm 1977, Vannote et al. 1980, Leopold 1994). These vectors produce spatial and temporal autocorrelation of environmental attributes and species occurrence along a continuum of environmental gradients within the stream network (Vannote et al. 1980). While breaks in the continuum (a component of structure) are identifiable as part of a spatial hierarchy (Frissell et al. 1988), species do not necessarily perceive these discontinuities as boundaries (*sensu* Holling 1992) that define suitable habitat (Campbell et al. 2007). However, ecologists' perceptions of how a population or assemblage is structured are influenced by the spatial and temporal scales at which samples are collected and models are built (Allen et al. 1984, Rahel et al. 1984). The apparent structure of stream fish populations typically reflects the fact that many of the habitat units that species respond to are spatially discrete (e.g., Frissell et al. 1988, Pringle et al. 1988, Hawkins et al. 1993). Because the dearth of knowledge of the life history and traits of most stream fishes precludes the precise description of their habitat, the delineation of suitable habitat may be best accomplished with less subjective methods that use empirical data. Empirical methods that delineate suitable habitat might describe ecologically meaningful characteristics of species occurrence that have conservation implications. For example, the predicted mean length and total length of patches of predicted species presence can be used to select the number, locations, and sizes of monitoring or conservation sites.

Complex models may be necessary to predict the occurrence of stream fish with sufficient accuracy and resolution to facilitate conservation decisions. Empirical relations between habitat descriptions and species occurrence are complex (Jackson et al. 2001), and may include measuring several environmental attributes at several spatiotemporal scales (Holling 1992, Karl et al. 2000). Fish ecologists often resort to calculating attributes at one or more scales that are functionally related to climate, geology, elevation, and/or landscape gradient that describe spatial position and using them to predict species occurrence (e.g., Gorman and Karr 1978, Kruse et al. 1997, Porter et al. 2000, Filipe et al. 2002). Poor model performance encourages the use of more complex models (as measured by the number of predictors incorporated) or alternative methods, but comparisons of different methods often produce equivocal results (Manel et al. 1999a, 1999b, Olden and Jackson 2002, Chapter 2). Models that predict species occurrence are often built with great concern for statistical validity and the avoidance of overfitting (Burnham and Anderson 2002). The naïve use of predictor selection methods (Olden and Jackson 2000) or model selection methods (Burnham and Anderson 2002) can provide overly optimistic assessments of prediction performance. Methods for building models and assessing their accuracy are well developed (Fielding and Bell 1997, Guisan and Zimmerman 2000, Scott et al.

2002), but few rigorous assessments of the effect of model complexity on prediction performance (including accuracy, generality, and resolution) are available to guide model building and application, especially for stream fishes. Prediction resolution, the ability of models to distinguish among neighboring prediction units, is infrequently assessed, but important when conservation and management decisions are made at fine grains. Models can accurately predict probabilities of species occurrence without sufficient resolution to distinguish among most prediction units. This is especially problematic for very rare and common species, for which prediction accuracy can be high for uninformative models that predicting similar probabilities of occurrence for most prediction units.

Species traits (e.g., maximum length) are expected to affect species' patterns of occurrence (Holling 1992, Poff 1997, Goldstein and Meador 2004) and vulnerability to anthropogenic extinction (e.g., Angermeier 1995). Thus, if models predict occurrence with sufficient accuracy and resolution, the accuracy and resolution of model predictions are expected to vary among species due to differences in their traits (Boone and Krohn 1999, Manel et al. 2001, Seoane et al. 2005). If relations between prediction characteristics and species traits can be described, model prediction characteristics might be anticipated and used to better guide conservation decisions (McPherson et al. 2006). For example, if species with similar traits have similar predicted patch sizes and observed frequency of absence from patches of suitable habitat, that information can be used to direct monitoring and to set habitat conservation goals. My goal is to describe characteristics of predictions of fish occurrence, and relations between those characteristics and species traits to facilitate the appropriate application of species occurrence predictions to conservation.

At large spatial extents and fine spatial grains, patterns of stream fish occurrence are poorly described and discontinuities in suitable habitat are difficult to discern. Therefore, I predict fish occurrence among stream reaches using statistical models. I used stream reaches as prediction units because they are hydrologically meaningful, easily identified, and temporally stable spatial units. Describing relations between characteristics of predicted fish occurrence and species traits requires accurate, high-resolution predictions of occurrence. Although fish occurrence is usually adequately described with few parameters (Hubert and Chamberlain 1996, Angermeier and Winston 1999), increasing model complexity (the number of predictors used) might improve prediction accuracy (Peterson and Cohoon 1999, Porter et al. 2000) and resolution, but might also reduce model generality. I assess the effects of model complexity on prediction accuracy, generality, and resolution by comparing the predictions of models built with 1 to 6 predictors. I then use the most accurate, parsimonious models to predict fish occurrence



and to describe the characteristics of those predictions. I expect prediction accuracy and resolution to increase and generality to decrease with increasingly complex models. I also expect the effect of increasing model complexity on prediction accuracy, generality, and resolution to be greater for common species than for rare species, because more presence data are available for common species. Predicted species occurrence (i.e., patches of predicted suitable habitat) is expected to be discontinuous for all species, and I expect mean patch lengths and total patch lengths to be negatively related to model complexity.

Characteristics of predicted fish occurrence might vary among species due to differences in their prevalence (Manel et al. 2001) and ecological traits (Boone and Krohn 1999, Seoane et al. 2005). I describe relations of species prevalence, maximum recorded body length, and nesting behavior with predicted mean patch length (MPL; mean length of contiguous reaches in which a species is predicted present) and total patch length (TPL; the sum of the lengths of all reaches in which the species is predicted present) for species in the upper Tennessee River basin (UTRB) for which accurate prediction of occurrence is possible and for which species trait information is available. Because species traits are constrained by phylogeny, I describe relations of species traits with predicted patterns of occurrence among and within taxonomic families.

Species prevalence is often positively related to the extent of geographic distribution (Brown 1995) and model prediction accuracy (Manel et al. 2001). Gaston and Blackburn (1996) describe several mechanisms that have been postulated to explain the relationship. These mechanisms can be used to predict relations among characteristics of prediction accuracy and species traits. The positive relation between prevalence and prediction accuracy may be due to ecological mechanisms or a statistical artifact of greater predicted MPL and TPL for common species. I predict that species prevalence is positively related to prediction accuracy, MPL and TPL. Pyron (1999) found that large-bodied suckers and sunfishes had larger geographic ranges. Fishes that grow larger are expected to have larger MPL and TPL because they might require larger patches of habitat, have greater breadth of suitable habitat, and have larger geographic ranges. Alternatively, larger size might be positively related to MPL and TPL because larger fish are more likely to be sampled if they are present at a location (Reynolds 1996), reducing sampling omission errors that bias predictions of species occurrence (McKenzie 2005, Royle et al. 2005). I predict that species that attain larger sizes will have larger MPL and TPL. Ensign et al. (1997) found that fishes that invested more in care of offspring re-colonized a disturbed reach more quickly than those that invested less care. Species that re-colonize quickly might occupy a greater proportion of suitable habitat and/or have fewer incidence omission errors (i.e., absence in suitable habitat) that bias predictions of occurrence than species that re-colonize slowly. I predict

that investment in care of offspring (e.g., nest building versus egg scattering) is positively related to prediction accuracy, MPL, and TPL.

Regardless of whether the mechanisms affecting predicted occurrence characteristics are ecological effects or statistical artifacts, describing occurrence characteristics and their relations to species traits will facilitate better application of predictions of species occurrence to conservation. For example, if the lengths of patches of predicted occurrence vary with species length, the length of conservation and monitoring units might be adjusted. I use the conclusions of the previous analyses to guide the building, assessment and application of models that predict fish occurrence in the UTRB. I compare the characteristics of these predictions with results from the peer reviewed literature and models built using habitat descriptions in reference books (e.g., Jenkins and Burkhead 1993) to assess the relative usefulness of my models for conservation.

## METHODS

### *Study Area*

Exceptional fish species richness, abundance of streams, and frequency and severity of threats to species persistence make predicting the occurrence of fishes in the upper Tennessee River basin a management and conservation priority. The UTRB is a 6<sup>th</sup> – order (Strahler 1952) watershed of approximately 4,474,000 ha that includes portions of Virginia, Tennessee, North Carolina, and Georgia (Figure 3.1). The UTRB also includes portions of the Ridge and Valley, Blue Ridge, and Appalachian Plateau physiographic provinces (Jenkins and Burkhead 1993, Fenneman and Johnson 1946) and comprises 5 subbasins: the Holston, Clinch, French Broad, Hiwassee, and Little Tennessee rivers. High environmental heterogeneity, abundance of stream habitat and absence of glaciations contribute to the high fish species richness of the UTRB. Although forests remain the dominant land cover, the UTRB is substantially altered by human activities. Agriculture and urbanization are important land uses and dams fragment many of the UTRB rivers and streams.

### *Predicting Fish Occurrence*

#### Fish Occurrence Data

Species occurrence records (samples) were procured from public databases (e.g., the Virginia Fish and Wildlife Information Service <http://128.173.240.58/build02/fwis/default.asp>) from the states of Virginia, Tennessee, and North Carolina. Most occurrence records were from collection permit and scientific reports. Errors of sampling omission (i.e., false absence) are likely and may be due to failure to sample the appropriate spatial or temporal strata, inadequate sampling effort, or ineffective or inappropriate methods (Angermeier et al. 2002). Additionally,

the objectives of the samplers may not have included capturing all of the species present. Errors of sampling commission (i.e., false presence) are less likely, but may occur due to species misidentification or errors in data transcription. Despite errors, these databases contain the most comprehensive information available describing fish distributions in the UTRB.

Predicting species occurrence using samples that are not representative of the region's assemblage biases results. I minimized the effect of sampling errors on predictions of occurrence by excluding samples with many likely sampling omission errors. This was accomplished by using only samples that appeared relatively complete (hereafter assemblage samples). Samples that recorded > 2 species, including  $\geq 1$  non-game species, and had reliable location information (i.e., could be plotted within 50 m of a stream reach) were defined as "assemblage samples" and assumed to have few omission errors. When multiple samples were available for a reach, the lists of species were combined and used as a single sample for that reach.

#### Fish Occurrence Predictors

I used the United States Geological Survey (USGS) National Hydrography Dataset (NHD, USGS 2000 <http://nhd.usgs.gov>) 1:100,000-scale stream-reach files to define streams. I identified reaches with the unique codes assigned in the NHD. The NHD defines reaches based on the location of confluences, major changes in stream gradient and changes in channel type (e.g., natural or channelized). Data from the National Elevation Dataset (NED, USGS 2000 <http://gisdata.usgs.gov/ned/default.asp>) were used to delineate the contributing watershed for each reach, and to calculate elevation and slope variables. Delineating watersheds for each reach described the contributing watershed of each reach that allowed me to avoid the need to define watersheds as prediction units (e.g., Dunham et al. 1999, 2002, Wall et al. 2004) and allowed the predictions of species occurrence to cross hydrologic unit boundaries. That is, this approach allows the data to define the predicted patches of occurrence without being constrained within pre-defined boundaries.

I calculated 9 environmental attributes to use as predictors of stream fish occurrence. State and physiographic province describe the geographic location of reaches and may account for differences in zoogeography. Strahler (1952) stream order and reach length index reach size. Downstream link magnitude (Osborne and Wiley 1992, Smith and Kraft 2005) measures proximity to larger streams, which is often correlated with fish species richness (Gorman 1986). Mean reach elevation and gradient are used because they are often correlates of fish species occurrence (Rahel and Hubert 1991, Reyes-Gavilan et al. 1996, Kruse et al. 1997). Reach length and sinuosity (Gordon et al. 1992) describe reach morphology and relative location of reaches in

the stream landscape (Fausche et al. 2002, Tockner et al. 2002). One predictor (F-type) describes human modification of reaches as natural or altered by humans.

I restricted my analyses to temporally stable predictors to accommodate the temporal extent of species occurrence sampling. Using predictors that have changed values subsequent to fish collections would confound predictions of fish occurrence. Predictors were assessed for outliers and collinearity by visual assessment of plots of data and Pearson correlations, respectively. Because remotely sensed predictors were used, data for the entire UTRB, including sites with and without fish samples were assessed. These analyses were used to identify outliers so they could be corrected, to assess the likely usefulness of predictors, and to interpret the results of model selection.

#### *Model Complexity and Selection*

To identify the appropriately complex models to use in subsequent analyses I assessed the effects of model complexity on prediction accuracy, generality, and resolution using a subset of the UTRB occurrence data. The Clinch River subbasin and Holston River subbasin provide an excellent opportunity to assess the effect of model complexity on model generality because both subbasins have many assemblage samples, share many species, and have similar environmental conditions because both occur mostly in the Ridge and Valley physiographic province. Models were built for 59 species present in 133 samples from the Clinch subbasin. Multiple logistic regression (MLR) is used to predict species occurrence because it is commonly used, relatively well developed and understood, and allows the use of both categorical and continuous predictor variables (Hosmer and Lemeshow 2000).

Manly et al. (2002), Burnham and Anderson (2002), and Hosmer and Lemeshow (2000) recommend careful selection of predictor variables prior to model building, assessment, and selection. However, little is known about the ecological traits and habitat preferences of most aquatic species and therefore the identity of the most useful predictor variables. Further, most available predictors of occurrence are indirect, distal gradients (Austin 2002) that preclude a priori model specification. Therefore, I assessed prediction accuracy for all possible combinations of predictor variables, excluding state, physiographic province, and F-type. Interaction terms were not considered due to the lack of information promoting their use and the desire for model parsimony. Transformation of continuous predictor variables was not necessary for model building since generalized linear models, like logistic regression, are not constrained by assumptions of the distribution of errors (McCullagh and Nelder 1989). However, categorical variables were converted to numeric values. For example, physiographic provinces were coded as Ridge and Valley = 1, Blue Ridge = 2 and Appalachian Plateau = 3.

I compared the prediction accuracy, generality, and resolution of logistic regression models that used from 1 to 6 predictors of occurrence. Akaike's Information Criteria (AIC; Akaike 1974) and the area under the receiver-operator characteristic curve (AUC) were used to rank candidate models and select a best model for each species at each level of complexity. AIC weights, the difference between the smallest AIC score and each candidate model score divided by the sum of the differences (Burnham and Anderson 2002), were used to select a best model(s) at each level of complexity. Burnham and Anderson (2002) advise caution in using the best model when its relative weight is less than 0.9. When AIC weights did not clearly identify a superior model (i.e., AIC weight < 0.9) for the candidate models for each species at each level of complexity, models were ranked by AIC and the "best" model was selected as the one with the largest AUC among the 10% with the smallest AIC. AUC measures a model's ability to discriminate between correct presence and absence predictions across the range of possible critical values. Criteria for assessing the usefulness of a model via AUC remain poorly developed, but AUC allows comparison of models without selection of critical values. Comparison of all possible models is not recommended as a model selection method (Burnham and Anderson 2002, Olden et al. 2004), but it can be a useful approach for assessing differences among models when meaningful a priori model selection is not possible due to insufficient knowledge of species occurrence-habitat relationships (Fleishman et al. 2001). Comparison of all possible models ensured that no alternative model predicted occurrence more accurately than the model selected for use in subsequent analyses.

#### *Model Complexity and Occurrence Characteristics*

The effect of model complexity on prediction accuracy was assessed by comparing mean cross-validation sensitivity (SE; percent of correct presence predictions), specificity (SP; percent of correct absence predictions), correct classification (CC; percent of correct occurrence predictions), Kappa statistic (Cohen 1960, Guggenmoos-Holzmann 1996), and AUC values for the best models for each of the 59 species at each level of complexity. The effect of model complexity on prediction generality (transferability) was assessed by comparing models built with 133 samples from the Clinch subbasin to predict the occurrence of species in 154 samples from the Holston subbasin using the same statistics. All species were present in > 4 samples from the Clinch and Holston subbasins. The effect of model complexity on prediction resolution was assessed by comparing mean patch length (MPL) and total patch length (TPL) of the best model for each species at each complexity. The effect of model complexity on accuracy, generality, and resolution were assessed for rare species (present in  $\leq 10\%$  of samples from the Clinch subbasin) and common species (present in  $\geq 40\%$  of samples from the Clinch subbasin). Presence and

absence predictions were assigned using species prevalence as the critical value (Hosmer and Lemeshow 2000, Liu et al. 2005).

#### *Relations Between Prediction Characteristics and Species Traits*

The characteristics of predictions of species occurrence are expected to differ among species due to differences in their prevalence (Manel et al. 2001) and ecological traits (Boone and Krohn 1999, Seoane et al. 2005). I assess the relation between species prevalence and prediction accuracy, generality, and resolution of the best model for each species using linear regression. I use cross-validation AUC to characterize accuracy and AUC of models built with 133 samples from the Clinch subbasin to predict the occurrence of species in 154 samples from the Holston subbasin to characterize generality, and MPL and TPL to characterize resolution. These analyses are conducted for all species for which significant (cross-validation likelihood ratio and Kappa p-values < 0.05) models could be built, because including insignificant models would lead to spurious results.

I assess the relations between species prevalence, maximum recorded body length, and nesting behavior and MPL and TPL using Kendall's  $\tau_c$  coefficient of rank correlation. I use Kendall's coefficient because it is a non-parametric statistic that is easily interpreted (Sokal and Rohlf 1995). Specifically, Kendall's  $\tau_c$  coefficient is the difference between the probabilities that the observed data are in the same order and in different orders for two variables, adjusted for ties. Species prevalence was calculated for the Clinch subbasin. Numerical codes assigned to nesting behaviors are other = 1, egg scatterer = 2, cavity or crevice spawner = 3, and manipulator = 4 (following Angermeier 1995). These analyses are conducted for all species for which significant (cross-validation likelihood ratio and Kappa p-values < 0.05) models could be built, and for which trait information were available. The results of these analyses guided the building, assessment, and application of models to predict the occurrence of stream fishes throughout the UTRB.

#### *System-wide Predictions and Assessment*

Standard criteria for assessing the sufficiency of model prediction accuracy and resolution for conservation applications have not yet been developed. While developing such criteria is beyond the scope of my work, comparison of the accuracy and resolution of my models with alternatives derived from habitat descriptions in reference books and others reported in peer-reviewed literature will facilitate assessment of their relative value and provide examples of the results that are likely in other studies. I built MLR models that predict the occurrence of stream fishes with > 4 presence records in the UTRB. Models were built using all possible combinations of 1 to 9 predictors. The best model for each species was selected based on the significance of

the model fit, AIC, and AUC. Significance of model fit was assessed using the likelihood ratio statistic ( $p < 0.05$ ). Among significant models, the model with the smallest AIC was selected as the best model. If AIC weights differed by  $< 0.9$ , the model with the largest cross-validation AUC was selected from the 10% of the models with the smallest AIC values as the best model. If no model for a species had a significant likelihood ratio test, the model with the smallest AIC and largest cross-validation AUC was selected as the best model. Cross-validation SE, SP, CC, Kappa, and AUC were used to assess model prediction accuracy and MPL and TPL were used to assess model resolution. I calculated several accuracy statistics to allow comparison of the accuracy of my models with “reference book models” for the same species and other model predictions described in peer-reviewed literature.

“Reference book models” were built by converting habitat preference descriptions in reference books to remotely sensed predictors and using them to predict fish occurrence in the UTRB. Reference book models were built for all species that had the necessary habitat preference descriptions. Habitat preference descriptions were obtained from Jenkins and Burkhead (1993) and Etnier and Starnes (1993). Predictors included stream size converted to a range of Strahler (1952) orders, and the subbasins and physiographic provinces with known presence. Applying reference book models does not require critical values, so AUC is not an appropriate assessment statistic. The prediction accuracy of these models was assessed using the SE, SP, CC, and Kappa statistics for occurrence data for the UTRB. The difference between the mean accuracy of MLR models and reference book models for species modeled by both methods was compared using paired t-tests. Prediction accuracy statistics were recorded from 14 peer-reviewed publications and compared with the accuracy of my prediction. Values of SE, SP, CC, Kappa, and AUC were recorded by species when possible. Means of accuracy statistics were used when statistics for individual species were not provided. Differences in the mean accuracy of MLR models, reference book models and those found in peer-reviewed publications were compared using graphical analyses. All statistical analyses were performed using SAS 9.0 and 9.1 (SAS Institute Inc., Cary, NC). All GIS analyses were performed using ARC/INFO 8.0.2 and ARC GIS 9.1 (Environmental Systems Research Institute, Inc. <http://www.esri.com>).

## RESULTS

### *Model Complexity and Selection*

Fifty-nine species were sufficiently prevalent in the Clinch and Holston subbasins to build models to assess the prediction accuracy, generality, and resolution of multiple logistic regression models that used from 1 to 6 predictors. Species prevalence varied but many of these

species were relatively common (mean prevalence = 31% in the Clinch and 25% in the Holston, respectively). Correlations between predictor variables were weak (Pearson correlation < 0.60) except between Strahler order and downstream link in the Appalachian Plateau (Table 3.1). All correlations were significant ( $p < 0.01$ ) due to the large number of reaches ( $N = 18,487$ ). Models were built from all possible combinations of 6 predictors (1,236 models) using 133 samples from the Clinch subbasin. For most species many (> 50) models had significant likelihood ratio statistics. For most species cross-validation AIC and AUC varied substantially among all possible models, but for most species many (> 50) models had similar (< 5% difference) AIC and AUC values.

#### *Model Complexity and Occurrence Characteristics*

For most species, cross-validation SE, SP, and CC increased with increasing model complexity (Figure 3.2). However, 95% confidence intervals for these statistics overlapped among most levels of complexity, which reflect the large inter-specific variation in model response to increasing complexity. Kappa increased with increasing complexity, especially for common species. Mean prediction accuracy was relatively high, but varied among species (mean AUC = 0.82, range = 0.66 to 0.98 for the most complex models). Accuracy (AUC) of the best model was generally high, varied among species, and slightly increased with each increase in complexity (mean change in AUC = 0.02), but > 3 predictors usually improved accuracy little (Figure 3.2).

Prediction generality was assessed using 154 species occurrence samples from the Holston subbasin and the best model built to predict fish occurrence in the Clinch subbasin. Generality as measured by SE, SP, and CC usually increased slightly with increasing complexity (mean change in AUC = 0.01), but 95% confidence intervals indicate few significant increases in generality with increasing model complexity. Sensitivity was low for rare species and high for common species, but varied little with increasing model complexity. Specificity was high for rare species and relatively low for common species, but increased with increasing model complexity for common species. Correct classification was high for rare species and low for common species and varied little with increasing model complexity. Models with > 3 predictors had similar SE, SP, CC, and AUC (Figure 3.2). Kappa was low and variable and did not increase with model complexity for rare species, but did increase with increasing model complexity for common species. Mean prediction generality was relatively low and varied among species (mean AUC = 0.67, range = 0.23 to 0.87 for the most complex models). All statistics describing model generality were lower than those describing cross-validation accuracy, but prediction generality



was similar to prediction accuracy and changes in accuracy and generality with increasing complexity were similar.

The resolution of predictive models usually increases with increasing model complexity, especially for common species (Figure 3.3). Mean patch length averaged 1126 km for all species at all levels of complexity and decreased with increasing species prevalence. Mean patch length was only 23 km (SD = 5) for common species (present in > 40% of samples), but was 3712 km (SD = 2267) for rare species (present in < 10% of samples). The greatest decrease in mean patch length occurred when model complexity increased from 1 to 2 predictors for common species (Figure 3.3). However, mean patch length usually decreases as model complexity increases. Total patch length (the sum of the length of all reaches where a species is predicted to be present) decreases slightly for rare species and increases markedly for common species (Figure 3.3) as model complexity increases.

#### *Relations Between Prediction Accuracy, Generality, and Resolution and Species Traits*

Although some accuracy statistics (i.e., SE, SP, and Kappa) are affected by species prevalence, prediction accuracy and generality as measured by AUC were not strongly affected by species prevalence. Prediction accuracy and generality moderately decreased as species prevalence increased, but the relations between accuracy and generality with prevalence are weak, in part due to large variation in AUC at low prevalence (Figure 3.4). Cross-validation AUC is significantly negatively correlated with species prevalence ( $p < 0.01$ ), but little variability in AUC is explained by prevalence ( $R^2 = 0.12$ ). Generality is not correlated with species prevalence ( $p < 0.18$ ) and little variability in AUC is explained by prevalence ( $R^2 = 0.01$ ). However, the removal of two outlier species results in a significant relationship ( $p = 0.01$ ), but little variability in AUC is explained by prevalence ( $R^2 = 0.11$ ). The outlier species (*Clinostomus funduloides* and *Etheostoma swannanoa*) were rare in the Clinch and Holston subbasins. Prevalence was 5% and 3% for *C. funduloides* and 8% and 4% for *E. swannanoa* for the Clinch and Holston subbasins, respectively.

Descriptions of maximum length and nesting behavior were available for 48 species for which I predicted occurrence and assessed prediction accuracy, generality, and resolution in the Holston and Clinch subbasins (Table 3.2). Comparison among families was possible for 5 families with > 2 species. Six species from 4 families were analyzed together. Assessing relations between occurrence characteristics and nesting behavior were not possible for catfishes and darters because their behaviors did not vary sufficiently among species.

Relations between mean patch length and total patch length and species traits were weak (Kendall's  $\tau < 0.6$ ) and insignificant for most comparisons (Table 3.2). A strong relation between

prevalence and MPL was found only for catfishes and sunfishes, but the relation was negative for catfishes, positive for sunfishes, and significant for only sunfishes. Prevalence was also strongly and significantly related to TPL for sunfishes, and significantly, but weakly related for minnows and all species. Maximum recorded length was negatively related to MPL for only “other species”, but the relation was marginally significant. Maximum recorded length was positively related to TPL for only catfishes, but the relation was not significant. Nesting behavior was not related to MPL or TPL for any family. The small number of species and low variation in traits within families may have hindered comparisons (Table 3.2).

#### *System-wide Prediction and Assessment*

At least 187 fish species are known to be present in the UTRB, but only 129 (69%) species were recorded in 524 occurrence samples and only 120 (64%) species were sufficiently prevalent (present in > 4 samples) to attempt statistical prediction of occurrences. Most of these 120 species are relatively rare (mean prevalence = 13.7), with only 9 (7.5%) species in  $\geq 50\%$  of the samples and 111 (92.5%) species in < 50% of the samples.

Statistically significant ( $p < 0.05$ ) logistic regression models were built for 111 of 120 (92.5%) species (Appendix A). Sensitivity was generally low, but varied among species (mean = 27.3, standard deviation = 31.0). Specificity and CC were high and less variable among species (mean = 91.2 and 90.0, standard deviation = 20.6 and 12.0, respectively). Kappa statistics were generally low, but varied among species (mean = 0.22, standard deviation = 0.23; Figure 3.4). Kappa statistics indicate that only 70 (58.3 %) models predicted occurrence more accurately than random. Cross-validation AUC statistics indicate that most models accurately predicted occurrence (mean = 0.85, standard deviation = 0.10). Cross-validation AUC and Kappa statistics were significantly correlated, but varied substantially (Pearson correlation = 0.271,  $p = 0.003$ ). Sensitivity, SP, CC, and Kappa are affected by species prevalence (Figure 3.5).

Models built based on habitat descriptions gleaned from reference books performed poorly (Figure 3.6). Although SE was high (mean = 61.4, standard deviation = 26.8), SP and CC were low (mean = 4.8, standard deviation = 13.2 for specificity and mean = 66.3, standard deviation = 28.0 for correct classification). Logistic regression models had significantly lower sensitivity ( $t = -8.94$ ,  $p < 0.001$ ), and significantly higher specificity ( $t = 37.99$ ,  $p < 0.001$ ) and correct classification ( $t = 7.83$ ,  $p < .0001$ ) than reference book models for 110 species modeled using both methods. The accuracy of reference book models is not affected by species prevalence, except at very low prevalence where all accuracy statistics vary among species (Figure 3.6). Models built based on habitat descriptions gleaned from reference books often had higher SE, but much lower SP and CC and lower Kappa than MLR models (Figures 3.5 and 3.6).

These differences are due to the low resolution of the reference book models that predict presence in most locations for most species regardless of their prevalence in sample data.

I located 13 studies in the peer-reviewed literature that provide model prediction accuracy statistics (Appendix B). Species occurrence was predicted for molluscs (1 study), birds (4 studies), fishes (3 studies), reptiles (1 study), plants (4 studies), mammals (1 study), and insects (1 study). Predictions were made based on MLR (10 studies), generalized regression analyses and spatial prediction (GRASP; 1 study), multivariate discriminant analysis (MDA; 1 study), artificial neural networks (ANNs; 2 studies), a classification tree (TRE; 1 study), multivariate adaptive regression splines (MARS; 1 study), a Bayesian MLR approach (BYS; 1 study), and a rule-based model (RB; 1 study). Accuracy was assessed using a variety of data, including validation (generality assessment; 8 studies), cross-validation (accuracy assessment; 5 studies) and resubstitution (6 studies). Resubstitution assesses a model's fit to the data used to build the model. Several studies used > 1 modeling method, assessment dataset, and/or assessment statistic, and few described species prevalence or how critical values were selected. Several studies provided only averages of assessment statistics for > 1 species.

Combined, these studies predicted the occurrence of > 140 species. The literature models had high SE (mean = 65.4, standard deviation = 5.3), but low SP (mean = 68.2, standard deviation = 5.6) and CC (mean = 73.9, standard deviation = 3.0). Kappa and AUC statistics were high and variable (mean = 95.0, standard deviation = 29.2 for Kappa and mean = 85.0, standard deviation = 2.4 for AUC). Most models described in the literature had higher SE, lower SP and CC, and much higher Kappa than my MLR models (Figure 3.7). Models described in the literature and my MLR models had similar AUC, suggesting that the models predict with similar accuracy, but that many models described in the literature predict absence poorly compared to my MLR models. Descriptions of the spatial resolution were not provided in any peer-reviewed model, preventing assessment of how the predictions of occurrence vary spatially.

## DISCUSSION

Knowledge of species occurrence is a prerequisite for efficient conservation, allowing conservation efforts to be directed to where they will be effective and away from where they will be ineffective. Predictions of occurrence are often used in lieu of survey information describing species occurrence. Unfortunately, prediction accuracy, generality and resolution are often low and may lead to poor conservation decisions, promoting the development of models that better predict species occurrence and the ability to predict the characteristics of model predictions. Increasing the complexity of the models does not often substantially improve prediction

performance. Most models that predict species occurrence are not rigorously assessed, but many likely predict with low accuracy, generality, and resolution. Prediction accuracy, generality, and resolution differ among species, due to differences in their detectability, prevalence, and other traits, suggesting that models that perform differently be applied differently. Improving the accuracy, generality, and resolution of predictions of species occurrence is a worthy endeavor, but scientists should also facilitate the appropriate applications of available predictions. Although the accuracy, generality and resolution of my models are limited, they provide the best descriptions of the likely occurrence of many species and they can provide useful information for management and conservation.

#### *Model Complexity and Selection*

I assessed the effect of model complexity on prediction accuracy to ensure that I used the most accurate, parsimonious models to assess relations between prediction characteristics and species traits. Unfortunately, sufficient occurrence data were not available to predict the occurrence of most species among stream reaches within specific sub-drainages where they were present. The paucity of species occurrence data for specific regions likely constrains the prediction of occurrence for most species to large spatial extents using predictors available at those extents, which likely decreased prediction accuracy (Osborne and Soares-Seane 2002, Zhang et al. 2008). Selecting a best model from among all possible models is not recommended as a model selection method (Burnham and Anderson 2002, Olden et al. 2004), but it did ensure that the most accurate candidate model was used to assess the effect of model complexity on prediction accuracy, generality, and resolution and relations between model characteristics and species traits. Comparing all possible models also demonstrated that for many species many statistically significant and similarly accurate models could be built. Interpretation of predictors and extrapolation of results should be done cautiously, but using integrative, remotely sensed predictors may reduce the importance of which predictors are used in predictive models. This information may benefit conservation efforts that need to predict the occurrence of many species for which little information describing their habitat preferences is available. A small number of integrative predictors can provide relatively accurate predictions of occurrence from many species.

#### *Model Complexity and Prediction Accuracy and Resolution*

Increasing model complexity beyond 2 or 3 predictors did not substantially improve prediction accuracy for most of my study species and I suggest that a similar pattern holds for most fish species. Although some scientists have moderately improved prediction accuracy by using more complex methods or by using additional predictors (e.g., Olden and Jackson 2002,

Segurado and Araujo 2004, Oakes et al. 2005), the results are often equivocal (Manel et al. 1999b, Peterson and Cohoon 1999, Porter et al. 2000, Chapter 2). Similar to the results of Beard et al. (1999) and Porter et al. (2000), I found that increasing model complexity beyond 2 or 3 predictors improved mean prediction accuracy little, regardless of species prevalence. Interestingly, Hakanson (1995) also found that regression models are usually optimally accurate when they use a small number of predictors, suggesting that the effect of model complexity is limited by the data and statistical tools used (Hastie and Tibsharini 1990) and how they are applied.

Using different predictors is also unlikely to substantially improve prediction accuracy. Porter et al. (2000) suggest that using predictors that describe habitat at fine grains (i.e., microhabitat predictors) can improve the accuracy of predictions of stream fish occurrence. However, collecting field data to predict species occurrence at fine grains and large extents is infrequently a viable option. Furthermore, the accuracy improvements reported by Porter et al. (2000) were similar to those I found by using more complex models that include only remotely sensed predictors. That is, using additional remotely sensed predictors or microhabitat predictors provide similar accuracy improvements. Several of the predictors I used describe factors that affect stream structure and function (Vannote et al. 1980, Leopold 1994) and are integrative, describing habitat at the spatial scale of dominant controlling processes (Strayer et al. 2006). Finding accurate predictors that describe watersheds and stream reaches, but that are not strongly correlated with one or more of the predictors I used seems unlikely.

The usefulness of fine-grain data to predict species occurrence is limited by their availability and temporal variability. Data that describe stream habitat at mesohabitat grains (e.g., individual pools), such as the existence of barriers to movement, water temperature, and dominant substrate (Schlosser 1991, Torgerson et al. 1999, Fausch et al. 2002, Gresswell et al. 2006) can predict fish occurrence or explain unexpected occurrences. For example, the presence of barriers might explain unexpected absences (Neves and Angermeier 1990, Winston and Taylor 1991, Quinn and Kwak 2003) and guide conservation efforts (Dunham et al. 2002). Methods for collecting such data at large extents are being developed (Torgerson et al. 1999, 2001, Fausch et al. 2002, Betts et al. 2006, Lorang et al. 2005), but are infrequently used. Using predictors collected at fine spatial grains does not guarantee accurate and high-resolution predictions. As the grain of habitat descriptors decreases (resolution increases) their measured temporal variability usually increases (Leopold and Langbein 1962, Leopold 1994). For example, reach-scale stream gradient changes over geological time scales, stream confluence locations and reach-scale meander patterns are relatively stable for decades, but pool depths and temperature can vary

seasonally and mesohabitat and microhabitat changing flows may occur with short recurrence intervals (Leopold 1994). Using highly temporally variable habitat data for large spatial extents may require frequent data and model updates. The temporally stable predictors that I used affect stream system structure and function (Leopold 1994, Knighton 1984) and provide coarse predictions that are appropriate for use with occurrence data collected over decades, but they may not facilitate predictions of occurrence at the fine spatiotemporal grains at which most individual stream fish likely perceive their environment (Holland et al. 2004, Olden et al. 2004).

Importantly, the short-term dynamics of species occurrence are not of particular interest in conservation; which focuses on long-term persistence of populations, often at large spatial extents. Whether the coarse predictors I used facilitate predictions of occurrence for biotic units at higher levels of organization (e.g., demes, populations, or assemblages) is not known.

The contention that higher resolution predictors might improve model predictions made by Fransen et al. (2006) is supported by temporally stable upstream extents of fish occurrence through their 2-year study (Cole et al. 2006, Fransen et al. 2006). However, discontinuous fish occurrence at fine grains and within large regions may be less temporally stable due to changes in habitat conditions and fish movement (Gowan et al. 1994, Fausch and Young 1995, Albanese et al. 2004, Taylor et al. 2006). Predictions of occurrence from coarse, temporally stable predictors might best be considered likely occurrences over the temporal extent of change of the predictors or zones of intermittent use due to disturbance and species interactions (Angermeier et al. 2002, Fausch et al. 2002) or changes in habitat conditions and fish movement (Fransen et al. 2006). Given the likely temporal variability in species occurrence, the application of static models of species occurrence should be tempered with rigorous assessments of prediction accuracy over time.

The temporal extent over which occurrence data were collected affects prediction accuracy and resolution. Using occurrence data that were collected over a long temporal extent likely increased SE, but reduced SP, CC, Kappa and prediction resolution for most species in the UTRB. Adding presence detections to assemblage samples when > 1 sample was collected from a reach increased the spatial extent and number of predicted presences. For example, if the spatial extent or frequency of detection of a species changes through time, the models will overestimate presence because all presence detections are used as a presence for building models regardless of the number of samples collected or the habitat conditions at the sample location. Overestimation of the extent of species occurrence due to temporally extensive sampling might be more likely for rare species because abundance strongly affects observed geographic extent

and frequency of detection. Such errors should be considered when assessing and applying predictions of occurrence (Cardillo et al. 1999).

The delineation of patches of species' habitats is a prerequisite for efficient conservation (Flebbe 1994) because most species are discontinuously distributed at some spatiotemporal scale (Weins 1989, Angermeier et al. 2002), possibly due to the spatial distribution of suitable habitat and temporal changes in the suitability of habitat (Brown 1984, Angermeier and Schlosser 1989, Kolasa 1989). The usefulness of patch-based models to stream fish conservation depends on how well patches can be identified (Dunham et al. 2002) and managed for conservation objectives. My models predicted discontinuous distributions for most species. Greater model complexity moderately improved model resolution, but the effect of model complexity on resolution was relatively small for many species. The ability to directly observe environmental attributes that species perceive as habitat often allows scientists to identify habitat patches of terrestrial animals (Boone and Krohn 2000, Holland et al. 2004), but a priori delineation of patches often imposes an anthropogenic scale of analysis that can bias results (Weins 1989). In some studies scientists could identify patches of stream fish habitat a priori (e.g., Rieman and McIntyre 1995, Dunham and Rieman 1999, Isaak and Thurow 2006), but identifying patches of stream fish habitat is often tenuous due to the absence of knowledge of habitat preferences and information describing habitat attributes that affect fish occurrence (e.g., temperature or water chemistry). My predictions of occurrence delineate patches of habitat that are defined by the data and can be used to guide research and conservation efforts. How well the patches of habitat are defined varies with model complexity and among species.

Greater model complexity increased the resolution of fish occurrence predictions, but mean patch lengths were long for most species and the effect of model complexity on resolution decreased with increasing species prevalence. I expected that increasing model complexity would better delineate the concurrence of environmental discontinuities and fish occurrence, reducing mean patch length (MPL) and the total length of stream in which species are predicted present (TPL). For common species, greater model complexity reduced MPL little, whereas TPL increased because more predictions of presence were made. More complex models allowed for a greater range of conditions with predictions of presence. Using only 2 or 3 predictors to predict the occurrence of common species may underestimate their range, possibly because they are able to use alternative habitat as defined by the more complex models. Greater model complexity reduced MPL and TPL of rare species, but their MPL and TPL are large compared to those of common species. Low resolution in models for rare species may often be due to failure of the predictors to describe important but unknown environmental gradients. However, a more

parsimonious explanation for the effect of prevalence on MPL and TPL is that low resolution is a statistical artifact of the small number of presence records for rare species in the data used to build the models and the critical values used to assign presence and absence predictions.

Complex models can improve prediction resolution, but they require sufficient presence and absence data to do so and effect of species prevalence should be accounted for in the application of model predictions. Although prediction accuracy may be similar for rare and common species for some statistics (e.g., AUC), low resolution for rare species may limit the usefulness of model predictions and should affect how predictions of occurrence of species that differ in prevalence are applied. Precise predictions of occurrence for rare species should not be expected without sufficient presence data, regardless of prediction accuracy. Model predictions for rare species should be interpreted and applied with greater caution than those for common species. Models that predict the presence of rare species might be accurately describing suitable habitat, but the presence of suitable habitat does not ensure the presence of the species. Managers should include additional and/or larger locations where rare species are predicted to be present in conservation and monitoring plans.

Reliance on accuracy statistics provided overly optimistic assessments of the usefulness of predictive models for many species. Assessing model resolution provided additional information that can be used to direct the appropriate application of model predictions. Predicted discontinuities of fish presence can be used to direct conservation actions. Further investigation of discontinuities in fish species occurrence and cautious application of the predictions to guide conservation efforts are warranted. Although consideration of the spatial pattern and relative spatial position of habitat patches is a frequent consideration for terrestrial conservation, the vector nature of lotic systems and difficulty in delimiting fish habitat, especially at large spatial extents, complicate their use for stream fish conservation (but see Isaak and Thurow 2006). The predicted occurrences of several species (e.g., *Cyprinella monacha*, *Cottus carolinae* and *Etheostoma zonale*) indicate that they are present in relatively large patches (contiguous reaches) of headwater stream (sensu Gomi et al. 2002) that are separated by patches with low predicted probability of presence. Other species (e.g., *Ambloplites rupestris*) occur in patches of large streams with discontinuities among and within large streams. Most predicted occurrences are discontinuous, but patch sizes are large for most species, indicating that the models do not differentiate among most adjacent reaches or that the species do not discriminate among many contiguous reaches. The coarseness of predictions of occurrence limits their use to relatively large spatial extents. It is unlikely that differences in predicted probabilities of occurrence among most adjacent reaches are large or ecologically significant. Despite small differences in predicted



probabilities of occurrence, observed occurrence is often discontinuous for rare species (Angermeier et al. 2002), supporting the notion that habitat is often not limiting the occurrence of rare species.

Concern for parsimony may be exaggerated for relatively simple pragmatic models (Loehle 1983, Starfield 1997) like those I built. For these data the cautious use of moderately overfit models is warranted due to increased prediction accuracy, generality, and resolution. Contrary to expectation (Levins 1966) and similar to the results found by Porter et al. (2000), increasing model complexity usually slightly improved model generality. Greater generality may be the result of including predictors that are important in the assessment region (Leftwich et al. 1997). Relatively low generality compared to accuracy suggests that, when possible, models should be built using data from the region in which they will be applied. However, prediction accuracy and generality varied among rare and common species more than among models of different complexity for those species (Figure 3.2), suggesting that additional occurrence data, especially presence data, might improve model performance more than increasing model complexity. Concern for model parsimony may be exaggerated, but the improvement in prediction accuracy, generality, and resolution achievable via increasing model complexity, is limited.

#### *Relations Between Prediction Accuracy, Generality and Resolution and Species Traits*

Scientists should expect the accuracy, generality and resolution of predictions of species occurrence to be related to species traits. The mechanisms that define lotic system and generate patterns of habitat that affect species occurrence produce a spatial and temporal continuum of environmental gradients within the stream network (Vannote et al. 1980). Discontinuities in the continuum are identifiable as parts of a spatial hierarchy (Frissell et al. 1988) and fish might perceive these discontinuities as boundaries (Holling 1992) that delineate populations or population segments. The apparent structure of stream fish populations typically reflects the fact that many of the habitat units that species respond to are spatially discrete (e.g., Frissell et al. 1988, Pringle et al. 1988, Hawkins et al. 1993). Observed prediction accuracy, generality and resolution can reflect real species occurrence and how scientists observe or perceive the system.

Because I define habitat patches based on environmental attributes and species occurrence data, I expect characteristics of suitable patches to be related to species traits that affect the species' spatial distribution and persistence in suitable habitat (Boone and Krohn 1999, Goldstein and Meador 2004, Seoane et al. 2005). That is, rather than impose an anthropocentric hierarchy of spatial units on streams (e.g., Frissell et al. 1986, Townsend and Hildrew 1994) that may affect perceived patterns of occurrence, I used empirical data and models to predict

occurrence along a continuum (e.g., Vannote et al. 1980, Wiens 1989), and looked for emergent patterns (i.e., patches and discontinuities). Effective use of occurrence predictions for conservation requires knowledge of species interactions with patches of habitat and the matrix in which the patches are present (Pringle et al. 1988, Gotelli and Taylor 1999). The concept of landscape filters suggests that environmental attributes “filter” or exclude certain species traits (and species with those traits) at a range of spatiotemporal scales and thus affect species occurrence patterns (Poff 1997). Lamouroux et al. (2002) found that traits of stream fish communities were related to characteristics of reach- and microhabitat-scale environmental gradients. Thus, the accuracy, generality and resolution of the predictions of occurrence might be due to relations between the environmental attributes used as predictors and species traits. Strong relations between prediction accuracy, generality and resolution and species traits would support the contention that predicted patches of occurrence delineate ecologically meaningful spatial units and boundaries that are perceived by species (at some spatiotemporal scale) and affect their occurrence. Strong relations between occurrence characteristics and species traits would also support the use of the patches as conservation and monitoring units and might facilitate the prediction of occurrence characteristics of different species.

The prediction of occurrence characteristics is precluded by the generally weak and inconsistent relations between predicted occurrence characteristics (MPL and TPL) and species prevalence, maximum recorded length, and nesting behavior among and across families (Table 3.2). A parsimonious explanation for the failure to find consistent relations between occurrence characteristics and species traits is that the models did not predict fish occurrence with sufficient accuracy and/or resolution to define ecologically meaningful patches. Inadequate or incorrect predictor variables, fish occurrence sampling errors, and/or improper model specification are possible causes of inaccurate and low-resolution model predictions. Accurate, high resolution prediction of patches is a prerequisite for assessing relations between prediction characteristics and species traits. Collecting species occurrence data that facilitate accurate prediction of occurrence, including spatial and temporal variability in occurrence and detectability, is a logical first step for additional research. Alternatively, the species traits I examined may not be important factors in determining occurrence of many species among reaches in the UTRB. Also, the statistics I used to describe prediction characteristics (i.e., MPL and TL) may not provide meaningful descriptions of prediction characteristics at the spatial extent of the UTRB (Goldstein and Meador 2004). Given accurate high-resolution predictions of species occurrence, hypotheses about how traits affect occurrence characteristics can be developed and tested using a variety of statistics.

Predictions of fish occurrence are made at many grains and the grain of prediction and how critical values are applied to define patches constrains prediction characteristics. For example, Mastrorillo et al. (1997) predicted the occurrence of 3 fish species within microhabitats, Porter et al. (2000) predicted the occurrence of 14 fishes among stream reaches, Wall et al. (2004) predicted the occurrence of Topeka shiner (*Notropis topeka*) in valley segments (several contiguous stream reaches), Dunham et al. (2002) predicted the occurrence of bull trout (*Salvelinus confluentus*) among watersheds, and Fransen et al. (2007) predicted the upstream extent of any fish species at 10-m increments. Using reaches as prediction units constrained the minimum grain of predictions for my models, likely to patches that are larger than the scale of an individual fish's perception of habitat (Holland et al. 2004). That is, while stream fish movement is increasingly acknowledged (Gowan et al. 1994, Fausch and Young 1995, Albanese et al. 2004) it is unlikely that most individuals of any given species respond to the overall conditions of these patches of habitat, complicating the detection of relations between prediction characteristics and species traits. The detection of relations between traits and occurrence characteristics might usefully be investigated by examining the responses of individuals as well as demes or populations. Because prediction models were built using occurrence data that were collected over a long temporal extent, the patches may be scaled to the response of populations or demes. Furthermore, because the characteristics of patches can vary substantially with the (often arbitrary) selection of critical values, relations between prediction characteristics and species traits may be obscured by the selection of critical values. If patch-based models are to be useful for stream fish conservation, the delineation of patches will necessarily be coarse until such conjectures can be evaluated.

Absence of strong relations between occurrence characteristics and species traits is not convincing evidence that such relations do not exist. However, regardless of the explanation, failure to detect the expected patterns emphasizes the need to rigorously assess the accuracy and resolution of predictions of occurrence and to cautiously apply the predictions to conservation decisions. Differences in the directions of relations between prediction characteristic and species traits suggest that how ecological attributes affect occurrence characteristics differs among fish families. More importantly, variation in prediction characteristics among species suggests that the value of predictions of occurrence vary with their accuracy and resolution. Whatever the reason for the weak and variable relations I observed, these models provide the best available descriptions of species occurrence and can inform some conservation decisions.

### *System-wide Prediction and Assessment*

Confidence in a model encourages its use, whereas rigorous model assessment encourages its appropriate use. The accuracy and generality of model predictions and the realism of model structure promote confidence in a model, but it is unlikely that all of these attributes can be simultaneously satisfactorily met (Levins 1966). Although methods for assessing models are well developed (Fielding and Bell 1997), the results of rigorous model assessments are infrequently reported, objective criteria for assessing model accuracy are not well developed and expectations for prediction accuracy and resolution are infrequently stated. Comparing the accuracy of statistical models built to predict fish occurrence in the UTRB with the accuracy of alternative models for the same species and with results from studies described in peer-reviewed literature substantiates the usefulness of statistical models for making some conservation decisions, identifies some common limitations, and demonstrates the need for the rigorous assessment and cautious and flexible application of statistical model predictions of occurrence to conservation problems.

Insufficient occurrence data was a problem for my work and likely for many other projects. Insufficient occurrence data precluded the prediction of occurrence of several species and limited the prediction accuracy (Stockwell and Peterson 2002) and resolution for some rare species. Although mean prediction accuracy was similar for common and rare species for which significant models could be built, prediction accuracy varied substantially among rare species and prediction assessments may be questionable for them. However, it is unlikely that additional occurrence data would substantially improve the accuracy of the predictions of occurrence for most species. Nevertheless, collecting additional occurrence data, especially presence data, could increase the number of species for which models can be built and might improve the accuracy of models for some rare species.

Low species prevalence is likely a common problem because most species are relatively uncommon throughout most of their range (Gaston 1994) and few regions are likely to be sufficiently sampled. Although my sample size was large ( $N = 524$ ), most species were rare (Figure 3.8). Low prevalence will make predicting occurrence difficult until a minimum number of presence detections are attained (Karl et al. 2000). This may require a very large number of samples and/or reducing the frequency of sampling omission errors. Methods to ensure sufficient sampling at samples sites and across the extent of prediction are available (Angermeier and Smogor 1995, Chapter 1), but the sufficiency of sampling is infrequently addressed. The results of a sampling simulation in which increasing numbers of samples were randomly collected from the 524 assemblage samples from the UTRB demonstrate that  $> 650$  samples are necessary to

collect  $\geq 95\%$  of the fish species likely present in the UTRB, given the sampling methods employed to collect the data used in this study (Figure 3.9). The fish assemblage in the UTRB is insufficiently sampled, precluding the prediction of occurrence for several rare species. Sampling omission errors also likely limit prediction accuracy (Thuiller et al. 2004). The large number of samples I used may have increased the likelihood of including uncommon or erroneous observations and decreased prediction accuracy (Clarkson and Wilson 1995). Methods are available to account for imperfect presence detection (Bayley and Peterson 2001, Peterson and Dunham 2003, McKenzie 2005, Royle et al. 2005), but the necessary repeated sampling data are not available for the UTRB.

In lieu of sufficient occurrence data, the opinions of experts are sometimes used to predict species occurrence (Pearce et al. 2001, Petit et al. 2003, Berg et al. 2004). Comparing the accuracy of models based on the opinions of experts to the accuracy of statistical models can improve the rigor of model assessment by providing objective assessment criteria. Models based on habitat descriptions from reference books performed very poorly compared to statistical models, supporting the use of statistical models. Expert opinion can improve prediction performance (Martin et al. 2005), but it does not ensure accurate predictions (Seoane et al. 2005). The paucity of quantitative data describing stream fish habitat at the scale of stream reaches and watersheds available in reference books precluded accurate prediction of occurrence. Most of the descriptions of habitat preference provided in reference books described habitat at grains too fine to allow prediction at large spatial extents. Assessments of micro- and meso-habitat (Frissell et al. 1988) do not adequately describe the habitat or predict the occurrence of many fishes (Richards et al. 1996, Schlosser 1991, 1995, Grossman et al. 1995). The description of quantitative habitat descriptors that are amenable to remote sensing should be promoted to enhance the value of reference books and to promote better understanding of the factors affecting fish occurrence.

The similarity of my results, especially for common species, to results reported in the peer-reviewed literature for a range of species and prediction methods suggests a common limitation to prediction accuracy. Although methods for building, assessing, and applying models are well developed (Fielding and Bell 1997, Guisan and Zimmermann 2000), the dearth of rigorous model assessments, especially assessments of prediction resolution, precludes conclusions regarding the usefulness of many predictions of species occurrence. However, the available accuracy assessment results demonstrate surprising similarities. The overall accuracy of most models that predict species occurrence is similar among models, regardless of species, data, or modeling method used. Most models predict occurrence correctly about 70 to 90% of the time

(i.e., range of CC is about 0.70 to 0.90 and AUC is usually about 0.80). Most variation in model accuracy appears to be due to the interaction of species prevalence, the critical value selected, and the statistic used to describe accuracy.

Unfortunately, few studies rigorously assess model predictions or describe how their predictions should be applied to conservation problems. Specifically, how critical values for presence and absence are selected given consideration of model resolution, species prevalence, likely sampling errors, and the application of the predictions, are infrequently described. Applying different critical values can produce substantially different depictions of a species likely geographic distribution and habitat use. For example, using the same model to predict the occurrence of snubnose darter (*Etheostoma simoterum*) in the UTRB, a critical value of 0.5 predicts their presence in 8,012 km whereas a critical value of 0.6 predicts their presence in 3,726 km. Using the larger critical value also results in a larger number of smaller patches of predicted presence. Rigorous assessment of model predictions, including the assessment of model accuracy, generality, and resolution of many alternative models for rare and common species provided insights into how model predictions can be appropriately assessed and applied. Several statistics were necessary to adequately describe prediction accuracy. SE, SP, and CC were informative, but are sensitive to species prevalence and the critical value used. Manel et al. (2001) recommend using Kappa to assess model predictions, but Kappa is also affected by prevalence and requires selection of a critical value (Byrt, et al. 1993, Guggenmoos-Holzmann 1996, Fielding and Bell 1997). Testing the significance of Kappa complemented other accuracy statistics by demonstrating that significant model fit (i.e., likelihood ratio  $< 0.05$ ) and high accuracy (e.g., CC and AUC  $> 0.90$ ) do not ensure that a model predicts better than random. However, the results of this test can change if a different critical value is used. The AUC was a useful model assessment statistic because it is less affected by prevalence and does not require the selection of a critical value (Fielding and Bell 1997), but it is insensitive to relatively small differences in prediction accuracy and does not describe the type of error encountered when the model is applied. Rigorous assessment of model predictions should guide their appropriate application.

Critical values can be adjusted to optimize the usefulness of predictions (Guisan and Zimmerman 2000, Hosmer and Lemeshow 2000). I adjusted critical values to increase sensitivity and to account for the effect of (usually low) prevalence on predictions from logistic regression. I used prevalence as a critical value to allow meaningful assessment of prediction accuracy and comparisons of prediction accuracy among species, but it did not fully account for the effect of prevalence on prediction accuracy. For conservation applications, critical values can be adjusted

to increase SE and increase the likelihood that the species entire habitat is delineated. This can inform conservation decisions if anthropogenic impacts are planned in locations of predicted presence. Alternatively, predicted probabilities of species occurrence can be mapped to provide additional information to those making conservation decisions (Elith et al. 2002, Miller et al. 2004). Although predicted probabilities of occurrence are not proportional to the probability of use by a species (Keating and Cherry 2004), they do provide information describing the locations of likely suitable habitat and differences in the probability of occurrence among locations, which can inform conservation decisions. Although the prediction accuracy, generality and resolution of many of my models were limited, they provide the best descriptions available of the likely occurrence of many species. While not perfect, such models are useful for management and conservation because they describe likely species occurrence that should be used to inform decisions, such as if species-specific conservation practices are warranted for a construction project.

## CONCLUSIONS

Static statistical models that use temporally stable remotely sensed predictors can predict fish occurrence among stream reaches with sufficient accuracy to facilitate many conservation decisions. However, low prediction accuracy and resolution, especially for rare species, limits their utility for making conservation decisions at fine spatial grains. Improving the accuracy, generality, and resolution of model predictions is a valuable goal. Additional occurrence data (especially presence detections) may improve the accuracy and resolution of models that predict the occurrence of rare species, but increasing model complexity is unlikely to substantially improve model prediction accuracy or resolution for most species. Substantial improvement of accuracy and resolution may not be possible, suggesting that protocols that facilitate the appropriate application of model predictions are needed.

Prediction accuracy and resolution vary among species, suggesting that prediction characteristics may be related to species ecological traits, and that anticipation of prediction characteristics is possible. However, I found little evidence of relations between prediction characteristics and species traits, which stymied efforts to anticipate prediction characteristics. Only species prevalence is strongly related to prediction characteristics, suggesting that prediction characteristics are statistical artifacts, at least for many species. Variation in prediction characteristics among species and the inability to anticipate prediction characteristics portend the need to rigorously assess model predictions to ensure their appropriate application to conservation problems.

The low accuracy and resolution of most models suggest that conservation of stream fishes must consider both the observed discontinuity of occurrences as well as the observed continuity of their habitat. Conservation actions may have to occur at large spatial extents and in several locations because we cannot accurately predict the occurrence of most species with high resolution.

#### ACKNOWLEDGMENTS

This work greatly benefited from the assistance of many people. In particular, I wish to thank L. Delarche Roghair, S. Klopfer, and Ken Convery of the Conservation Management Institute at Virginia Tech for providing outstanding GIS and database assistance. E. P. Smith, J. T. Peterson, C. T. Moore, and T. O. Brenden provided helpful statistical advice. K. Mattson generously provided figures of the UTRB. J. Copeland and R. Greenlee of the Virginia Department of Game and Inland Fisheries provided valuable advice and encouragement. The Virginia Department of Game and Inland Fisheries and the United States Geological Survey funded this work.



## LITERATURE CITED

- Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19: 716–723.
- Albanese, B., P. L. Angermeier, and S. Dorai-Raj. 2004. Ecological correlates of fish movement in a network of Virginia streams *Canadian Journal of Fisheries and Aquatic Sciences* 61: 857-869.
- Allen, T. F. H., R. V. O'Neill, and T. W. Hoekstra. 1984. Interlevel relations in ecological research and management: some working principles from hierarchy theory. *USDA Forest Service General Technical Reprint RM-110*. 11 p.
- Angermeier, P. L. 1995. Ecological attributes of extinction-prone species: loss of freshwater fishes of Virginia. *Conservation Biology* 9: 143-158.
- Angermeier, P. L., and I. J. Schlosser. 1989. Species-area relations for stream fishes. *Ecology* 70: 1450-1462.
- Angermeier, P. L., and M. R. Winston. 1999. Characterizing fish community diversity across Virginia landscapes: prerequisite for conservation. *Ecological Applications* 9: 335-349.
- Angermeier, P. L., and R. A. Smogor. 1995. Estimating number of species and relative abundances in stream-fish communities: effects of sampling effort and discontinuous spatial distributions. *Canadian Journal of Fisheries and Aquatic Sciences* 52: 936-949.
- Angermeier, P. L., K. L. Krueger, and C. A. Dolloff. 2002. Discontinuity in stream-fish distributions: Implications for assessing and predicting species occurrence. In Scott, J. M., P. J. Heglund, M. Morrison, M. Raphael, J. Haufler, and B. Wall, Editors. *Predicting species occurrences: issues of accuracy and scale*. Island Press, Covello, CA. 840 p.
- Austin, M. P. 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling* 157:101-118.
- Bayley, P. B. and J. T. Peterson. 2001. Species presence for zero observations: an approach and an application to estimate probability of occurrence of fish species and species richness. *Transactions of the American Fisheries Society* 130: 620-633.
- Beard, K. H., N. Hengartner, and D. K. Skelly. 1999. Effectiveness of predicting breeding bird distributions using probabilistic models. *Conservation Biology* 13: 1108-1116.
- Berg, A., U. Gardenfors, and T. von Proschwitz<sup>a</sup>. 2004. Logistic regression models for predicting occurrence of terrestrial mollusks in southern Sweden – importance of environmental data quality and model complexity.
- Betts, M. G., A. W. Diamonda, G. J. Forbes, M. A. Villard, and J. S. Gunne. 2006. The importance of spatial autocorrelation, extent and resolution in predicting forest bird occurrence *Ecological Modelling* 191: 197-224.
- Boone, R. B., and W. B. Krohn. 1999. Modeling the occurrence of bird species: are the errors predictable? *Ecological Applications* 9: 835-848.
- Boone, R. B., and W. B. Krohn. 2000. Predicting broad-scale occurrences of vertebrates in patchy landscapes. *Landscape Ecology* 15: 63-74.
- Brotons, L., W. Thuiller, M. B. Araujo, and A. H. Hirzel<sup>b</sup>. 2004. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography* 27: 437-448.
- Brown, J. H. 1984. On the relationship between abundance and distribution of species. *The American Naturalist* 124: 255-279.
- Burnham, K. P., and D. R. Anderson. 2002. *Model selection and multimodel inference: a practical-theoretic approach*, 2nd Edition. Springer, New York.
- Byrt, T., J. Bishop, and J. B. Carlin 1993. Bias, prevalence and kappa. *Journal of Clinical Epidemiology* 46: 423-429.
- Campbell Grant, E. H., W. H. Lowe, and W. F. Fagan. 2007. Living in the branches: population

- dynamics and ecological processes in dendritic networks. *Ecology Letters* 10: 165-175.
- Cardillo, M., D. W. Macdonald, and S. P. Rushton. 1999. Predicting mammal species richness and distributions: testing the effectiveness of satellite-derived land cover data. *Landscape Ecology* 14: 423-435.
- Clarkson, R. W., and J. R. Wilson. 1995. Trout biomass and stream habitat relationships in the White Mountains area, East-central Arizona. *Transactions of the American Fisheries Society* 124: 599-612.
- Cohen J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20: 37-46.
- Cole, M. B., D. M. Price, and B. R. Fransen. 2006. Change in the upper extent of fish distribution in eastern Washington streams between 2001 and 2002. *Transactions of the American Fisheries Society* 135: 634-642.
- Dunham, J. B., and B. E. Rieman. 1999. Metapopulation structure of bull trout: influences of physical, biotic, and geometrical landscape characteristics. *Ecological Applications* 9: 642-655.
- Dunham, J. B., B. E. Rieman, and J. T. Peterson. 2002. Patch-based models to predict species occurrence: lessons from salmonid fishes in streams. Pages 327 to 344 in J. M. Scott et al., editors. *Predicting Species Occurrence: Issues of Accuracy and Scale*, Island Press, Washington, D. C.
- Elith, J., M. A. Burgman, H. M. Regan. 2002. Mapping epistemic uncertainties and vague concepts in predictions of species distribution. *Ecological Modelling* 157: 313-329.
- Etnier, D. A. 1997. Jeopardized southeastern freshwater fishes: a search for causes. *In* G. W. Benz and D. E. Collins, editors. *Aquatic fauna in peril: the southeastern perspective*. Special Publication 1, Southeastern Aquatic Research Institute, Lenz Design & Communications, Decatur, Georgia.
- Etnier, D.A., and W.C. Starnes. 1993. *The Fishes of Tennessee*. The University of Tennessee Press, Knoxville, Tennessee. 704 p.
- Fausch, K. D., and M. K. Young. 1995. Evolutionary significant units and movement of resident stream fishes: a cautionary tale. Pages 360 - 370 in Nielson, editor. *Evolution and the Aquatic Ecosystem*, American Fisheries Society Symposium 17.
- Fausch, K. D., C. E. Torgersen, C. V. Baxter, and H. W. Li. 2002. Landscapes to riverscapes: bridging the gap between research and conservation of stream fishes. *BioScience* 52:483-498.
- Fenneman, N. H., and D. W. Johnson. 1946. *Physical division of the United States*, U.S. Geological Survey, Physiography Committee Special Map.
- Fielding, A. H. and J. F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation* 24: 38-49.
- Filipe, A. F., I. G. Cowx, and M. J. Collares-Pereira<sup>c</sup>. 2002. Spatial modeling of freshwater fish in semi-arid river systems: a tool for conservation. *River Research and Applications* 18: 123-136.
- Flebbe, P. A. 1994. A regional view of the margin: salmonid abundance and distribution in the Southeastern Appalachian Mountains of North Carolina and Virginia. *Transactions of the American Fisheries Society* 123: 657-667.
- Fleishman, E., R. MacNally, J. P. Fay, and D. D. Murphy. 2001. Modeling and predicting species occurrence using broad-scale environmental variables: an example with butterflies of the great basin. *Conservation Biology* 15: 1674-1685.
- Fransen, B. R., S. D. Duke, L. G. McWethy, J. K. Walter, and R. E. Bilby. 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.
- Frissell, C. A., W. J. Liss, C. E. Warren, and M. D. Hurley. 1986. A hierarchical framework for stream habitat classification: viewing streams in a watershed context. *Environmental*

- Management 10: 199-214.
- Gaston, K. J. 1994. *Rarity*. Chapman and Hall, London, U.K.
- Gaston, K. J., and T. M. Blackburn. 1996. *Rarity and Body Size: Importance of Generality*. Conservation Biology 10: 1295-1298.
- Goldstein, R. M., and M. R. Meador. 2004. Comparisons of fish species traits from small streams to large rivers. Transactions of the American Fisheries Society 133: 971-983.
- Gomi, T., R. C. Sidle, and J. S. Richardson. 2002. Understanding processes and downstream linkages of headwater systems. BioScience 52: 905-916.
- Gordon, N. D., T. A. McMahon, and B. L. Finlayson. 1992. *Stream hydrology: an introduction for ecologists*. John Wiley & Sons. New York.
- Gorman, O. T. 1986. Assemblage organization of stream fishes: the effect of rivers on adventitious streams. The American Naturalist 128: 611-615.
- Gorman, O. T., and J. R. Karr. 1978. Habitat structure and stream fish communities. Ecology 59: 507-515.
- Gotelli, N. J., and C. M. Taylor. 1999. Testing metapopulation models with stream-fish assemblages. Evolutionary Ecology Research 1: 835-845.
- Gowan, C., M. K. Young, K. D. Fausch, and S. C. Riley. 1994. Restricted movement in resident stream salmonids: a paradigm lost? Canadian Journal of Fisheries and Aquatic Sciences 51: 2626-2637.
- Gresswell, R. E., C. E. Torgersen, D. S. Bateman, T. J. Guy, S. R. Hendricks, and J. E. B. Wofford. 2006. A spatially explicit approach for evaluating relationships among coastal cutthroat trout, habitat and disturbance in small Oregon streams. American Fisheries Society Symposium 48: 457-471.
- Grossman, G. D., J. Hill, and J. T. Petty. 1995. Observations on habitat structure, population regulation, and habitat use with respect to evolutionary significant units: a landscape perspective for lotic systems. American Fisheries Society Symposium 17: 381-391.
- Guggenmoos-Holzmann, I. 1996. The meaning of Kappa: probabilistic concepts of reliability and validity revisited. Journal of Clinical Epidemiology 49: 775-782.
- Guisan, A., and U. Hofer<sup>d</sup>. 2003. Predicting reptile distributions at the mesoscale: relation to climate and topography. Journal of Biogeography 30: 1233-1243.
- Guisan, A., and N. E. Zimmermann. 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135: 147-186.
- Hakanson, L. 1995. Optimal size of predictive models. Ecological Modelling 78: 195-204.
- Hastie, T. J., and R. J. Tibshirani. 1990. *Generalized additive models*. Chapman and Hall. New York.
- Hawkins, C. P., J. L. Kershner, P. A. Bisson, M. D. Bryant, L. M. Decker, S. V. Gregory, D. A. McCullough, C. K. Overton, G. H. Reeves, R. J. Steedman, and M. K. Young. 1993. A hierarchical approach to classifying stream habitat features. Fisheries 18: 3-10.
- Holland, J. D., D. G. Bert, and L. Fahrig. 2004. Determining the spatial scale of species' response to habitat. BioScience 54: 227-233.
- Holling, C. S. 1992. Cross-scale morphology, geometry, and dynamics of ecosystems. Ecological Monographs 62: 447-502.
- Hosmer D. W., Jr., and S. Lemeshow. 2000. *Applied Logistic Regression*, 2nd Edition. John Wiley and Sons, New York.
- Hubert, W. A., and C. B. Chamberlain. 1996. Environmental Gradients Affect Rainbow Trout Populations among Lakes and Reservoirs in Wyoming. Transactions of the American Fisheries Society 125: 925 – 932.
- Isaak, D. J., and R. F. Thurow. 2006. Network-scale spatial and temporal variation in Chinook salmon (*Oncorhynchus tshawytscha*) redd distributions: patterns inferred from spatially continuous replicate surveys. Canadian Journal of Fisheries and Aquatic Sciences 63: 285-296.

- Jackson, D. A., P. R. Peres-Neto, and J. D. Olden. 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.
- Jenkins, R. E., and N. M. Burkhead. 1993. *Freshwater Fishes of Virginia*. American Fisheries Society, Bethesda, Maryland.
- Karl, J. W., P. J. Heglund, E. O. Garton, J. M. Scott, N. M. Wright, and R. L. Hutto. 2000. Sensitivity of species-habitat relationship model performance to factors of scale. *Ecological Applications* 10: 1690-1705.
- Keating, K. A., and S. Cherry. 2004. Use and interpretation of logistic regression in habitat selection studies. *Journal of Wildlife Management* 68: 774-789.
- Knighton, R. 1984. *Fluvial forms and processes*. Routledge, Chapman and Hall, Inc. New York.
- Kolasa, J. 1989. Ecological systems in hierarchical perspective: breaks in community structure and other perspectives. *Ecology* 70: 36-47.
- Kruse, C. G., W. A. Hubert, and F. J. Rahel. 1997. Geomorphic influences on the distribution of Yellowstone cutthroat trout in the Absaroka mountains, Wyoming. *Transactions of the American Fisheries Society* 126: 418-427.
- Lamouroux, N., and P. L. Angermeier. 2002. Intercontinental convergence of stream fish community traits along geomorphic and hydrologic gradients. *Ecology* 83: 1792-1807.
- Leftwich, K. N., P. L. Angermeier, and C. A. Dolloff. 1997. Factors influencing behavior and transferability of habitat models for a benthic stream fish. *Transactions of the American Fisheries Society* 126: 725-734.
- Lehmann, A., J. M. Overton, J. R. Leathwick<sup>e</sup>. 2002. GRASP: generalized regression analysis and spatial prediction. *Ecological Modelling* 157: 189-207.
- Leopold, L. B. 1994. *A View of the River*. Harvard University Press, Cambridge, MA.
- Leopold, L. B., and W. B. Langbein. 1962. The concept of entropy in landscape evolution. *Theoretical Papers in the Hydrologic and Geomorphic Sciences*, Geological Survey Professional Paper 500-A. US Government Printing Office, Washington D.C.
- Levins, R. 1966. The strategy of model building in population biology. *American Scientist* 54: 421-431.
- Liu, C., P. M. Berry, T. P. Dawson, and R. G. Pearson. 2005. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography* 28: 385-393.
- Loehle, C. 1983. Evaluation of theories and calculation tools in ecology. *Ecological Modelling* 19: 239-247.
- Lorang, M. S., D. C. Whited, F. R. Hauer, J. S. Kimball, and J. A. Stanford. 2005. Using airborne multispectral imagery to evaluate geomorphic work across floodplains of gravel-bed rivers. *Ecological Applications* 15: 1209-1222.
- MacKenzie, D. I., J. D. Nichols, N. Sutton, K. Kawanishi, and L. L. Bailey. 2005. Improving inferences in population studies of rare species that are detected imperfectly. *Ecology* 86: 1101-1113.
- Manel, S., H. C. Williams, and S. J. Ormerod. 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology* 38: 921-931.
- Manel, S., J. M. Dias, S. T. Buckton, and S. J. Ormerod<sup>f</sup>. 1999a. Alternative methods for predicting species distribution: an illustration with Himalayan river birds. *Journal of Applied Ecology* 36: 734-747.
- Manel, S., J-M. Dias, and S. J. Ormerod. 1999b. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modeling* 120: 337-347.
- Manly, B.F., L. McDonald, D. L. Thomas, T. L. McDonald, W. P. Erickson. 2002. *Resource Selection by Animals: Statistical Design and Analysis for Field Studies* 2nd ed., Wiley, 240 p.

- Martin, T. G., P. M. Kuhnert, K. Mengersen, and H. P. Possingham. 2005. The power of expert opinion in ecological models using Bayesian methods: impacts of grazing on birds. *Ecological Applications* 15: 266–280.
- Mastrorillo, S., S. Lek, F. Dauba, and A. Belaud. 1997. The use of artificial neural networks to predict the presence of small-bodied fish in a river. *Freshwater Biology* 38: 237-246.
- McCleary, R. J. and M. A. Hassan. 2008. Predictive modeling and spatial mapping of fish distributions in small streams of the Canadian Rocky Mountain foothills. *Canadian Journal of Fisheries and Aquatic Sciences* 65: 319-333.
- McCullagh, P., and J. A. Nelder. 1989 *Generalized Linear Models*, 2nd edition, Chapman & Hall, London.
- McPherson, J. M., W. Jetz, and D. J. Rogers. 2006. Using coarse-grained occurrence data to predict species distributions at finer spatial resolutions – possibilities and limitations. *Ecological Modelling* 192: 499-522.
- Miller, J. R., M. G. Turner, E. A. H. Smithwick, C. L. Dent, and E. H. Stanley. 2004. Spatial extrapolation: the science of predicting ecological patterns and processes. *BioScience* 54: 310-320.
- Muñoz, J., and A. M. Felicísimo<sup>e</sup>. 2004. Comparison of statistical methods commonly used in predictive modelling. *Journal of Vegetation Science* 15: 285-292.
- Neves, R. J., and P. L. Angermeier. 1990. Habitat alteration and its effects on native fishes in the upper Tennessee River system, east-central U.S.A. *Journal of Fish Biology* 37: 45-52 Supplement A.
- Oakes, R., M., K. B. Gido, J. A. Falke, J. D. Olden, and B. L. Brock. 2005. Modelling stream fishes in the great plains, USA. *Ecology of Freshwater Fish* 14: 361-374.
- Olden, J. D. and D. A. Jackson. 2000. Torturing data for the sake of generality: how valid are our regression models? *Ecoscience* 7: 301-310. (not used yet)
- Olden, J. D., and D. A. Jackson<sup>h</sup>. 2002. A comparison of statistical approaches for modeling fish species distributions. *Freshwater Biology* 47: 1976-1995.
- Olden, J. D., R. L. Schooley, J. B. Monroe, and N. L. Poff. 2004. Context-dependent perceptual ranges and their relevance to animal movements in landscapes. *Journal of Animal Ecology* 73: 1190-1194.
- Orrock, J. L., J. F. Pagels, W. J. McShea, and E. K. Harper<sup>i</sup>. 2000. Predicting presence and abundance of a small mammal species: the effect of scale and resolution. *Ecological Applications* 10: 1356-1366.
- Osborne, L. L., and M. J. Wiley. 1992. Influence of tributary spatial position on the structure of warmwater fish communities. *Canadian Journal of Fisheries and Aquatic Sciences* 49: 671-681.
- Osborne, P. E., J. C. Alonso, and R. G. Bryant<sup>j</sup>. 2001<sub>j</sub>. Modelling landscape-scale habitat use using GIS and remote sensing: a case study with great bustards. *Journal of Applied Ecology* 38: 458-471.
- Pearce, J. L., K. Cherry, M. Drielsma, S. Ferrier, and G. Whish. 2001. Incorporating expert opinion and fine-scale vegetation mapping into statistical models of faunal distribution. *Journal of Applied Ecology* 38: 412-424.
- Peterson, A. T., and K. P. Cohoon. 1999. Sensitivity of distributional prediction algorithms to geographic data completeness. *Ecological Modelling* 117: 159-164.
- Peterson, J. T. and J. Dunham. 2003. Combining inferences from models of sampling efficiency, detectability, and suitable habitat to classify landscapes for conservation of threatened bull trout, *Salvelinus confluentus*. *Conservation Biology* 17: 1070-1077.
- Petit, S., D. Chamberlain, K. Haysom, R. Pywell, J. Vickery, L. Warman, D. Allen, and L. Firbank<sup>k</sup>. 2003. Knowledge-based models for predicting species occurrence in arable conditions. *Ecography* 26: 626-640.

- Poff, N. L. 1997. Landscape filters and species traits: towards mechanistic understanding and prediction in stream ecology. *Journal of the North American Benthological Society* 16: 391-409.
- Porter, M. S., J. Rosenfeld, and E. A. Parkinson<sup>1</sup>. 2000. Predictive models of fish species distribution in the Blackwater Drainage, British Columbia. *North American Journal of Fisheries Management* 20: 349-359.
- Pringle, C. M., R. J. Naiman, G. Bretschko, J. R. Karr, M. W. Oswood, J. R. Webster, R. L. Welcomme, and M. J. Winterbourn. 1988. Patch dynamics in lotic systems: the stream as a mosaic. *Journal of the North American Benthological Society* 7: 503-524.
- Quinn, J. W., and T. J. Kwak. 2003. Fish assemblage changes in an Ozark River after impoundment: a long-term perspective. *Transactions of the American Fisheries Society* 132: 110-119.
- Rahel, F. J., and W. A. Hubert. 1991. Fish assemblages and habitat gradients in a Rocky Mountain-Great Plains stream: biotic zonation and additive patterns of community change. *Transactions of the American Fisheries Society* 120: 319-332.
- Rahel, F. J., J. D. Lyons, and P. A. Cochran. 1984. Stochastic or deterministic regulation of assemblage structure? It may depend on how the assemblage is defined. *American Naturalist* 124: 583-589.
- Reyes-Gavilan, F. G., R. Garrido, A. G. Nicieza, M. M. Toledo, and F. Brana. 1996. Fish community variation along physical gradients in short streams in northern Spain and the disruptive effects of dams. *Hydrobiologia* 321: 155-163.
- Reynolds, J. B. 1996. Electrofishing. Pages 221-253 in B. R. Murphy and D. W. Willis, editors. *Fisheries Techniques*, Second Edition. American Fisheries Society, Bethesda, Maryland.
- Richards, C., L. B. Johnson, and G. E. Host. 1996. Landscape-scale influences on stream habitats and biota. *Canadian Journal of Fisheries and Aquatic Sciences* 53: 295-311 Supplement 1.
- Rieman, B. E., and J. D. McIntyre. 1995. Occurrence of bull trout in naturally fragmented habitat patches of varied size. *Transactions of the American Fisheries Society* 124: 285-297.
- Royle, J. A., J. D. Nichols, and M. Kery. 2005. Modelling occurrence and abundance of species when detection is imperfect. *Oikos* 110: 353-359.
- Schlosser, I. J. 1991. Stream fish ecology: a landscape perspective. *BioScience* 41: 704-712.
- Schlosser, I. J. 1995. Critical landscape attributes that influence fish population dynamics in headwater streams. *Hydrobiologia* 303: 71-81.
- Schumm, S. 1977. *The Fluvial System*. John Wiley, New York.
- Scott, J. M., F. Davis, B. Csuti, R. Noss, B. Butterfield, C. Groves, H. Anderson, S. Caicco, F. D'Erchia, T. C. Edwards Jr., J. Ulliman, and G. Wright. 1993. Gap Analysis: a geographic approach to protection of biodiversity. *Wildlife Monographs* 123: 1-41.
- Scott, J. M., P. A. Heglund, M. L. Morrison, J. B. Haufler, M. B. Raphael, W. A. Wall, and F. B. Samson. 2002. *Predicting species occurrences: issues of accuracy and scale*. Island Press, Washington, D.C. 868 p.
- Segurado, P., and M. B. Araujo. 2004. An evaluation of methods for modelling species distributions. *Journal of Biogeography* 31: 1555-1568.
- Seoane, J., L. M. Carrascal, C. L. Alonso, and D. Palomino. 2005. Species-specific traits associated to prediction errors in bird habitat suitability modeling. *Ecological Modelling* 185: 299-308.
- Smith, T. A., and C. E. Kraft. 2005. Stream fish assemblages in relation to landscape position and local habitat variables. *Transactions of the American Fisheries Society* 134: 430-440.
- Sokal, R. R., and F. J. Rohlf. 1995. *Biometry: the Principles and Practices of Statistics in Biological Research*, third edition. W. H. Freeman and Company, New York.
- Sowa, S. P. 1998. Gap analysis in riverine environments. *Gap Analysis Bulletin* 7: 18-20.

- Starfield, A. M. 1997. A pragmatic approach to modeling for wildlife management. *Journal of Wildlife Management* 61: 261-270.
- Stockwell, D. R. B., and A. T. Peterson. 2002. Effects of sample size on accuracy of species distribution models. *Ecological Modelling* 148: 1-13.
- Strahler, A. N. 1952. Dynamic basis of geomorphology. *Geological Society of America Bulletin*. 63: 923-938.
- Strayer, D. L., H. M. Malcom, R. E. Bell, S. M. Carbotte, and F. O. Nitsche. 2006. Using geophysical information to define benthic habitats in a large river. *Freshwater Biology* 51: 25-38.
- Taylor, C. M., T. L. Holder, R. A. Fiorillo, L. R. Williams, R. B. Thomas, and M. L. Warren Jr. 2006. Distribution, abundance, and diversity of upland stream fishes under variable environmental conditions. *Canadian Journal of Fisheries and Aquatic Sciences* 63: 43-54.
- Thuiller, W., L. Brotons, M. B. Araujo, and S. Lavorel. 2004. Effects of restricting environmental range of data to project current and future species distributions. *Ecography* 27: 165-172.
- Tockner, K., J. V Ward, P. J Edwards, and J. Kollmann. 2002. Riverine landscapes: an introduction. *Freshwater Biology* 47: 497-500.
- Torgersen, C. E., D. M. Price, H. W. Li, and B. A. McIntosh. 1999. Multiscale thermal refugia and stream habitat associations of chinook salmon in northeastern Oregon. *Ecological Applications* 9: 301-319.
- Torgersen, C. E., R. N. Faux, B. A. McIntosh, N. J. Poage, and D. J. Norton. 2001. Airborne thermal remote sensing for water temperature assessment in rivers and streams. *Remote Sensing of Environment* 76: 386-398.
- Townsend, C. R. & Hildrew, A. G. 1994. Species traits in relation to a habitat templet for river systems. *Freshwater Biology* 31: 265-275.
- Van den Berg, M. S., W. Joosse, and H. Coops<sup>m</sup>. 2003. A statistical model predicting the occurrence and dynamics of submerged macrophytes in shallow lakes in the Netherlands. *Hydrobiologia* 509: 611-623.
- Vannote, R. L., G. W. Minshall, K. W. Cummins, J. R. Sedell, and C. E. Cushing. 1980. The river continuum concept. *Canadian Journal of Fisheries and Aquatic Sciences* 37: 130-137.
- Wall, S. S., C. R. Berry, Jr., C. M. Blausey, J. A. Jenks, and C. J. Kopplin. 2004. Fish-habitat modeling for gap analysis to conserve the endangered Topeka shiner (*Notropis topeka*). *Canadian Journal of Fisheries and Aquatic Sciences* 61: 954-973.
- Weins, J. A. 1989. Spatial scaling in ecology. *Functional Ecology* 3: 385-397.
- Williams, J. E., J. E. Johnson, D. A. Hendrickson, S. Contreras-Balderas, J. D. Williams, M. Navarro-Mendoza, D. E. McAllister, and J. E. Deacon. 1989. Fishes of North America endangered, threatened, or of special concern: 1989. *Fisheries* 14:2-3.
- Winston, M. R., C. M. Taylor. 1991. Upstream extirpation four minnow species due to damming of a prairie stream. *Transactions of the American Fisheries Society* 120: 98-105.
- Zhang, H., T. Thieling, S. C. Bates Prins, E. P. Smith, and M. Hudy. 2008. Model-based clustering in a brook trout classification study within the eastern United States. *Transactions of the American Fisheries Society* 137: 841-851.

Table 3.1. Pearson correlations among Strahler stream order, downstream link magnitude, mean reach elevation, mean reach gradient, reach length and reach sinuosity for 18,487 reaches in 3 physiographic provinces and all physiographic provinces of the upper Tennessee River basin. All correlations were significant ( $p < 0.01$ ).

	Order	Link	Elevation	Gradient	Length
<b>Appalachian Plateau (N = 1,216)</b>					
Link	0.64				
Elevation	-0.39	-0.42			
Gradient	-0.12	-0.01	0.14		
Length	-0.31	-0.11	0.17	0.13	
Sinuosity	0.07	0.02	-0.03	-0.08	0.35
<b>Blue Ridge (N = 8,790)</b>					
Link	0.58				
Elevation	-0.36	-0.40			
Gradient	-0.47	-0.26	0.34		
Length	-0.34	-0.08	0.14	0.22	
Sinuosity	0.14	0.13	-0.13	-0.02	0.36
<b>Ridge and Valley (N = 8,085)</b>					
Link	0.45				
Elevation	-0.18	-0.27			
Gradient	-0.38	-0.19	0.44		
Length	0.01	0.05	0.02	0.03	
Sinuosity	0.13	0.07	-0.08	-0.05	0.19
<b>All Physiographic Provinces Combined (N = 18,487)</b>					
Link	0.50				
Elevation	-0.23	-0.31			
Gradient	-0.39	-0.22	0.45		
Length	-0.11	0.02	0.01	0.07	
Sinuosity	0.13	0.09	-0.13	-0.06	0.24



Table 3.2. Kendall's  $\tau$  c rank correlations and probabilities between species prevalence (Prev), maximum recorded length of species (Max), and nesting behavior (Nes), and total length of predicted presence (TPL) and mean length of predicted patches of presence (MPL). Number of species compared are in parentheses and significant correlations ( $p \leq 0.05$ ) are in bold font. Identical nesting behaviors within groups precluded some comparisons.

	Prev	Max	Nes	Prev	Max	Nes	Prev	Max	Nes
	Catfishes (3)			Darters (8)			Minnows (17)		
TPL	0.33	1.0		-0.36	0.26		<b>-0.41</b>	-0.30	-0.23
	0.60	0.12		0.22	0.37		0.02	0.09	0.25
MPL	-1.0	-0.33		-0.29	0.42		-0.28	-0.27	0.02
	0.12	0.60		0.32	0.16		0.12	0.14	0.92
	Suckers (7)			Sunfishes (7)			Other Species (6)		
TPL	-0.05	-0.19	-0.036	<b>0.62</b>	0.33		-0.20	0.07	-0.39
	0.88	0.54	0.32	0.05	0.29		0.57	0.85	0.30
MPL	-0.05	-0.19	-0.36	<b>0.62</b>	-0.05		-0.07	-0.60	-0.54
	0.88	0.54	0.32	0.05	0.88		0.85	0.09	0.15
	All Species (48)								
TPL	-0.22	0.02	-0.18						
	<b>0.03</b>	0.79	0.12						
MPL	-0.17	-0.07	-0.02						
	0.08	0.49	0.85						

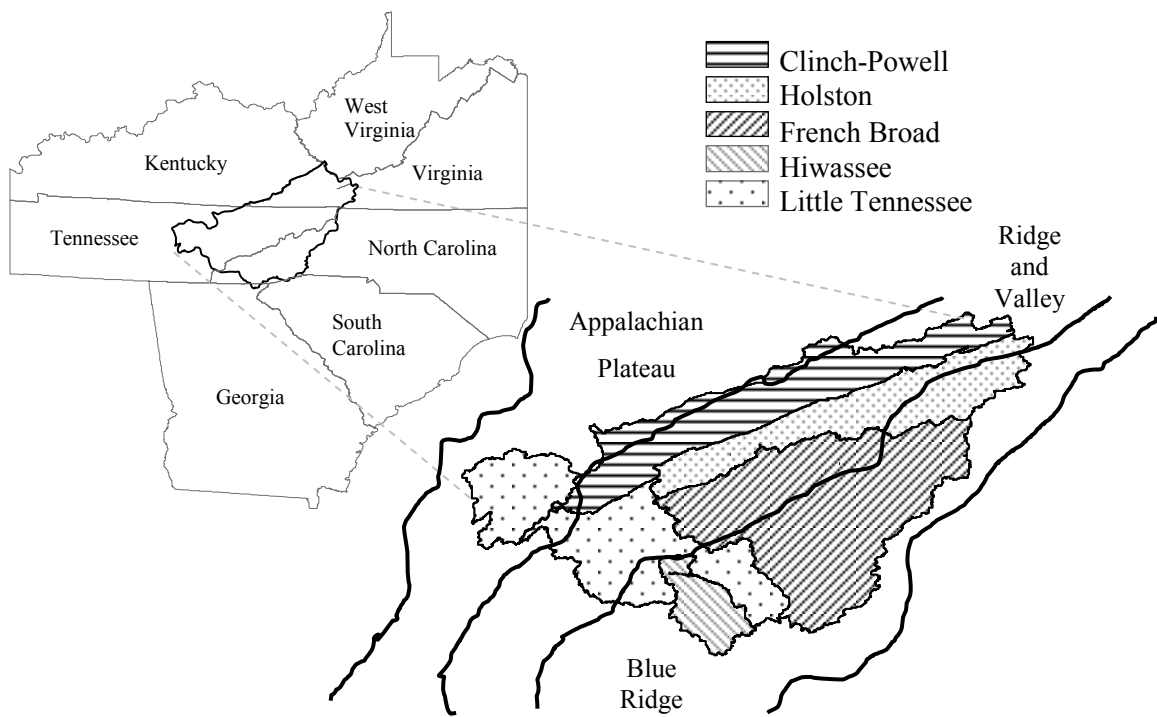


Figure 3.1. Location of the upper Tennessee River basin and its 5 major sub-basins and 3 physiographic provinces.

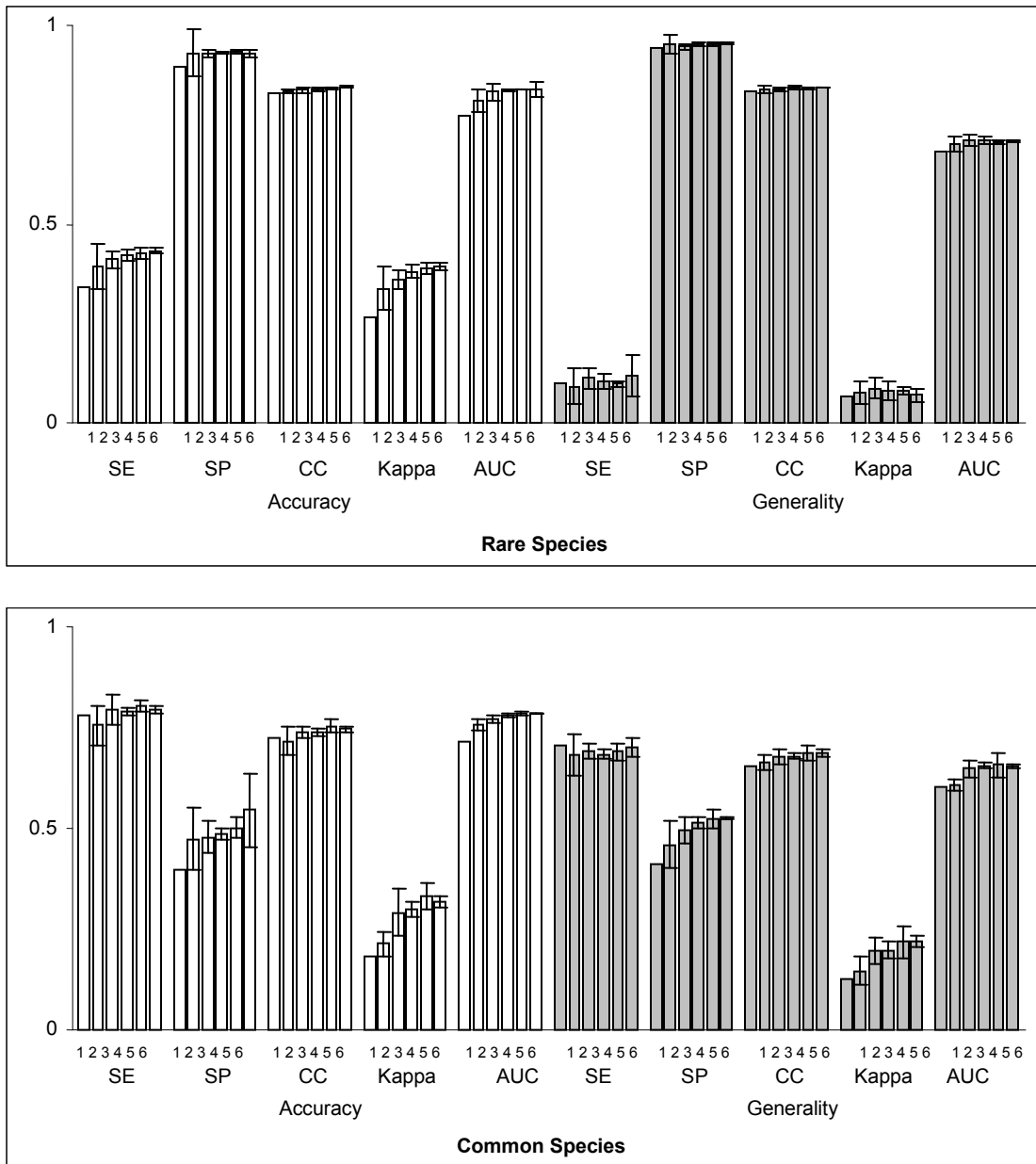


Figure 3.2. Mean and 95% confidence intervals for sensitivity (SE), specificity (SP), total correct classification (CC), Kappa statistic and area under the receiver operator characteristic curve (AUC) for logistic regression models built with 1 to 6 predictors to predict the occurrence of 59 stream fish in the upper Tennessee River basin. Statistics describing the prediction accuracy (unfilled) are cross-validation results for models built in the Clinch subbasin whereas those describing generality (filled) are the same models applied to the Holston subbasin. Statistics are calculated for rare (prevalence  $\leq 10\%$ ) and common (prevalence  $> 40\%$ ) species.

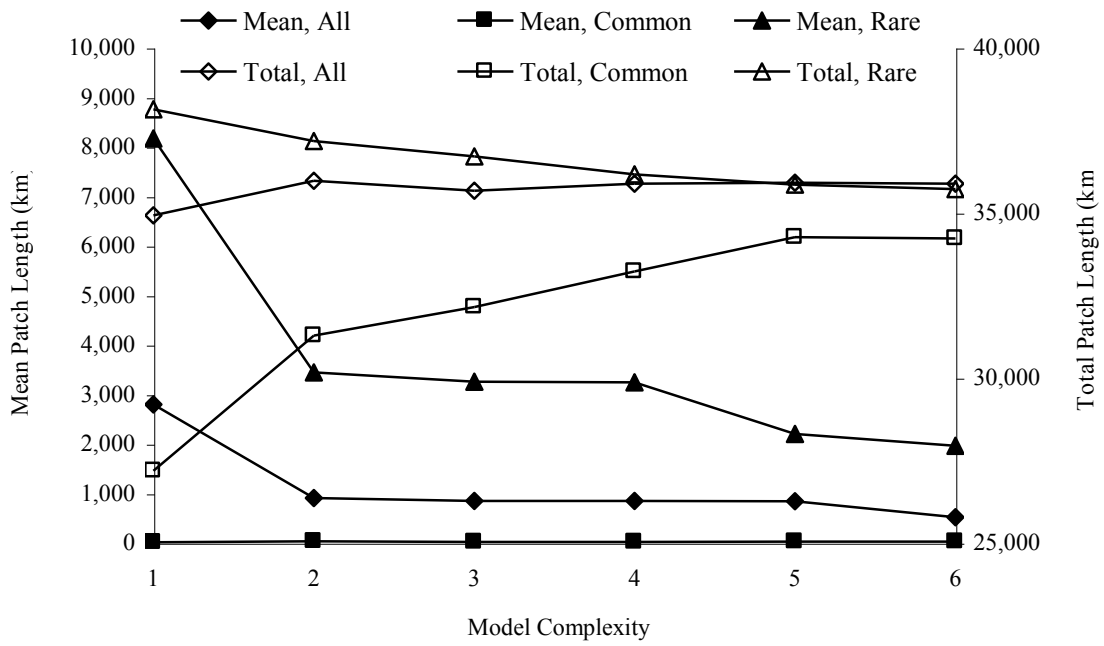


Figure 3.3. Relationships between model complexity (number of predictors), mean length of patches (filled symbols; contiguous reaches with predicted presence), and total length of patches (unfilled symbols) for 59 species, 13 rare (< 10% prevalence) species and 11 common (> 40% prevalence) species of stream fish for which predictions of occurrence were made in the Clinch subbasin.

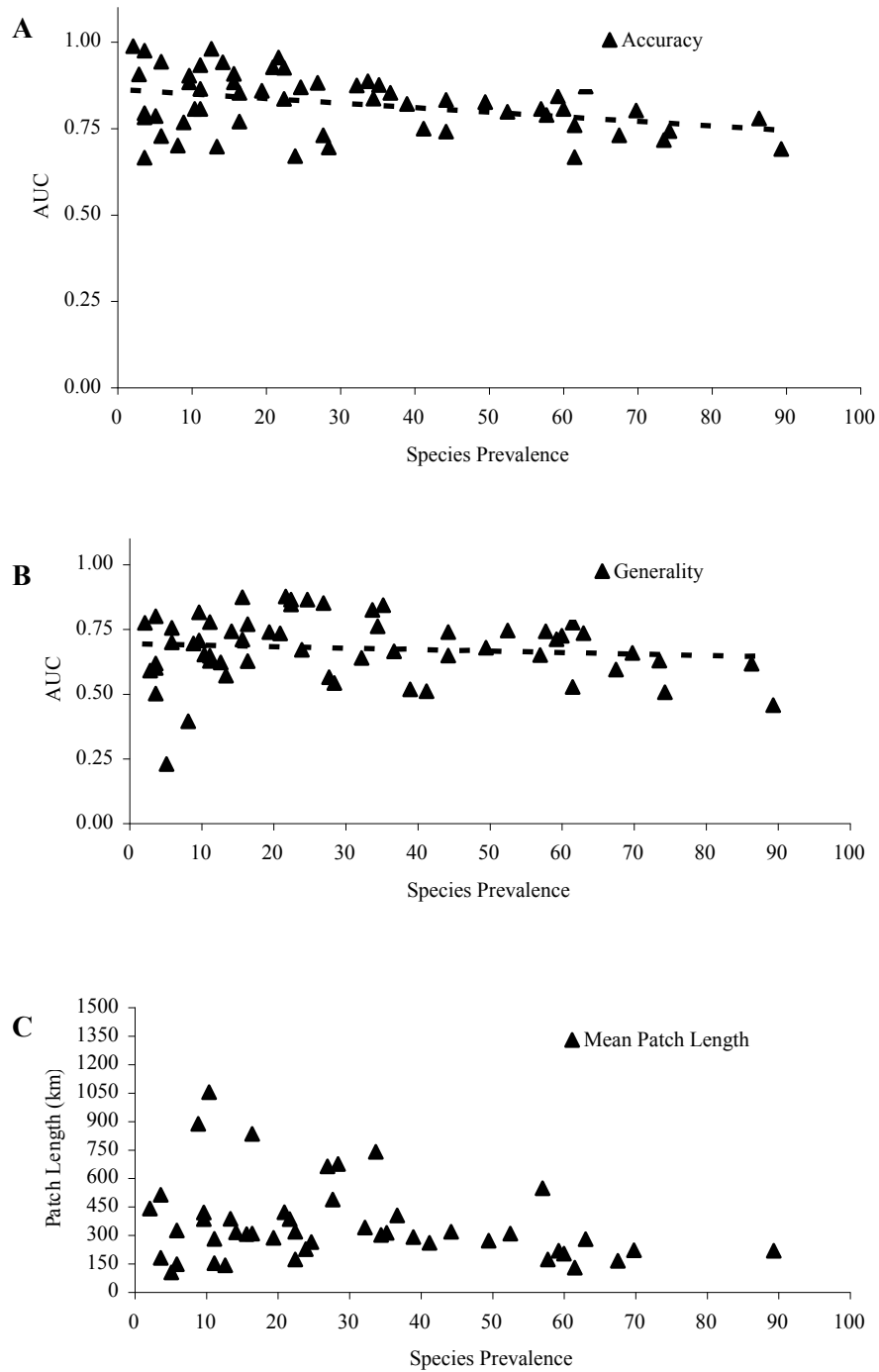


Figure 3.4. The relations between species prevalence in the Clinch subbasin and prediction accuracy (A; cross-validation AUC), species prevalence in the Holston subbasin and model generality (B; AUC for models built for the Clinch and applied in the Holston), and the relation between species prevalence in the upper Tennessee River basin and mean patch length (C). Dashed lines are the regression lines for the Clinch accuracy assessment and the Holston generality assessment. The model with the largest AUC for each species is used.

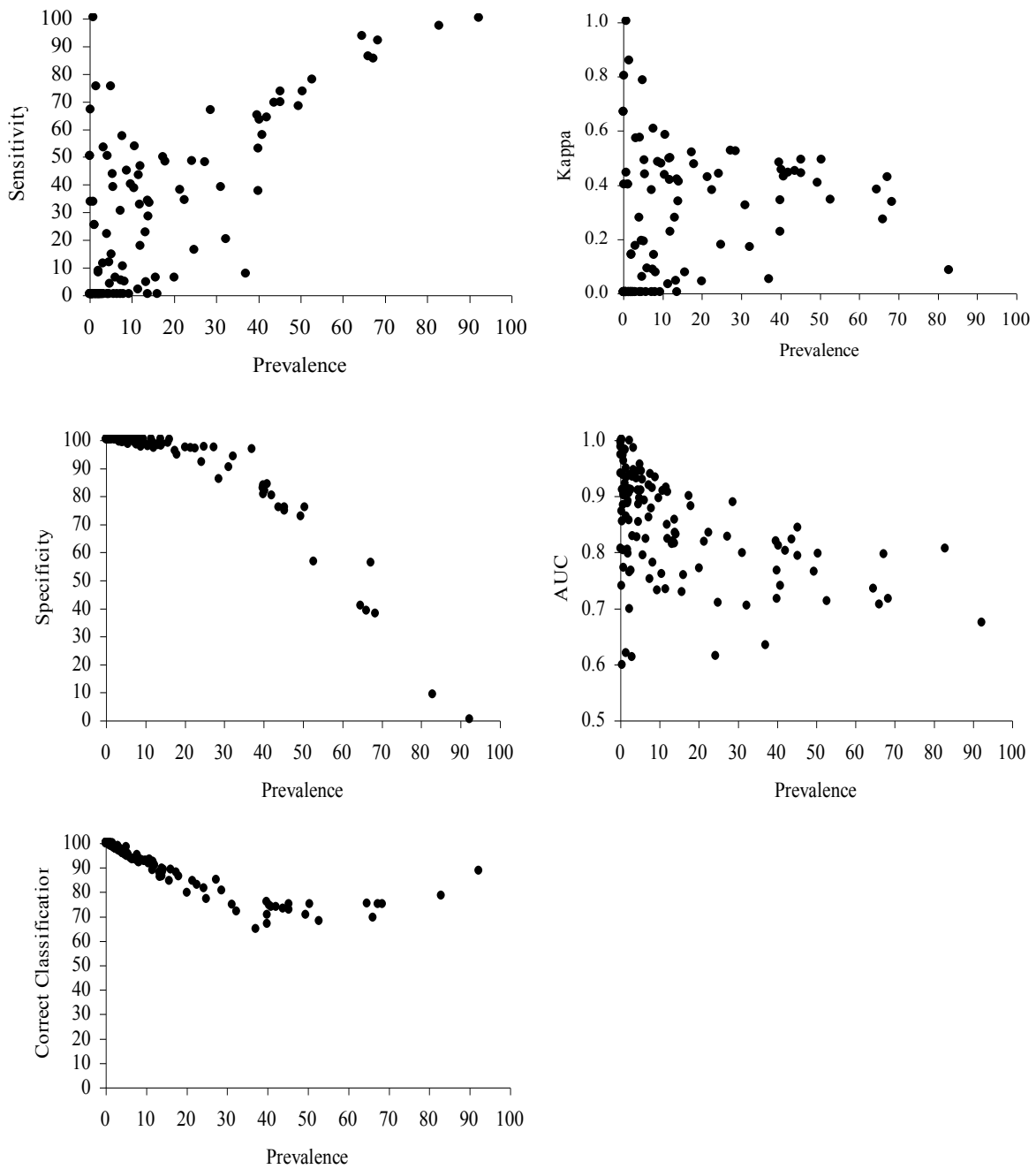


Figure 3.5. Relationships between species prevalence and sensitivity, specificity, correct classification, Kappa, and area under the receiver operating characteristic (AUC) prediction accuracy statistics for 120 logistic regression models that predict the occurrence of stream fishes in the upper Tennessee River basin.

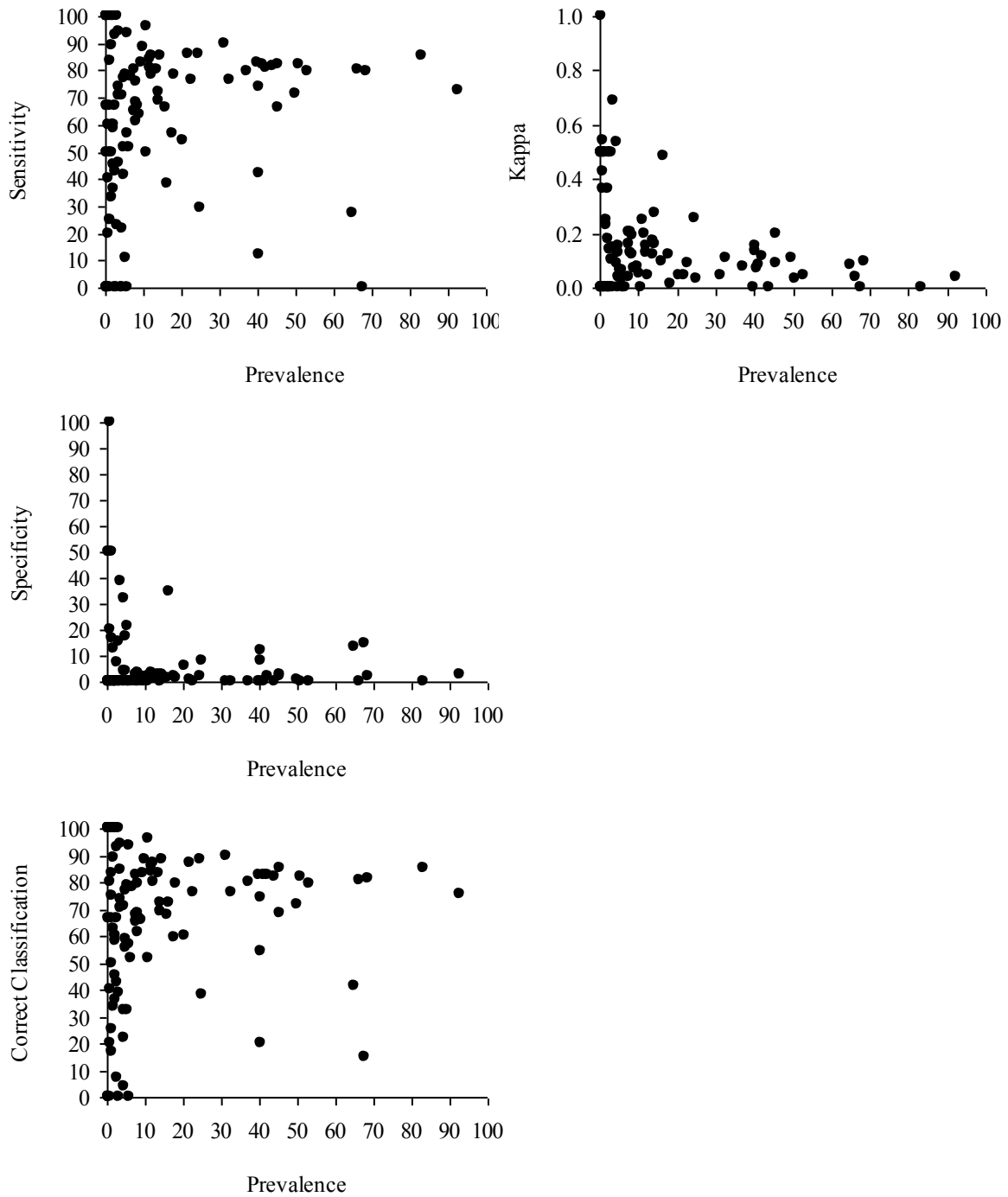


Figure 3.6. Relationships between species prevalence and sensitivity, specificity, correct classification, and Kappa prediction accuracy statistics for 104 reference book models that predict the occurrence of stream fishes in the upper Tennessee River basin.

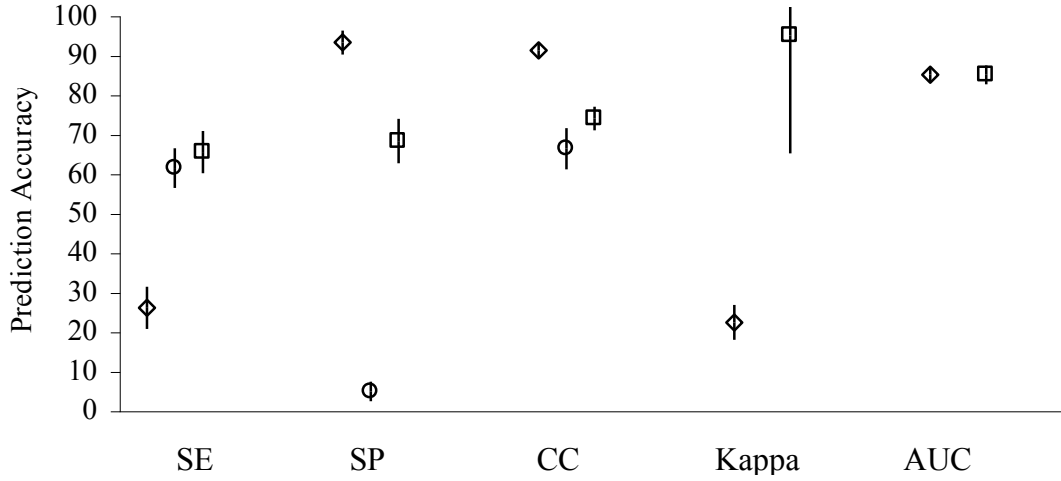


Figure 3.7. Mean and 95% confidence intervals for sensitivity (SE), specificity (SP), total correct classification (CC), Kappa statistic, and area under the receiver-operator characteristic curve (AUC) for models that predict species occurrence. Multiple logistic regression models (diamonds) and models based on the opinions of experts (circles) predict the occurrence of 120 fish species in the upper Tennessee River drainage. A variety of methods were used in 14 peer reviewed publications (squares) to predict the occurrence of >140 species.



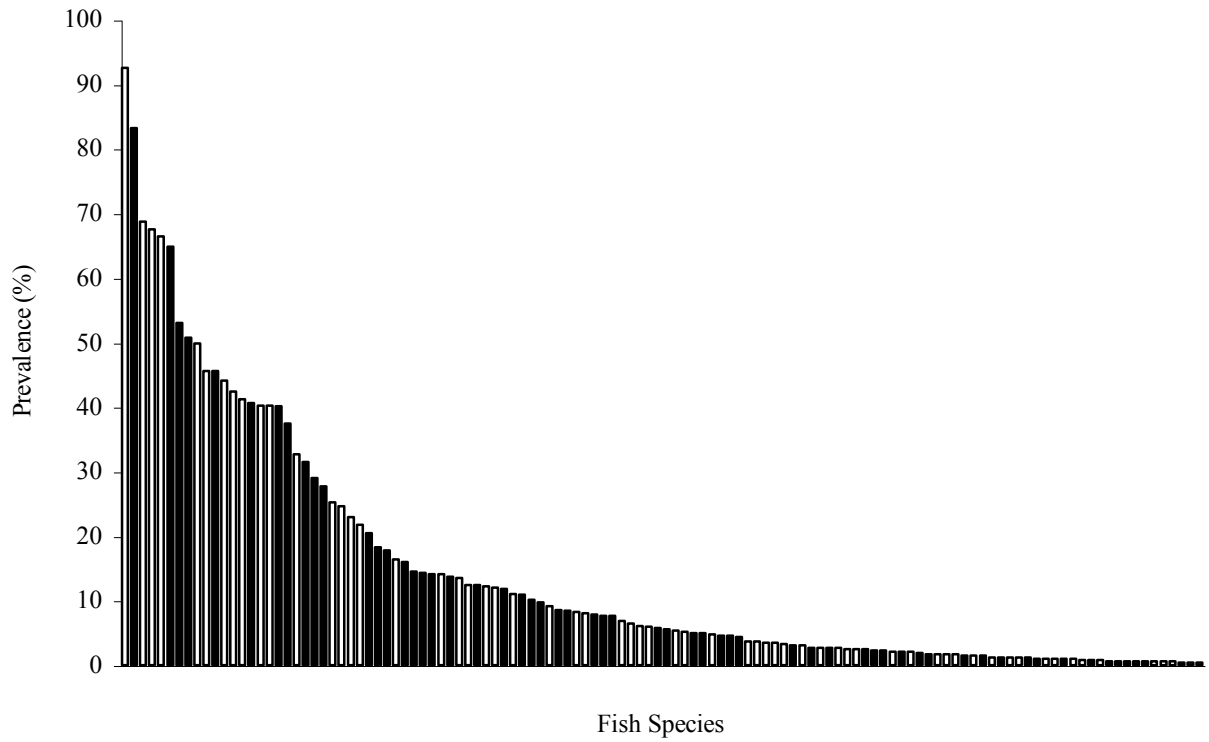


Figure 3.8. Prevalence of 120 fish species in the upper Tennessee River basin. Bars represent individual species. The height of each bar is the percent of 524 samples in which a species was present. Filled bars (N = 59) represent species used to assess the effect of model complexity on prediction accuracy, resolution and generality.

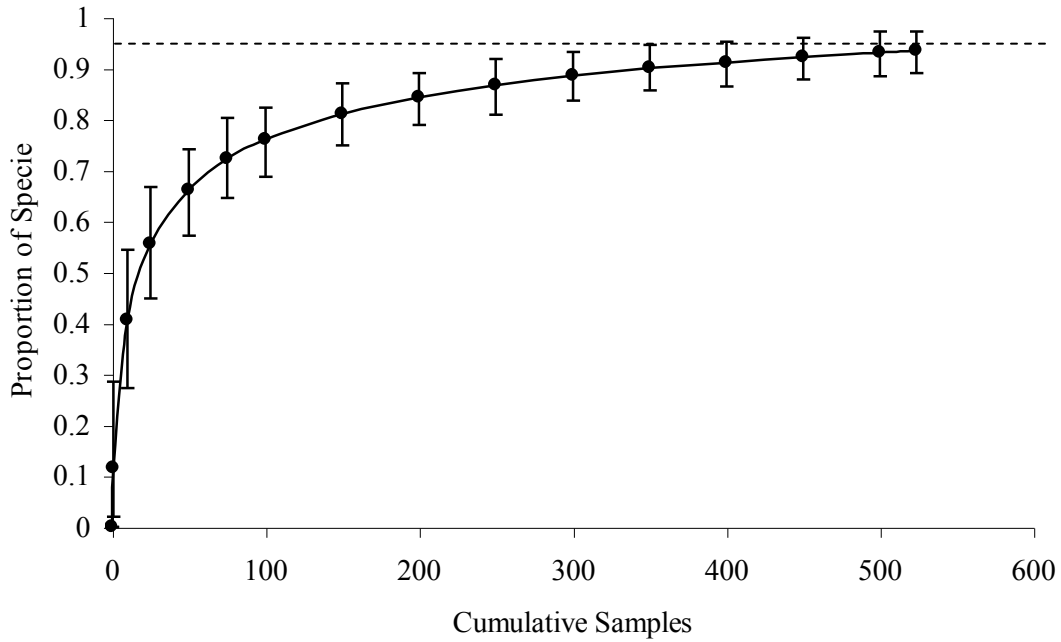


Figure 3.9. Sampling simulation results for 524 fish occurrence samples from the upper Tennessee River basin. Values of the ordinate are the mean proportion of the estimated species richness collected in 10,000 sample iterations at the sample size on the abscissa. Error bars are 95% bootstrap confidence bounds. The dashed line represents 95% of the estimate of species richness ( $N = 136$ ) that was calculated from the asymptote of the sampling simulation results.

Appendix A. Number of predictors (d.f.), likelihood ratio (l.r.) and probability, area under the receiver-operator characteristic curve (AUC), cross-validation sensitivity (SE), Specificity (SP) and total correct classification (CC), Kappa and p-value for logistic regression models that predict the occurrence of fish in the upper Tennessee River basin. For each species the model with the smallest AIC score and /or largest AUC among all models compared is described. Superscripts correspond to analytical groups in Table 3.2.

Species	d.f.	l.r.	p-value	AUC	SE	SP	CC	Kappa	p-value
Ambloplites rupestris	3	66.46	<0.001	0.71	85.97	38.62	68.89	0.27	<0.001
Ameiurus melas	2	8.06	0.005	0.80	0.00	100.00	99.62	0.00	0.500
Ameiurus natalis	2	41.22	<0.001	0.73	1.61	100.00	88.36	0.03	0.031
Ameiurus nebulosus	2	2.40	0.301	0.91	0.00	100.00	99.81	0.00	0.500
Ammocrypta clara	6	26.13	<0.001	1.00	33.33	99.81	99.43	0.40	<0.001
Aplodinotus grunniens	4	71.47	<0.001	0.93	21.74	98.80	95.42	0.27	<0.001
Campostoma anomalum	4	26.39	<0.001	0.67	99.78	0.00	88.17	0.00	0.358
Carassius auratus	1	4.14	0.126	0.77	0.00	100.00	98.85	0.00	0.500
Carpionodes cyprinus	4	21.85	<0.001	0.91	0.00	100.00	98.28	0.00	0.500
Catostomus commersoni	4	22.96	<0.001	0.63	7.41	96.42	64.31	0.05	0.026
Clinostomus funduloides	3	20.28	<0.001	0.83	0.00	100.00	97.52	0.00	0.500
Cottus baileyi <sup>6</sup>	5	65.17	<0.001	0.89	99.39	5.88	93.32	0.09	0.001
Cottus bairdii <sup>6</sup>	6	43.89	<0.001	0.76	0.00	99.79	88.74	0.00	0.362
Cottus carolinae	6	107.70	<0.001	0.73	93.26	40.44	74.81	0.38	<0.001
Cyprinella galactura <sup>3</sup>	6	151.32	<0.001	0.80	63.85	79.74	73.28	0.44	<0.001
Cyprinella monacha	5	35.05	<0.001	0.90	7.69	100.00	97.71	0.14	<0.001
Cyprinella spiloptera <sup>3</sup>	6	133.14	<0.001	0.82	37.72	96.83	83.97	0.42	<0.001
Cyprinus carpio <sup>3</sup>	5	88.00	<0.001	0.81	22.22	97.79	87.40	0.32	<0.001
Dorosoma cepedianum <sup>6</sup>	6	149.74	<0.001	0.91	53.45	97.64	92.75	0.58	<0.001
Dorosoma petenense	3	10.96	0.012	0.99	0.00	100.00	99.62	0.00	0.500
Erimystax cahni	1	18.91	<0.001	1.00	0.00	99.81	99.43	0.00	0.475
Erimystax dissimilis	7	152.58	<0.001	0.91	46.15	96.73	90.46	0.49	<0.001
Erimystax insignis	3	108.39	<0.001	0.85	32.26	98.70	90.84	0.41	<0.001
Esox masquinongy	4	15.16	0.002	0.90	100.00	100.00	100.00	0.00	0.500
Esox niger	4	18.02	0.001	1.00	93.26	40.44	74.81	0.00	0.500
Etheostoma acuticeps	7	31.97	<0.001	0.96	0.00	100.00	98.85	0.50	<0.001
Etheostoma blennioides <sup>2</sup>	4	224.96	<0.001	0.84	73.26	75.56	74.43	0.50	<0.001
Etheostoma caeruleum	4	17.15	0.0018	0.86	0.00	100.00	98.28	0.50	<0.001
Etheostoma camurum <sup>2</sup>	4	91.87	<0.001	0.93	38.71	98.17	94.66	0.43	<0.001
Etheostoma chlorbranchium	4	29.48	<0.001	0.91	8.33	99.80	97.71	0.14	<0.001
Etheostoma flabellare	6	56.85	<0.001	0.70	19.75	93.73	71.56	0.17	<0.001
Etheostoma jessiae	4	33.93	<0.001	0.79	0.00	100.00	93.89	0.00	0.500
Etheostoma kennicotti	3	42.62	<0.001	0.94	0.00	99.80	97.14	0.00	0.500
Etheostoma percnum	4	33.93	<0.001	0.95	0.00	100.00	98.28	0.00	0.500
Etheostoma rufilineatum <sup>2</sup>	6	183.37	<0.001	0.82	69.13	75.51	72.71	0.45	<0.001
Etheostoma simoterum	6	82.57	<0.001	0.72	91.60	37.72	74.43	0.33	<0.001
Etheostoma stigmaeum	4	52.21	<0.001	0.82	0.00	99.80	92.94	0.00	0.393
Etheostoma swannanoa <sup>2</sup>	3	31.92	<0.001	0.75	5.00	99.79	92.56	0.08	<0.001
Etheostoma tippecanoe	6	42.53	<0.001	0.98	25.00	100.00	98.85	0.40	<0.001
Etheostoma vulneratum <sup>2</sup>	2	55.53	<0.001	0.83	0.00	100.00	93.32	0.00	0.500
Etheostoma zonale	4	180.65	<0.001	0.86	48.06	91.65	80.92	0.09	<0.001

Appendix A. Continued.

Species	d.f.	l.r.	p-value	AUC	SE	SP	CC	Kappa	p-value
<i>Fundulus catenatus</i> <sup>6</sup>	3	9.02	0.011	0.61	0.00	100.00	91.60	0.00	0.500
<i>Gambusia affinis</i>	4	37.14	<0.001	0.91	0.00	100.00	97.33	0.00	0.500
<i>Hemitremia flammea</i>	6	18.18	0.001	1.00	50.00	100.00	99.81	0.67	<0.001
<i>Hybopsis amblops</i>	5	166.31	<0.001	0.94	38.18	97.44	91.22	0.43	<0.001
<i>Hypentelium nigricans</i>	4	71.96	<0.001	0.76	96.84	8.93	78.05	0.08	0.004
<i>Ichthyomyzon bdellium</i> <sup>6</sup>	5	96.80	<0.001	0.94	50.00	99.00	96.76	0.57	<0.001
<i>Ichthyomyzon greeleyi</i>	3	4.85	0.089	0.61	0.00	100.00	98.47	0.00	0.500
<i>Ictalurus punctatus</i> <sup>1</sup>	2	134.81	<0.001	0.93	44.68	97.27	92.56	0.48	<0.001
<i>Ictiobus bubalus</i>	3	16.13	0.0003	0.99	50.00	100.00	99.81	0.67	<0.001
<i>Ictiobus niger</i>	6	30.39	<0.001	1.00	66.67	100.00	99.81	0.80	<0.001
<i>Labides thessiculus</i>	2	27.75	<0.001	0.92	0.00	99.81	98.28	0.00	0.450
<i>Lampetra appendix</i>	3	10.39	0.016	0.80	0.00	100.00	98.09	0.00	0.500
<i>Lepisosteus osseus</i>	5	125.76	<0.001	0.96	75.00	99.19	97.90	0.78	<0.001
<i>Lepomis auritus</i> <sup>5</sup>	6	119.38	<0.001	0.77	52.58	80.30	70.04	0.34	<0.001
<i>Lepomis cyanellus</i> <sup>5</sup>	2	26.26	<0.001	0.73	0.00	100.00	92.18	0.00	0.500
<i>Lepomis gibbosus</i> <sup>5</sup>	4	10.24	0.006	0.76	0.00	100.00	97.33	0.00	0.500
<i>Lepomis gulosus</i>	2	27.16	<0.001	0.89	0.00	100.00	97.90	0.00	0.500
<i>Lepomis macrochirus</i> <sup>5</sup>	5	96.18	<0.001	0.74	57.42	83.81	73.28	0.42	<0.001
<i>Lepomis megalotis</i> <sup>5</sup>	8	122.75	<0.001	0.86	33.78	98.22	89.12	0.42	<0.001
<i>Lepomis microlophus</i>	3	13.96	<0.001	0.90	0.00	100.00	98.85	0.00	0.500
<i>Luxilus chrysocephalus</i>	7	81.56	<0.001	0.71	77.42	56.33	67.56	0.34	<0.001
<i>Luxilus coccogenis</i>	6	147.10	<0.001	0.80	73.26	75.56	74.43	0.49	<0.001
<i>Luxilus cornutus</i>	3	97.66	<0.001	0.81	0.00	100.00	85.88	0.00	0.500
<i>Lythrurus ardens</i>	1	8.95	0.030	0.87	0.00	100.00	99.43	0.00	0.500
<i>Lythrurus lirus</i>	5	43.96	<0.001	0.78	4.44	99.79	91.60	0.07	<0.001
<i>Micropterus dolomieu</i> <sup>5</sup>	6	150.58	<0.001	0.81	63.03	81.47	74.05	0.45	<0.001
<i>Micropterus punctulatus</i> <sup>5</sup>	4	101.54	<0.001	0.83	28.00	97.55	87.60	0.34	<0.001
<i>Micropterus salmoides</i> <sup>5</sup>	5	48.12	<0.001	0.73	6.02	98.64	83.97	0.07	0.003
<i>Minytrema melanops</i>	2	9.92	0.002	0.97	0.00	100.00	99.62	0.00	0.500
<i>Morone chrysops</i>	1	22.63	<0.001	0.97	0.00	100.00	99.05	0.00	0.500
<i>Moxostoma anisurum</i> <sup>4</sup>	2	28.61	<0.001	0.89	0.00	100.00	97.71	0.00	0.500
<i>Moxostoma carinatum</i> <sup>4</sup>	6	105.41	<0.001	0.94	43.33	98.38	95.23	0.49	<0.001
<i>Moxostoma duquesnei</i> <sup>4</sup>	5	122.29	<0.001	0.80	38.75	89.84	74.24	0.32	<0.001
<i>Moxostoma erythrurum</i> <sup>4</sup>	5	136.93	<0.001	0.83	33.90	9.61	15.08	0.38	<0.001
<i>Moxostoma macrolepidotum</i> <sup>4</sup>	5	138.08	<0.001	0.94	57.14	97.93	94.66	0.60	<0.001
<i>Nocomis leptocephalus</i>	1	4.31	0.116	0.88	0.00	100.00	99.62	0.00	0.500
<i>Nocomis micropogon</i> <sup>3</sup>	5	112.50	<0.001	0.76	67.87	72.36	70.23	0.40	<0.001
<i>Notemigonus crysoleucas</i>	3	3.16	0.075	0.62	0.00	100.00	98.47	0.00	0.500
<i>Notropis amblops</i>	3	262.07	<0.001	0.89	66.45	85.75	80.15	0.52	<0.001
<i>Notropis ariommus</i> <sup>3</sup>	6	61.73	<0.001	0.88	11.54	99.80	95.42	0.19	<0.001
<i>Notropis atherinoides</i>	1	18.41	<0.001	0.99	0.00	100.00	99.43	0.00	0.500
<i>Notropis buccatus</i>	1	1.60	0.206	0.60	0.00	100.00	99.24	0.00	0.500
<i>Notropis leuciodus</i> <sup>3</sup>	7	171.72	<0.001	0.82	64.53	82.55	75.57	0.48	<0.001
<i>Notropis lutipinnis</i>	3	0.81	0.369	0.80	0.00	100.00	99.24	0.00	0.500
<i>Notropis photogenis</i> <sup>3</sup>	6	79.83	<0.001	0.82	4.23	98.45	85.69	0.04	0.062
<i>Notropis rubellus</i> <sup>3</sup>	3	163.56	<0.001	0.88	47.92	94.39	85.88	0.47	<0.001

Appendix A. Continued.

Species	d.f.	l.r.	p-value	AUC	SE	SP	CC	Kappa	p-value
<i>Notropis rubricroceus</i> <sup>3</sup>	4	69.97	<0.001	0.77	5.88	96.92	79.20	0.04	0.087
<i>Notropis spectrunculus</i> <sup>3</sup>	3	29.89	<0.001	0.85	0.00	100.00	96.76	0.00	0.500
<i>Notropis stramineus</i>	5	56.80	<0.001	0.93	11.11	99.60	96.56	0.17	<0.001
<i>Notropis telescopus</i> <sup>3</sup>	5	147.62	<0.001	0.79	69.33	74.48	72.14	0.44	<0.001
<i>Notropis volucellus</i> <sup>3</sup>	7	119.21	<0.001	0.89	39.62	98.30	92.37	0.48	<0.001
<i>Noturus eleutherus</i> <sup>1</sup>	3	48.75	<0.001	0.91	0.00	100.00	95.23	0.00	0.500
<i>Noturus flavipinnis</i>	3	45.28	<0.001	0.95	0.00	99.80	96.18	0.00	0.500
<i>Noturus insignis</i>	1	32.62	<0.001	0.80	0.00	100.00	97.71	0.00	0.500
<i>Oncorhynchus mykiss</i>	4	166.94	<0.001	0.83	47.73	96.94	84.54	0.52	<0.001
<i>Perca flavescens</i>	4	4.25	0.039	0.74	0.00	100.00	99.43	0.00	0.500
<i>Percina aurantiaca</i>	6	103.92	<0.001	0.92	30.00	98.55	93.32	0.38	<0.001
<i>Percina burtoni</i> <sup>2</sup>	4	30.98	<0.001	0.83	0.00	100.00	95.42	0.00	0.500
<i>Percina caprodes</i> <sup>2</sup>	4	108.33	<0.001	0.83	32.89	98.21	88.74	0.41	<0.001
<i>Percina copelandi</i>	4	37.69	<0.001	0.98	33.33	99.81	99.05	0.44	<0.001
<i>Percina evides</i> <sup>2</sup>	6	143.46	<0.001	0.91	42.86	97.86	91.98	0.49	<0.001
<i>Percina macrocephala</i> <sup>2</sup>	2	23.38	<0.001	0.86	0.00	100.00	97.52	0.00	0.500
<i>Percina squamata</i>	4	11.75	0.003	1.00	0.00	100.00	99.81	0.00	0.500
<i>Phenacobius crassilabrum</i>	1	40.75	<0.001	0.91	0.00	99.80	97.33	0.00	0.500
<i>Phenacobius uranops</i> <sup>3</sup>	4	187.46	<0.001	0.90	49.46	95.82	87.60	0.52	<0.001
<i>Phoxinus oreas</i>	4	4.65	0.031	0.85	0.00	100.00	99.24	0.00	0.500
<i>Phoxinus tennesseensis</i>	3	15.18	0.004	0.77	0.00	100.00	96.95	0.00	0.500
<i>Pimephales notatus</i> <sup>3</sup>	3	56.35	<0.001	0.71	15.91	97.19	76.72	0.18	<0.001
<i>Pimephales promelas</i>	3	6.21	0.102	0.70	0.00	100.00	97.71	0.00	0.500
<i>Pimephales vigilax</i>	2	19.53	<0.001	0.94	0.00	100.00	99.05	0.00	0.500
<i>Pomoxis annularis</i>	6	46.82	<0.001	0.98	25.00	100.00	98.85	0.40	<0.001
<i>Pomoxis nigromaculatus</i> <sup>5</sup>	6	74.13	<0.001	0.91	14.29	98.79	94.27	0.19	<0.001
<i>Pylodictis olivaris</i> <sup>1</sup>	5	66.23	<0.001	0.89	3.70	99.60	94.66	0.06	0.013
<i>Rhinichthys atratulus</i>	4	133.86	<0.001	0.79	85.03	55.79	74.43	0.42	<0.001
<i>Rhinichthys cataractae</i>	6	54.34	<0.001	0.88	10.00	98.99	93.89	0.14	<0.001
<i>Salmo trutta</i>	3	76.42	<0.001	0.82	17.31	98.31	90.27	0.22	<0.001
<i>Salvelinus fontinalis</i>	4	89.60	<0.001	0.98	52.94	99.01	97.52	0.57	<0.001
<i>Semotilus atromaculatus</i> <sup>3</sup>	3	65.96	<0.001	0.72	37.31	83.38	66.41	0.22	<0.001
<i>Stizostedion vitreum</i>	4	56.98	<0.001	0.93	75.00	100.00	99.62	0.46	<0.001

Appendix B. Statistics describing the performance of models that predict species occurrence gleaned from peer-reviewed publications. Predictions were made for > 140 species and several taxonomic groups. Some statistics describe mean prediction performance for > 1 species. N is the number of samples reported for that study. Area under the receiver operator characteristic curve (AUC), sensitivity (SE), specificity (SP), correct classification (CC), and Kappa statistics for models assessed using resubstitution, cross-validation, and validation on an independent dataset are presented. Model methods include multiple logistic regression (MLR), generalized regression analysis and spatial prediction (GRASP), multivariate discriminant analysis (MDA), artificial neural networks (ANN), classification tree (TRE), multivariate adaptive regression splines (MARS), Bayesian-MLR model (BYS), and rule-based models (RB). Entries are blank where the appropriate statistics were not published. Authors are identified by superscripts.

Group	N	Method	Assessment	AUC	SE	SP	CC	Kappa
Molluscs <sup>a</sup>	545	MLR	resubstitution		43.9	92.4	87.5	
Molluscs <sup>a</sup>	541	MLR	validation		36.9	89.7	84.8	0.29
Bird <sup>b</sup>	1,550	MLR	validation	0.93				
Bird <sup>b</sup>	1,550	MLR	validation	0.94				
Bird <sup>b</sup>	1,550	MLR	validation	0.57				
Bird <sup>b</sup>	1,550	MLR	validation	0.91				
Bird <sup>b</sup>	1,550	MLR	validation	0.76				
Bird <sup>b</sup>	1,550	MLR	validation	0.79				
Bird <sup>b</sup>	1,550	MLR	validation	0.91				
Bird <sup>b</sup>	1,550	MLR	validation	0.79				
Bird <sup>b</sup>	1,550	MLR	validation	0.91				
Bird <sup>b</sup>	1,550	MLR	validation	0.68				
Bird <sup>b</sup>	1,550	MLR	validation	0.86				
Bird <sup>b</sup>	1,550	MLR	validation	0.78				
Bird <sup>b</sup>	1,550	MLR	validation	0.75				
Bird <sup>b</sup>	1,550	MLR	validation	0.89				
Bird <sup>b</sup>	1,550	MLR	validation	0.81				
Bird <sup>b</sup>	1,550	MLR	validation	0.68				
Bird <sup>b</sup>	1,550	MLR	validation	0.78				
Bird <sup>b</sup>	1,550	MLR	validation	0.84				
Bird <sup>b</sup>	1,550	MLR	validation	0.84				
Bird <sup>b</sup>	1,550	MLR	validation	0.81				
Bird <sup>b</sup>	1,550	MLR	validation	0.83				
Bird <sup>b</sup>	1,550	MLR	validation	0.79				
Bird <sup>b</sup>	1,550	MLR	validation	0.80				
Bird <sup>b</sup>	1,550	MLR	validation	0.90				
Bird <sup>b</sup>	1,550	MLR	validation	0.82				
Bird <sup>b</sup>	1,550	MLR	validation	0.81				
Bird <sup>b</sup>	1,550	MLR	validation	0.81				

Appendix B. Continued.

Group	N	Method	Assessment	AUC	SE	SP	CC	Kappa
Bird <sup>b</sup>	1,550	MLR	validation	0.90				
Bird <sup>b</sup>	1,550	MLR	validation	0.80				
Bird <sup>b</sup>	1,550	MLR	validation	0.85				
Fish <sup>c</sup>	306	MLR	resubstitution		84.8	72.7	78.2	
Fish <sup>c</sup>	306	MLR	resubstitution		63.3	77.8	72.0	
Fish <sup>c</sup>	306	MLR	resubstitution		75.9	74.4	75.0	
Fish <sup>c</sup>	306	MLR	resubstitution		80.0	84.2	83.3	
Fish <sup>c</sup>	306	MLR	resubstitution		84.6	70.0	79.0	
Fish <sup>c</sup>	306	MLR	resubstitution		81.7	63.9	75.0	
Reptile <sup>d</sup>	4,138	MLR	cross-validation	0.75				
Reptile <sup>d</sup>	4,138	MLR	validation	0.78				
Plant <sup>e</sup>	19,875	GRASP	cross-validation	0.94				
Plant <sup>e</sup>	19,875	GRASP	validation	0.94				
Bird <sup>f</sup>	180	MDA	cross-validation		41.0	84.0	86.0	0.27
Bird <sup>f</sup>	180	MLR	cross-validation		41.0	92.0	82.0	0.36
Bird <sup>f</sup>	180	ANN	cross-validation		31.0	84.0	94.0	0.15
Plant <sup>g</sup>	1,285	MLR	cross-validation	0.96				
Plant <sup>g</sup>	1,285	TRE	cross-validation	0.97				
Plant <sup>g</sup>	1,285	MARS	cross-validation	0.98				
Plant <sup>g</sup>	103,181	MLR	cross-validation	0.79				
Plant <sup>g</sup>	103,181	TRE	cross-validation	0.98				
Plant <sup>g</sup>	103,181	MARS	cross-validation	0.90				
Plant <sup>g</sup>	1,285	MLR	validation	0.95				
Plant <sup>g</sup>	1,285	TRE	validation	0.91				
Plant <sup>g</sup>	1,285	MARS	validation	0.93				
Plant <sup>g</sup>	103,181	MLR	validation	0.78				
Plant <sup>g</sup>	103,181	TRE	validation	0.95				
Plant <sup>g</sup>	103,181	MARS	validation	0.91				
Fish <sup>h</sup>	128	ANN	cross-validation		41.7	82.5	67.2	2.58
Fish <sup>h</sup>	128	ANN	cross-validation		64.3	68.1	66.4	3.60
Fish <sup>h</sup>	128	ANN	cross-validation		95.2	9.1	65.6	0.46
Fish <sup>h</sup>	128	ANN	cross-validation		35.6	90.7	64.8	1.70
Fish <sup>h</sup>	128	ANN	cross-validation		70.9	78.1	75.0	5.18

Appendix B. Continued.

Group	N	Method	Assessment	AUC	SE	SP	CC	Kappa
Fish <sup>h</sup>	128	ANN	cross-validation		58.8	63.3	60.9	2.47
Fish <sup>h</sup>	128	ANN	cross-validation		94.8	29.4	68.8	1.44
Fish <sup>h</sup>	128	ANN	cross-validation		50.0	90.6	80.5	3.62
Fish <sup>h</sup>	128	ANN	cross-validation		93.1	29.3	72.7	2.28
Fish <sup>h</sup>	128	ANN	cross-validation		71.7	54.5	69.1	2.59
Fish <sup>h</sup>	128	ANN	cross-validation		23.3	26.5	6.0	1.31
Fish <sup>h</sup>	32	ANN	validation		90.5	18.2	65.6	0.45
Fish <sup>h</sup>	32	ANN	validation		90.0	86.7	84.5	3.88
Fish <sup>h</sup>	32	ANN	validation		41.7	58.3	78.1	2.58
Fish <sup>h</sup>	32	ANN	validation		88.9	75.0	62.5	0.89
Fish <sup>h</sup>	32	ANN	validation		55.6	64.3	78.1	2.96
Fish <sup>h</sup>	32	ANN	validation		90.5	78.3	71.9	1.57
Fish <sup>h</sup>	32	ANN	validation		53.3	18.2	65.6	0.45
Fish <sup>h</sup>	32	ANN	validation		95.8	88.2	71.9	2.34
Fish <sup>h</sup>	32	ANN	validation		80.4	37.5	81.3	1.53
Fish <sup>h</sup>	32	ANN	validation		18.7	54.4	73.3	1.85
Fish <sup>h</sup>	32	ANN	validation		18.7	25.1	7.7	1.11
Mammal <sup>i</sup>	349	MLR	resubstitution				80.1	
Bird <sup>j</sup>	71	MLR	resubstitution	0.90				
Bird <sup>j</sup>	92	MLR	resubstitution	0.93				
Bird <sup>j</sup>	92	MLR	resubstitution	0.93				
Bird <sup>j</sup>	92	BYS	resubstitution	0.97				
Plant <sup>k</sup>	31	RB	resubstitution		100.0	97.0	97.0	0.87
Plant <sup>k</sup>	31	RB	resubstitution		50.0	88.0	79.0	0.42
Plant <sup>k</sup>	31	RB	resubstitution		100.0	36.0	58.0	0.28
Plant <sup>k</sup>	31	RB	resubstitution		65.0	69.0	67.0	0.33
Plant <sup>k</sup>	31	RB	resubstitution		67.0	73.0	70.0	0.40
Plant <sup>k</sup>	31	RB	resubstitution		67.0	73.0	70.0	0.40
Plant <sup>k</sup>	31	RB	resubstitution		78.0	40.0	61.0	0.18
Plant <sup>k</sup>	31	RB	resubstitution		84.0	36.0	64.0	0.21
Plant <sup>k</sup>	31	RB	resubstitution		76.0	25.0	58.0	0.01
Plant <sup>k</sup>	31	RB	resubstitution		78.0	30.0	64.0	0.09
Insect <sup>k</sup>	85	RB	resubstitution		73.0	81.0	80.0	0.38
Insect <sup>k</sup>	85	RB	resubstitution		39.0	97.0	88.0	0.44
Insect <sup>k</sup>	85	RB	resubstitution		25.0	93.0	74.0	0.22



Appendix B. Continued.

Group	N	Method	Assessment	AUC	SE	SP	CC	Kappa
Insect <sup>k</sup>	85	RB	resubstitution		80.0	88.0	86.0	0.67
Insect <sup>k</sup>	85	RB	resubstitution		60.0	82.0	75.0	0.41
Insect <sup>k</sup>	85	RB	resubstitution		69.0	82.0	77.0	0.51
Insect <sup>k</sup>	85	RB	resubstitution		52.0	78.0	65.0	0.30
Insect <sup>k</sup>	83	RB	resubstitution		61.0	94.0	88.0	0.61
Insect <sup>k</sup>	83	RB	resubstitution		50.0	86.0	78.0	0.38
Insect <sup>k</sup>	83	RB	resubstitution		49.0	79.0	67.0	0.29
Insect <sup>k</sup>	83	RB	resubstitution		87.0	54.0	72.0	0.42
Insect <sup>k</sup>	83	RB	resubstitution		87.0	58.0	78.0	0.48
Insect <sup>k</sup>	83	RB	resubstitution		98.0	5.0	80.0	0.05
Insect <sup>k</sup>	83	RB	resubstitution		96.0	7.0	80.0	0.04
Insect <sup>k</sup>	83	RB	resubstitution		91.0	8.0	78.0	0.00
Bird <sup>k</sup>	53	RB	resubstitution		8.0	96.0	89.0	0.04
Bird <sup>k</sup>	53	RB	resubstitution		9.0	96.0	92.0	0.05
Bird <sup>k</sup>	53	RB	resubstitution		33.0	92.0	90.0	0.01
Bird <sup>k</sup>	53	RB	resubstitution		22.0	84.0	78.0	0.05
Bird <sup>k</sup>	53	RB	resubstitution		38.0	81.0	76.0	0.12
Bird <sup>k</sup>	53	RB	resubstitution		50.0	64.0	59.0	0.14
Bird <sup>k</sup>	53	RB	resubstitution		54.0	66.0	61.0	0.19
Bird <sup>k</sup>	53	RB	resubstitution		50.0	63.0	59.0	0.12
Bird <sup>k</sup>	53	RB	resubstitution		45.0	64.0	58.0	0.09
Fish <sup>l</sup>	48	MLR	resubstitution		72.0	87.0	81.0	
Fish <sup>l</sup>	48	MLR	resubstitution		68.0	76.0	73.0	
Fish <sup>l</sup>	48	MLR	resubstitution		53.0	86.0	73.0	
Fish <sup>l</sup>	48	MLR	resubstitution		79.0	88.0	85.0	
Fish <sup>l</sup>	48	MLR	resubstitution		86.0	77.0	81.0	
Fish <sup>l</sup>	48	MLR	resubstitution		79.0	94.0	90.0	
Fish <sup>l</sup>	48	MLR	resubstitution		63.0	83.0	75.0	
Fish <sup>l</sup>	48	MLR	resubstitution		96.0	86.0	92.0	
Fish <sup>l</sup>	48	MLR	resubstitution		95.0	93.0	94.0	
Fish <sup>l</sup>	48	MLR	resubstitution		85.0	83.0	83.0	
Fish <sup>l</sup>	48	MLR	resubstitution		94.0	95.0	94.0	
Fish <sup>l</sup>	48	MLR	resubstitution		98.0	43.0	90.0	
Fish <sup>l</sup>	48	MLR	resubstitution		86.0	89.0	88.0	
Fish <sup>l</sup>	94	MLR	validation				70.0	
Fish <sup>l</sup>	94	MLR	validation				73	

Appendix B. Continued.

Group	N	Method	Assessment	AUC	SE	SP	CC	Kappa
Fish <sup>l</sup>	94	MLR	validation				67	
Fish <sup>l</sup>	94	MLR	validation				53.0	
Fish <sup>l</sup>	94	MLR	validation				73.0	
Fish <sup>l</sup>	94	MLR	validation				65.0	
Fish <sup>l</sup>	94	MLR	validation				72.0	
Plants <sup>m</sup>	14,112	MLR	validation	0.87				
Plants <sup>m</sup>	14,112	MLR	validation	0.77				
Plants <sup>m</sup>	14,112	MLR	validation	0.67				
Plants <sup>m</sup>	14,112	MLR	validation	0.62				
Plants <sup>m</sup>	14,112	MLR	validation	0.89				
Plants <sup>m</sup>	14,112	MLR	validation	0.94				
Plants <sup>m</sup>	14,112	MLR	validation	0.85				

Berg and Proschwitz 2004<sup>a</sup>  
 Brotons, et al. 2004<sup>b</sup>.  
 Filipe, et al. 2002<sup>c</sup>  
 Guisan and Hofer 2003<sup>d</sup>  
 Lehmann et al. 2002<sup>e</sup>  
 Manel et al. 1999a<sup>f</sup>  
 Muñoz and Felicísimo 2004<sup>g</sup>  
 Olden and Jackson 2002<sup>h</sup>  
 Orrock et al. 2000<sup>i</sup>  
 Osborne et al. 2001<sup>j</sup>  
 Petit et al. 2003<sup>k</sup>  
 Porter et al. 2000<sup>l</sup>  
 Van den Berg et al. 2003<sup>m</sup>

## General Conclusions

### Sampling Sufficiency and the Use of Predictions of Species Occurrence for Conservation

Limited resources for conservation require efficient conservation – directing actions to where they are most needed and most likely to succeed. The paucity of knowledge of the geographic distribution of stream fish limits the efficiency of their conservation and management. I found that insufficient sampling of stream fish assemblages in Virginia is common. The effort required to sufficiently sample a region varies with the survey objectives, analytical methods, sampling design, and the sufficiency of individual samples. The geographic distribution of species also affects the number of samples required. Although many samples have been collected in most regions of Virginia, additional sampling will likely collect additional species in many regions. The number of samples required to sufficiently sample most regions is large, but well planned and organized sampling and reducing errors of sampling omission could decrease the number of samples required to accurately describe the species composition of regions of Virginia.

Models are often used to predict the occurrence of species in lieu of sampling data that is sufficient to inform conservation and management decisions. In the absence of data from well-planned and organized sampling, multipurpose species occurrence data are often used to predict species occurrence. Many methods are used to improve the prediction accuracy, generality, and resolution from models that use multipurpose data. I found that several modeling methods provide useful predictions of species occurrence, even for rare species, for which management and conservation are most urgent and knowledge of occurrence is least available. The accuracy and resolution of predictions of species occurrence differ among species and are low for most species. Additional occurrence data might improve the accuracy and resolution of models that predict the occurrence of rare species, but increasing model complexity is unlikely to substantially improve model prediction accuracy or resolution for most species. Substantial improvement of model prediction accuracy and resolution may not be possible, suggesting that protocols that facilitate the appropriate application of model predictions are needed.

### Management Implications

My work suggests that insufficient knowledge of species occurrence is a common problem for the conservation and management of stream fish and likely a problem for other taxonomic groups. The large number of historical samples of the stream fish assemblages in Virginia does not ensure sufficient sampling to describe species occurrences at the spatial extents

at which conservation and management decisions are made. Given the importance of data that describe the geographic distribution of species to conservation and management, implementing well-planned and organized surveys across large spatial extents should be a high priority. Such survey data would facilitate more efficient conservation and management and might prove useful for improving the accuracy and resolution of model predictions of species occurrence.

My work also suggests that many models that predict species occurrence lack sufficient prediction accuracy and resolution to inform many conservation and management decisions. Improving the accuracy and resolution of model predictions is a worthy goal, but scientists and managers must also better understand the limitations of predictive models and learn to apply model predictions more appropriately. Future studies should quantify the limitations of model prediction accuracy and resolution and describe specific implications for conservation and management. Exploration of the accuracy and resolution limits of models that use temporally stable predictors of species occurrence, based on ecological theory is also warranted. For example, biological interaction and disturbance mechanisms as well as environmental conditions affect species occurrence. Quantifying the relative importance of these mechanisms among different assemblages and ecosystems warrants further study because it has profound implications for predictive modeling, conservation, and management.