Assessing predictive performance and transferability of species distribution models for freshwater fish in the United States

Jian Huang

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

Doctor of Philosophy
In
Fisheries and Wildlife

Emmanuel A. Frimpong
Donald J. Orth
Yan Jiao
Jie Li

April 17, 2015
Blacksburg, Virginia

Keywords: Species distribution models, discrimination power, calibration, spatial transferability, temporal transferability, climate change, machine learning, New River, stream fish

Assessing predictive performance and transferability of species distribution models for freshwater fish in the United States

Jian Huang

Abstract

Rigorous modeling of the spatial species distributions is critical in biogeography, conservation, resource management, and assessment of climate change. The goal of chapter 2 of this dissertation was to evaluate the potential of using historical samples to develop high-resolution species distribution models (SDMs) of stream fishes of the United States. I explored the spatial transferability and temporal transferability of stream–fish distribution models in chapter 3 and chapter 4 respectively. Chapter 2 showed that the discrimination power of SDMs for 76 non-game fish species depended on data quality, species' rarity, statistical modeling technique, and incorporation of spatial autocorrelation. The area under the Receiver-Operating-Characteristic curve (AUC) in the cross validation tended to be higher in the logistic regression and boosted regression trees (BRT) than the presence-only MaxEnt models. AUC in the cross validation was also higher for species with large geographic ranges and small local populations. Species prevalence affected discrimination power in the model training but not in the validation. In chapter 3, spatial transferability of SDMs was low for over 70% of the 21 species examined. Only 24% of logistic regression, 12% of BRT, and 16% of MaxEnt had AUC > 0.6 in the spatial transfers. Friedman's rank sum test showed that there was no significant difference in the performance of the three modeling techniques. Spatial transferability could be improved by using spatial logistic regression under Lasso regularization in the training of SDMs and by matching the range and location of predictor variables between training and transfer regions. In chapter 4, testing of temporal SDM transfer on independent samples resulted in discrimination power of the moderate to good range, with AUC > 0.6 for 80% of species in all three types of models. Most cool water species had good temporal transferability. However, biases and misspecified spread occurred frequently in the temporal model transfers. To reduce under- or over-estimation bias, I suggest rescaling the predicted probability of species presence to ordinal ranks. To mitigate inappropriate spread of predictions in the climate change scenarios, I recommended to use large training datasets with good coverage of environmental gradients, and fine-tune predictor variables with regularization and cross validation.

Acknowledgments

A PhD is a long, tough but wonderful journey that no one can proceed without the support and fellowship of professional and personal network. First, I want to express my sincere gratitude to my advisor and Committee Chair Dr. Emmanuel A. Frimpong. I have tremendous respect for Dr. Frimpong for his knowledge, wisdom, passion for research, and support for students. I appreciate his guidance in my research projects, his suggestions in my study of natural resources and statistics, his patience in answering my questions and revising my manuscripts, and his support for me to attend international conferences. My graduate committee provided great suggestions for me to develop ideas in this dissertation and helped me think critically by posing challenging questions. I am grateful to members of my graduate committee, Dr. Donald Orth (Department of Fish and Wildlife Conservation), Dr. Yan Jiao (Department of Fish and Wildlife Conservation) and Dr. Jie Li (Department of Statistics), for their time, insights and assistance in my study, research and career. They helped develop me into the quantitative ecologist that I would like to be.

This work was made possible by the help of many students and staff at Virginia Tech, all of whom I owe a debt of gratitude. My fellow graduate students and friends, Brandon Peoples, Stephen Floyd, Joe Buckwalter, Steve Watkins, helped me with field work, taught me skills in fish sampling and identification, shared food with me on the road trips, and cared about me when I had trouble in water, for which I am indebted. I also wish to thank venerable field technicians Kaylie Fitzgerald, Caitlin Worsham, Christina Bolton, Paige Crane, Charles Olinger, Tommy Russell, and June Shrestha for their assistance with field data collections in the New River during 2012 -2014. My praise is also due to Liang Lu, Kaylie Fitzgerald, Pascaline Okongo, Alicia Mercer, Caitlin Worsham, Joe Buckwalter, and Iris Fynn for their hard work on processing thousands of fish distribution maps for the *IchthyMaps* project. Hai-lung Cheng, Huiquan Jiang, Yaw Ansah, Bonnie Myers and Dan Hua helped me in various ways in my work plan and dissertation. I also am indebted to two kind ladies Dana Keith and Terri Waid in VT-FWC for their help with administrative issues. I also thank my family for their unconditional love and encouragement throughout my academic pursuits and all my friends for their support.

Attribution

My academic advisor Dr. Emmanuel Frimpong contributed significantly to this dissertation. He co-authored all three chapters. He wrote proposals to get funds from U.S. Geological Survey, and contributed vitally in the project scheme, sampling design, statistical analysis, and revisions of manuscripts. Dr. Donald Orth provided great suggestions to improve chapter 4. Chapter 2 was submitted to PLOS ONE in January 2015 for review. Chapter 3 is to be submitted to Ecography for publication. Chapter 4 is to be submitted to the Global Ecology and Biogeography.

Table of Contents

List of Figures

List of Tables

Chapter 1: General introduction

Rigorous modeling of species-habitat relationships and the spatial pattern of species distributions is critical in biodiversity conservation and resource management (Franklin and Miller 2009). Species distribution models (SDMs) include a series of statistical and related methods used with mapped biological and environmental data to model spatial distribution of species or other biogeographic patterns (Figure 1.1). The general steps to develop SDMs include: (1) constructing a conceptual model, (2) selecting modeling approach, (3) evaluating and calibrating the model, and (4) making predictions based on the fine-tuned model. SDMs are nearly the same as habitat suitability models (HSMs) except that the HSMs output the measures of habitat suitability without separating suitable and occupied unit from suitable but unoccupied units (Franklin and Miller 2009). SDMs could be used as a tool to reintroduce and recover declining or extirpated species (e.g., Pearce and Lindenmayer 2000, Hirzel et al. 2004), designate critical habitats and reserves for species as legally mandated by environmental legislations (Fielding and Haworth 1995, Sindt et al. 2012), predict and mitigate the effect of potential climate and landscape changes on economic or threatened species (Chu et al. 2005, Lyons et al. 2010, Comte and Grenouillet 2013), and to control biological invasions (Wang and Jackson 2014).

However, gathering species occurrences is usually time- and budget-constrained, especially for rare and endangered freshwater fish species (Franklin and Miller 2009, Elith and Leathwick 2009). Lacking reliable species absence, most SDMs studies have been constrained to the presence-only Maximum-Entropy Species-Distribution Modeling (Phillips et al. 2006) or Genetic Algorithm for Rule-set Production or GARP (Stockwell and Peters 2002). However, presence-only models lack the ability to estimate species prevalence (i.e., the proportion of

species presence in all occurrence observations) and to evaluate model performance via commonly-used validation approaches (Brotons et al. 2004, Franklin and Miller 2009, Elith et al. 2011).

Samplings for stream fishes, non-game species particularly, have been predominantly community-based, making it promising to apply the meta-community theories to solve the problem of data scarcity in SDMs. Communities in a defined metacommunity are assumed spatially connected by migrating species (Gilpin and Hanski 1991, Wilson 1992), and compositions of local communities are presumably determined by the regional species pool and regulated by local environmental factors (Leibold et al. 2004, Niu et al. 2012). The basic idea of developing a metacommunity database is collating fish historical occurrences from different sources (e.g., published atlases, databases of state agencies) and deriving absences for one species from known presences of other species (Figure 1.2). The probability of presence for each species observed in a unit =1 contains significant information about the probability of absence of other unobserved species in the community-based fish sampling (Figure 1.2). Game species that are mostly sampled in target-species surveys should be excluded from inferring the absence of other species. Inferred absence of a species from historical occurrences of multiple other species in multiple sampling units are more reliable than pseudo absences randomly sampled from the background which are primarily unsampled areas, given no information on sampling effort or detection probability of the species. It is worth noting that the commonly used presence-only MaxEnt model randomly samples pseudo absences from the 'background' (i.e., all units except for those with confirmed presence of a particular species) by default (Phillips et al. 2006). One of the main hypotheses tested in this dissertation was that presence-absence models based on presences and inferred absences in the metacommunity database would outperform presence-

only models in terms of discrimination power (e.g., the area under the Receiver-Operating-Characteristic curve).

Historical occurrences of most freshwater fishes in river segments are documented in national and state atlases (e.g., Lee at al. 1980, Jenkins and Burkhead 1994). Additionally the Multi-State Aquatic Resources Information System (MARIS, www.marisdata.org) and FishNet2 (www.fishNet2.org) have also compiled electronic fish collections for a number of states. Advances in GIS technology allows for converting paper maps to electronic maps accurately and efficiently. After synthesizing data from different sources, one could construct a metacommunity matrix of species presences over stream segments (Figure 1.2), and then infer absence of one species by the presence of other (non-game) species. Essentially, the inferred absence will be assigned to stream segments where the species has not been recorded during the time period defined for the metacommunity. The species presences and absences used to develop species distribution models in this study were derived from the *IchthyMaps* database. This database contains 606,550 fish presence records sampled predominantly during 1950-1990. These records were published in atlases of freshwater fishes throughout the conterminous United States. The steps to develop metacommunity database from atlas data were: 1) scanning each atlases of species presences, 2) geo-referencing digitized atlases, 3) extracting presence records with GIS (geographic information system) tools, 4) joining the presence records to stream networks, and 5) integrating presence records to form the metacommunity database. To infer absences for units in a given region, one needs to first sample the presences of the species in the region from the metacommunity database. The sampled units with no presence record of a focal species but at least a presence of any other non-game species were designated as absences for that species.

Game fish are usually sampled in targeted species surveys so they were not used to infer the absence of other species.

Presence-absence models are recommended as long as the absence data are available (Brotons 2004). A series of presence-absence models such as logistic regression (Gumbel 1961) and boosted regression trees (Friedman 2001) can be developed by using presences and absences derived from the metacommunity database. Classification tree approaches has been increasingly used in SDMs of stream fish when absence data are available (e.g., Steen et al. 2008, Lyons et al. 2010, Bond et al. 2011). The high popularity of tree-based models is largely due to their capability of handling nonlinear ecological response-environment relationship (De'ath and Fabricius 2000) and complex interaction among the predictor variables without being plagued by multicollinearity (Breiman 2001). Nevertheless, one would arbitrarily 'dump' whatever predictors are available into the "black box" of machine learning and end up with a set of environmental predictors of which the ecological associations with the biological response might be difficult to interpret.

The types of environmental predictors selected also vary considerably among fish species distribution studies. Climatic, hydrological and geomorphological factors are the three most popular categories of predictor variables in the SDMs of stream fish (Chu et al. 2005, Leathwick et al. 2005, Mugodo et al. 2006, Chen et al. 2007, Lassalle et al. 2008, Steen et al. 2008, Lyons et al. 2010, Steen et al. 2010, Bond et al. 2011, Grenouillet et al. 2011, Sindt et al. 2012, Zarkami et al. 2012, Yu et al. 2013). The number and scale of predictors in each category (i.e., climate, landscape and disturbance, hydrology, geology, geomorphology, riparian zone, and water chemistry) are not consistent. These variations in the conceptual framework of SDMs render it difficult to validate and compare studies in different regions or years. Additionally, in a few

studies, only the predictors related to the targeted disturbance (e.g., climate change, landscape modification) are considered in their SDMs (e.g., Chu et al. 2005, Yu et al. 2013). Araújo and Guisan (2006) argued that it might be valid to use solely climate predictors while ignoring local environmental heterogeneity in the large-extent SDM studies. However, much more stringent evidence is needed to support this idea in assessing the change in distribution of fish species at the regional scale (e.g., intermediate size watershed). Temperature might be a key constraint, but unlikely to be the only constraint, to the dispersal, survival, growth and reproduction for some fish species (e.g., cold-water species). It is also possible that the changed temperature would not exceed the thermal tolerance for the focal species, which results in no conspicuous distribution shift, but indeed the distribution would be affected by other disturbances such as altered flow regime or landscape modification. A range map delineated based solely on a single type of predictor likely mismatches the true suitable habitat for the focal species. Performance of SDMs can be improved if we can identify a complete and consistent set of informative environmental predictors for stream fish species.

Resolution (i.e., study unit grain) also varies among studies of stream fish distribution. The NHD (National Hydrography Dataset) inter-confluence segment has predominantly been chosen as the study unit in species distribution models for stream fish in the United States (e.g., Steen et al. 2008, Lyons et al. 2010, Steen et al. 2010), largely owing to the intensive collation of environmental variables by the US EPA (Environmental Protection Agency) and USGS (US Geological Survey) and now the NFHAP (National Fish Habitat Action Plan) program. The study unit in studies of fish SDMs in other nations (Chu et al. 2005, Leathwick et al. 2005, Mugodo et al. 2006, Lassalle et al. 2008, Grenouillet et al. 2011, Zarkami et al. 2012, Yu et al. 2013) varied from occurrence points, stream reach, inter-confluence stream segment, to large

river basins. This happens largely because the data used in these studies are observational and opportunistic, rather than based on robust sampling designs. Another point worth noting is that terminology (e.g., defining stream segment, reach and section) seems inconsistent among these studies, which adds to difficulty in interpreting and comparing results among studies. In this study, stream segment is used as the basic unit and a stream segment is defined as the stream between two adjacent confluences or a first-order stream (Frissell et al. 1986).

Deriving absences from metacommunity database also enables us to incorporate spatial autocorrelation in the SDMs. Map-based species or trait distribution often exhibit spatial autocorrelation (i.e., the closer the study units, the more the similarity of habitat and biological process or response). It is widely suggested to explicate spatial autocorrelation and association with environmental predictors in modeling species distribution (Magalhães et al. 2002, Diez and Pulliam 2007, Dormann et al. 2007) and assembling patterns (e.g., Stewart-Koster et al. 2007). Violating the assumption of independent and identically-distributed residuals in the generalized linear models would yield biased parameter estimation and misspecified organism-habitat relationships (Dormann et al 2007). Spatial autocorrelation has been explicated differently, depending on the study objective, data quality, focal species, and source of spatial dependence. Three techniques are often used to incorporate spatial autocorrelation in the SDMs: 1) Zeros of non-diagonal elements in the covariance matrix of a GLM are replaced by among-site distance measures to account for the spatial dependence of response variable and residuals, 2) Utilizing the spatial autocorrelation, conditional autoregressive model (CAR) and simultaneous autoregressive model (SAR) respectively incorporate a response-related and residual-related neighborhood matrices in the GLM framework (Dormann et al. 2007). 3) Another strategy recently proposed is extracting spatial components (i.e., eigenvectors) out of topology matrices.

The spatial eigenvectors in turn can be flexibly used as predictors akin to environmental variables (Griffith and Peres-Neto 2006). This approach, spatial eigenvector mapping, is found robust in model simulations and outstanding in comparative studies (Dormann et al. 2007) and is promising in non-parametric models.

The performances of SDMs are usually measured in terms of discrimination power (a model's ability to discriminate occupied and unoccupied sites) and calibration (a model's reliability in predicting the probability of a site being occupied). Common measures of discrimination power include sensitivity (true positive or presence rate), specificity (true negative or absence rate), accuracy (correct classification rate), and AUC (area under the Receiver-Operating-Characteristic (ROC) curve). An ROC curve is a plot of sensitivity against 1− specificity with varying discrimination thresholds for the predicted probability of presence. AUC averages diagnostic accuracy across the whole range of thresholds of the probability of presence (Hanley and McNeil 1982). In model calibration, systematic under-estimation or over-estimation bias, and spread of predicted probability of presence are examined (Miller et al. 1991). In a well calibrated model, if the predicted probabilities of species presence was 0.4, then approximately 40 out of 100 sites would actually be occupied by the species. The discrimination power and calibration can be evaluated in the model fit, k-fold cross validation, and external evaluation. In the model fit, predicted probability of species presences are calculated from the training data on which the species distribution models are based. In the k-fold cross validation, the whole dataset is randomly partitioned into k subsets. A total of k sub-models are built in the k-fold cross validation. In each of k rounds, the k^{th} subset of data is used as testing data and the rest k-1 subsets are used as training data to develop a sub-model. Performance measures of a sub-model are computed based on the observed values and predicted values of testing data. In the external

validation, predictions are made on the independent data and performance measures are computed by contrasting these predictions with the observed values.

Predicting the probability of species occurrence in poorly-sampled areas or under future climatic conditions is an appealing function of SDMs. Yet, such transfers of SDMs are risky because of habitat heterogeneity and stochastic temporal dynamics. However, the assessments of model transferability have been neglected in many studies. The spatial and temporal transferability of SDMs could be influenced by species trait (Bulluck et al. 2006, Randin et al. 2006), data quality (Strauss and Biedermann 2007, Barbosa et al. 2009, Wang and Jackson 2014), modeling approach (Meynard and Quinn 2007, Peterson et al. 2007, Wenger and Olden 2012) and variable selection (Strauss and Biedermann 2007, Wenger and Olden 2012).

In this dissertation, I assessed predictive performance and transferability of species distribution models for freshwater fishes based on historical atlas data in the United States. In chapter 2, I compared presence-only MaxEnt model and presence-absence models (i.e., boosted regression trees (BRT) and logistic regression), where the absence of a species was inferred from accumulated evidence of the presence of other non-game fish species. Additionally, I investigated how discrimination power of SDMs is affected by data resolution, species' rarity, observed prevalence, and spatial autocorrelation. I assessed the spatial transferability and temporal transferability of species distribution models for stream fishes respectively in chapter 3 and chapter 4. I used the Friedman's rank sum test to compare the transferability of three widely used modelling techniques (i.e., logistic regression, BRT, and MaxEnt). Discrimination power was used as the primary measure of transferability in both chapters. Specifically in chapter 3, I also measured the cross–basin consistency of variable selections in the boosted regression trees, and compared the fish–habitat relationships described in different basins using partial

dependence plots. Specifically in chapter 4, I extended the evaluation of transferability by means of calibration.

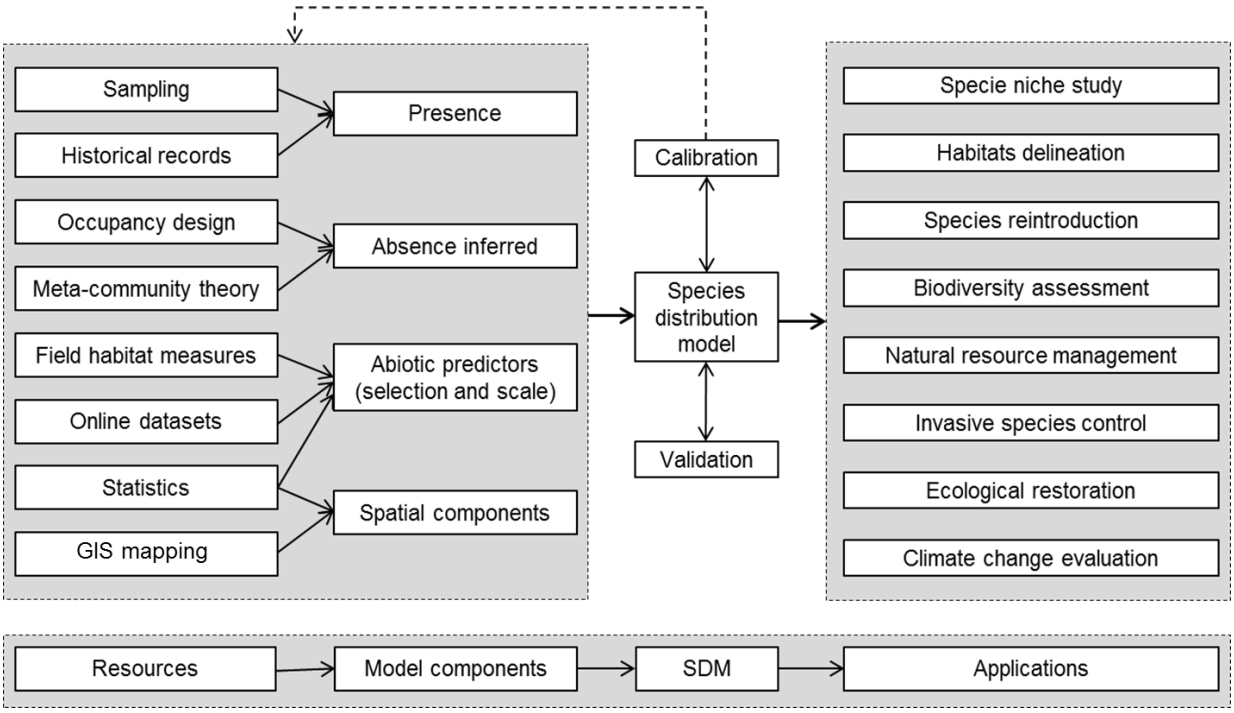
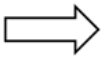


Figure 1.1. A framework of developing species distribution models and its application.

Species	Unit	Type
S1	U1	N
S1	U2	N
S1	U3	N
S2	U2	N
S2	U3	N
S3	U3	G
S3	U4	G
S3	U5	G
S4	U5	N
S5	U1	N
S6	U1	G
S6	U2	G
S6	U4	G
S6	U5	G
S7	U1	N
S7	U2	N
S7	U5	N
S8	U1	N
S8	U2	N
S8	U5	N

Matrix A



Unit	Species							
	S1	S2	S3	S4	S5	S6	S7	S8
U1	1	0	0	0	1	1	1	1
U2	1	1	0	0	0	1	1	1
U3	1	1	1	0	0	0	0	0
U4			1			1		
U5	0	0	1	1	0	1	1	1

Matrix B

Figure 1.2. An illustration of inferring absences based on historical fish presence records. Each row of Matrix A represents a presence record of a fish species in a unit. Unit in Matrix A can be NHD (National Hydrography Dataset) inter-confluence stream segment or coarser unit. Type N and G distinguish non-game and game species. The 1's in the Matrix B are fish presences, and 0's are fish absences inferred from presences of one or more other non-game species. In the example, only two game species (S3 and S6) were found in unit U4, which were not informative to infer absence of non-game species, so the cells at unit U4 for all non-game species (S1, S2, S4, S5, S7, and S8) were left blank.

Chapter 2: The potential to use historical atlas data to develop high-resolution distribution models of freshwater fishes of the United States

Jian Huang, Emmanuel A. Frimpong*

Department of Fish and Wildlife Conservation, Virginia Polytechnic Institute and State University, 100 Cheatham Hall, Blacksburg, VA 24061, USA

* Corresponding author; E-mail: frimp@vt.edu; Tel.: +1-540-231-6880; Fax: +1-540-231- 7580

Abstract: The availability of large amounts of historical presence records for freshwater fishes of the United States provides an opportunity for deriving reliable absences from data reported as presence-only, when sampling was predominantly community-based. In this study, we used boosted regression trees (BRT), logistic regression, and MaxEnt models to assess the performance of a historical metacommunity database with inferred absences, for modeling fish distributions, investigating the effect of model choice and data properties thereby. With models of the distribution of 76 native, non-game fish species of varied traits and rarity attributes in four river basins, we show that model accuracy depends on data quality (e.g., sample size, location precision), species' rarity, statistical modeling technique, and consideration of spatial autocorrelation. The cross-validation area under the Receiver-Operating-Characteristic curve (AUC) tended to be high in the spatial presence-absence models at the highest level of resolution for species with large geographic ranges and small local populations. Prevalence affected training but not validation AUC. The fish-habitat relationships were evaluated through partial dependence plots. The community-based SDM framework broadens our capability to model species distributions by innovatively removing the constraint of lack of species absence data, thus providing a robust prediction of distribution for stream fishes in other regions where

historical data exist, and for other taxa (e.g., benthic macroinvertebrates, birds) usually observed by community-based sampling designs.

Key words: species distribution models; metacommunity database; inferred absence; species prevalence; spatial autocorrelation; boosted regression trees; cross-validation; stream fish; rarity.

INTRODUCTION

Understanding species-habitat relationships and the spatial pattern of species distributions is critical in biogeography, biodiversity conservation, and resource management (Elith et al. 2006, Franklin and Miller 2009). Through modeling historical ranges, suitable locations could be determined for reintroducing and recovering declining or extirpated species (Pearce and Lindenmayer 1998, Hirzel et al. 2004). Based on current biological sampling surveys, species distribution models (SDMs) could be used to design conservation or management plans (Bani et al. 2002, Esselman and Allan 2010, Zarkami et al. 2012). Conservation managers could predict and mitigate the effect of potential climate and landscape changes on economic or threatened species (Chu et al. 2005, Brown et al. 2009, Bond et al. 2011), and find strategies to control species invasions (Kelly and Meetemeyer 2002, Anderson et al. 2004, Herborg et al. 2007) by updating habitat variables to future scenarios in calibrated models.

One component now limiting the progress of biodiversity conservation and resource management is biological data to support rigorous SDMs (Brotons et al. 2004, Franklin and Miller 2009, Elith and Leathwick 2009). Species occurrence data of high resolution, particularly at large spatial extents, are usually not available or not synthesized into readily usable forms. For example, NatureServe provides the most up-to-date electronic species distribution maps of US

freshwater fauna and flora with low resolution (http://www.natureserve.org/), but neither species-habitat relationships nor subtle temporal shifts in distribution are discernible from maps at such coarse resolutions. This limitation exists largely because gathering occurrence data by sampling each species' entire habitat range can be time-consuming and costly (Brotons et al. 2004). Observations of presence for rare, cryptic, and migratory freshwater fishes tend to be particularly spatially sparse, let alone the absence data that ideally require multiple-visit occupancy-based sampling designs. Constrained by data availability, most previous SDMs studies have focused on common or economic species (e.g., Clark et al. 2001, Perry et al. 2005), or developed models with only presence observations such as the Maximum-Entropy Species-Distribution Modeling or MaxEnt (Phillips et al. 2004). However, presence-only models can only estimate realized niche when the assumptions of known prevalence and sampling bias are valid (Soberón and Nakamura 2009), and usually yield less accurate species-habitat associations and species distributions than presence-absence models (Brotons et al. 2004, Yackulic et al. 2012).

Atlases have been the most common approach to present species occurrences at large spatial extents (Donald and Fuller 1998, Brotons et al. 2004). However, most distribution maps derive data from reports of the occurrence of species (i.e., a snapshot of presences), thus they only provide limited information on species abundances and relative habitat suitability. It is easy to underestimate presence consistently in interpreting these maps, because a species is considered absent in locations subjected to no or very low sampling effort. Particularly, non-game species of fish that have not been the focus of any specific conservation studies and species whose detection depend strongly on sampling gear, effort, or habitat type will tend to show higher numbers of false absences. Alternatively, researchers have used museum records to

evaluate species distribution across multiple states or such large sampling units. Yet, some common limitations of museum data have been identified, including: 1) they may not accurately locate the position of records collected before the era of GPS (Zimmermann et al. 2010), 2) they are usually collected with varied sampling approaches and intensities, 3) they span long time periods in which the habitats might have changed substantially, 4) and they are not sufficient in quantity to delineate full distribution ranges of species and develop robust models (Herborg et al. 2007). These aspects of sampling biases tend to inflate false negative or positive rates in the less sampled areas, and underestimate species' dispersal and invasion ability in prediction studies (Elith et al. 2006, 2011, Zimmermann et al. 2010). A framework that can appropriately synthesize species occurrence from field surveys and literature would provide an avenue to fill the gaps in data for modeling and predicting the spatial distribution of species.

We propose a framework for modeling species distributions using historical presences of species recorded in high-resolution atlases and absences inferred from locations where historical presences have been recorded for other species known to be typically sampled as part of a community. Applying the framework to freshwater fishes of the United States, non-game species are better indicators of community sampling. Unlike non-game species, the presence of game species in a sample can be of questionable utility in inferring habitat suitability because populations of game species exist in many suboptimal habitats due to repeated stocking. In addition, whereas game species tend to be targeted for recreation and oversampled, non-game species appear in presence records predominantly as part of community samples. Accumulated over many years, we propose that such samples offer a strong evidence of absence where a species has never been observed but presence records of other species exist. The presences and inferred absences used in this study were derived from historical atlas on species distribution

published during 1950-1990. It is not possible to go back to the sampling period to verify whether these presences and inferred absences were true directly. Instead, we tested the accuracy of these presences and absences using species distribution models (SDMs). These presences and absences are useful and probably accurate if the performance of these SDMs have good performances and the species-habitat relationships described by these models corroborate with studies based on field observations.

Species observed over multiple spatial and temporal scales in a defined geographic area belong to a metacommunity (Gilpin and Hanski 1991, Wilson 1992, Leibold et al. 2004). In practice, developing such a metacommunity sample involves collating historical occurrences of fish species from different sources and deriving absences for one species from known presences of other species. Communities in a defined metacommunity are assumed spatially connected by migrating and dispersing individuals and species (Gilpin and Hanski 1991, Wilson 1992), and local community compositions are determined by the regional species pool and regulated by local environmental factors according to two of the prevailing perspectives of metacommunities (Leibold et al. 2004, Niu et al. 2012). Whereas species present in a sampling unit belong to the same regional pool, they may not all have co-existed in that unit at any point in time, and coexisting species may not be observed in a single sampling visit either, due to the variability in sampling technique, timing, effort, and detection rate of different species (MacKenzie et al. 2002, Huang et al. 2011, Pacifici et al. 2012, Pritt and Frimpong 2014). The temporal and spatial dependencies of occurrence are particularly strong for vagile species, such as fish, which regularly move among feeding, breeding, and over-wintering or summer habitats (Angermeier and Schlosser 1995). The compilation and documentation of the metacommunity database ware provided in the Appendix A.

In this study, we used boosted regression trees (BRT), logistic regression and MaxEnt models to assess the performance of a historical metacommunity database, with the overarching objective of 1) comparing presence-only and presence-absence models, where we infer the absence of a species from accumulated evidence of the presence of other fish species. Additionally, we investigated the effect of 2) data resolution at two levels (i.e., inter-confluence segment level and the coarser watershed level), 3) species' rarity and sampling prevalence, and 4) spatial autocorrelation on model performance. We modeled habitat suitability and distribution of 76 selected freshwater fish species (representing approximately 10% of described freshwater fish species of the United States) exhibiting a range of rarity in four basins of the United States. We used principal coordinate analysis of neighbor matrices (PCNM; Borcard and Legendre 2002) to incorporate spatial autocorrelation into the species distribution models as a means to evaluate the effects of spatial autocorrelation on model performance. We assessed specific habitat requirements for the selected species through partial dependence plots in the BRT models. Data resolution, species' rarity and sampling prevalence, and spatial autocorrelation, are major known but not fully understood factors affecting the behavior and performance of SDMs, and likely to corrupt inference if not properly controlled in the quest to investigate any major hypothesis. Our main hypothesis was that presence-absence models developed with inferred absences would outperform presence-only models in terms of discrimination power (i.e., the area under the Receiver-Operating-Characteristic curve). If this hypothesis is supported, then the existence of vast historical freshwater fish presences for the entire United States, synthesized into a single metacommunity database, constitutes an enormous resource for SDMs to help address myriad ecological, conservation, and resource management problems.

METHODS

Selecting study basins and fish species

It is imperative for the evaluation of a database and for comparison of different models, to include a variety of regions and a range of common and rare species so that limitations of the proposed modeling approach can be uncovered and explicated. We selected four basins in the United States for this study: New River, Illinois River, Brazos River, and Snake River, meeting criteria of data availability and geographic diversity (Figure 2.1). The four selected river basins spanned a range of climate, physiography, and anthropogenic influences (e.g., hydrological alterations, agriculture, and urbanization). The Brazos River is warmer than other 3 basins and it showed narrower range and smaller variation in temperature (PRISM climate group 2004). The dominant landscape in the New River, Illinois River, Brazos River, and Snake River basins were forest/agriculture, agriculture/urban, forest/grassland, and grassland/forest, respectively. We selected 76 fish species (Table 2.1; Table A.1) with different rarity and distributional characteristics from the four river basins to develop habitat suitability and species distribution models. The 76 freshwater species belong to 15 families, and together represent approximately 10% of all currently described freshwater fish species of the United States and a phylogenetically diverse subset of species. The attributes considered in the species selection included a variety of macrohabitat preferences, body size, migration ability, and temperature tolerances (Frimpong and Angermeier 2009) and the three common dimensions of rarity—range size, habitat breadth, and local population size (Rabinowitz 1981, Pritt and Frimpong 2010). The presence of these species were derived from the *IchthyMaps* database and spatially joined to NHDplusV2 segments (Appendix A). The absences of these species were inferred in the metacommunity matrix (e.g., Matrix B in the Figure 1.2) using the approach described in Appendix A. A metacommunity matrix contains 1 (species presence) and 0 (inferred asbence of a focal species

from the presence(s) of other non-game species). Row *i* of a metacommunity matrix represents co-occurrences of species belonging to a regional pool at unit *i*, and column *j* lists the presences and absences of species *j* at the study units. A total of 13,955 fish presence records occurring on 1,933 inter-confluence segments (study units) were collated to produce species distribution models for the 76 species in the four basins (Table 2.1). Total number of study units (N) was 293, 852, 575, and 195 in Brazos River, Illinois River, New River and Snake River basin respectively.

We used the inter-confluence segments of the enhanced 1:100,000 resolution National Hydrographic Dataset (NHDplusV2, available from http://www.horizon-systems.com/nhdplus/NHDPlusV2_home.php) as the primary study units. NHDplusV2 is a geographic and hydrologic framework dataset that has been widely applied to the environmental assessment and stream habitat management by the US Environmental Protection Agency (USEPA), US Geological Survey (USGS), and other agencies. Matching the NHDplusV2 resolution (1:100K) allowed for convenient retrieval of numerous environmental (habitat) variables organized by stream segments and network accumulated attributes and for predicting species distribution at high resolution. We also coarsened the habitat and fish occurrence data to HUC-12 (12-digit hydrologic unit code) watershed level to examine the effect of data resolution on model performance, and more comprehensively compare different modeling approaches (i.e., presence-absence model versus presence-only model at both stream segment and watershed level).

Developing species distribution models

The species distribution models we developed in this study are habitat suitability models.

We used the definitions of Kearney (2006) for environment-"the biotic and abiotic phenomena surrounding and potentially interacting with an organism" and for habitat-"a description of a physical place, at a particular scale of space and time, where an organism either actually or potentially lives". Among over 50 available statistical approaches for SDMs, we selected to compare logistic regression under the Lasso (least absolute shrinkage and selection operator) regularization (Tibshirani 1996), boosted regression tree (BRT) model (Friedman 2001), and the Maximum-Entropy, MaxEnt (Phillips et al. 2004). Logistic regression has been conventionally used in SDM studies (e.g., Chu et al. 2005, Mugodo et al. 2006); using Lasso allows mitigation of multicollinearity, and selection of an optimal set of predictor variables. BRT is a more recent machine-learning approach that has outperformed counterparts in few comparative studies and reviews (e.g., Elith et al. 2006, Franklin and Miller 2009). MaxEnt was found superior to other presence-only models (e.g., GARP and bioclim) in previous comparative studies (e.g., Elith et al. 2006, Phillips et al. 2006).

In logistic regression, the probabilities of a defined success (e.g., presence of species at a site in this study) can be modeled with a set of the predictor variables, using a logistic link function as follows:

$$log\,\frac{p(y_i=1|x_i)}{p(y_i=0|x_i)} = \sum_{j=0}^{k}\beta_j x_{ij} \quad (1)$$

where $p(y=1|x_i)$ is the probability of presence at site i, β's are regression coefficients, x's are values of predictor variables. The coefficients are usually estimated by optimizing the likelihood function:

$$L(\beta) = p(\beta|Data) = \prod_{i=1}^{n} p(x_i)^{y_i}(1-p(x_i))^{1-y_i} \quad (2)$$

Under the Lasso regularization, the objective function is:

$$\log(L(\beta)) + \lambda \sum_{j=0}^{k}|\beta_j| \quad (3)$$

where $\sum_{j=0}^{k}\left|\beta_j\right| \leq S$ is the constraint added on the maximum likelihood optimization, and λ is the regularization or penalty parameter that needs to be tuned through validation. We implemented the Lasso-version logistic models in the R statistical program (R core team 2014) with the package 'glmnet' (Friedman et al. 2010).

Boosted regression trees (BRT) developed by Friedman et al. (2001) have gained popularity in recent studies of species distribution models (e.g., Elith et al. 2006). Boosting is the algorithm that ensembles individual classifiers (e.g., classification trees, regression trees) and sequentially fine-tunes the model by using weighted average of predictions (Friedman 2001). The optimal number of trees were determined through minimizing the loss function in terms of deviance reduction, while achieving a good balance between tree complexity and learning rate (Elith et al. 2006). Most currently used BRT models also incorporate bagging algorithms. Bagging strategies (i.e., both samples and predictors are randomly sub-sampled without replacement from the full dataset) are applied at each iteration to control overfitting (by bagging samples) and incorporate complex non-linear relationships (by bagging predictors) (Breiman 2001). Analogous to other tree-based models, BRT models do not require pre-selecting or re-scaling predictor variables; instead, contribution (or importance) of each predictor variable are calculated based on the frequency of a variable being selected for splitting, weighted by the squared improvement to the model from each split across all trees (Friedman and Meulman 2003). Other appealing features of BRT models include resistance to outliers and multicollinearity, and applicability to data of small sample size but many predictors (i.e., the n<<p problem) (Friedman 2001, Elith et al. 2006). We implemented the boosted regression tree models with the R package 'dismo' (Hijmans et al. 2013).

MaxEnt is specialized from the statistical mechanics theories for species distribution model with only presence data (Phillips et al. 2004). Entropy maximizes as the system disperses to equilibrium over time (Jaynes 1957). From the ecological perspective, MaxEnt essentially searches the probability distributions of entropy that satisfies all constraints (i.e., the expectation of each environmental variable conditional on species presence needs to match its sampled mean). Environmental variables that have sound ecological basis generally impose strong constraints, which serves as a criterion to measure variable importance and variable selection in MaxEnt. As a generative machine learning approach, MaxEnt could fit complex species-habitat relationships and incorporate multiple types of predictors and interactions thereof. MaxEnt (Phillips et al. 2004) has been developed as a shareware that can be downloaded from www.cs.princeton.edu/~schapire/maxent/. We used the inferred absences from the metacommunity matrices for the MaxEnt absences, instead of pseudo absences randomly drawn from the background (a default setting in the MaxEnt). This change in setting should lower the false negative error rate and make the MaxEnt models comparable to the other models.

We evaluated the performance of logistic regression, BRT and MaxEnt models in terms of AUC (i.e., the area under the receiver-operating-characteristic (ROC) curve) in both training and 5-fold cross-validation processes. AUC is a threshold-independent measure of a model's discrimination power (i.e., the ability to discriminate occupied from unoccupied units). An ROC curve is a plot of sensitivity (true positive rate) against 1− specificity (true negative rate) at varying discrimination thresholds (Figure A.1). A discrimination threshold on the predicted probability of species presence is used to classify each unit as occupied (1) or unoccupied (0). For example, if the discrimination threshold is set to 0.36, a unit with predicted probability of species presence higher than 0.36 would be classified as an occupied unit. The area under the

ROC curve (AUC) ranging from 0 to 1 averages diagnostic accuracy across the whole range of threshold on the probability of presence (Hanley and McNeil 1982). AUC is a measure of discrimination power rather than goodness of fit (e.g., R^2 or pseudo R^2). The higher the AUC, the more accurately a species distribution model can rank observation sites or prediction sites from suitable to unsuitable for the focal species. Such index can be used to guide to delineate reserve for species. It is common to use the classification proposed by Swets (1988) to interpret AUC: excellent $AUC > 0.90$, good $0.80 < AUC < 0.90$, fair $0.70 < AUC < 0.80$, poor $0.60 < AUC < 0.70$; fail $0.50 < AUC < 0.60$. AUC has been used to a primary index in many studies to select modeling approach, to measure predictive performance, and evaluate transferability of species distribution models. For example, Bouska et al. (2015) used species distribution models with AUC higher than 0.75 to delineate suitable habitats for threatened amphibian species. We found other performance measures (e.g., correlations of observed occurrence and predicted occurrence probability, and deviance) to be significantly correlated with AUC, so for brevity only AUC is shown in the results.

The habitat factors considered in this study were in seven categories: climate, geology, hydrology, stream morphology, land use/land cover, disturbance, and water chemistry (Table 2.2 A). The climate data (e.g., temperature, precipitation) were obtained from the PRISM climate group (2004). The land cover data in 1980's for each NHD inter-confluence catchment and watershed were derived from the USGS Land Cover Institute (US Geological Survey 1986). Other environmental variables of biological importance to stream fish identified in the literature were retrieved from NHDplusV1 and NHDplusV2 (EPA, USGS and Horizon Systems Corporations 2010, 2012; McKay et al. 2012). In addition, we obtained the habitat condition index from the National Fish Habitat Action Plan (NFHAP) databases (Esselman et al. 2011,

National Fish Habitat Board 2012). The ranges of these predictor variables in each of the four river basins were listed in the Table 2.2 B. For each set of highly correlated variables (Pearson's $|r| > 0.8$), only one was kept to minimize multicollinearity. We examined the species-habitat relationships with the partial dependence plots of the optimized boosted regression model for each species.

We tested whether incorporating spatial autocorrelation would improve the performance of the species distribution models, using the principal coordinate analysis of neighbor matrices (PCNM) approach (Borcard and Legendre 2002, Borcard et al. 2004) in the R package 'PCNM' (Legendre et al. 2012). In the PCNM procedure, we used latitude and longitude to create a Euclidean distance matrix among all sampled stream segments in each of the four basins. We then truncated the distance matrices to a lower triangular matrix (i.e., elements above the diagonal are set to 0). Mutually orthogonal eigenvectors were then extracted from the truncated matrix, and those spatial eigenvectors associated with positive eigenvalues and significant Moran's I were kept to form the spatial matrix. Moran's I (Moran 1950) measures spatial autocorrelation based on both the values and locations of a variable. The null hypothesis in the Moran's I test is that there is no spatial autocorrelation in the tested variable. This null hypothesis is rejected if there is a strong clustered or dispersed pattern in the tested variable. The decision for the Moran's I test is usually based on the *p*-value calculated by a permutation on the values of the tested variable among the study units, or by approximating the Moran's I value to normal score. To incorporate the spatial information into the environmental predictors, we built multivariate regression models with environmental variables as responses and spatial matrix as predictors (Brind'Armour et al. 2005). We then used the predicted (i.e., 'spatialized') values of

the environmental variables from the multivariate regression as the model matrix instead of the raw environmental matrix in the spatial models.

After developing species distribution models with procedures described above, we used ANCOVA (Wildt and Ahtola 1977) to examine the effect of model choice (logistic regression, BRT, and MaxEnt), data resolution (inter-confluence segment and watershed level), species' rarity, and spatial autocorrelation on AUC. Species and basin were treated as blocking factors and family number (Nelson et al. 2006) as a covariate. The rarity classification of the modeled species was obtained from Pritt and Frimpong's (2010) implementation of Rabinowitz (1981). Rarity types are eight combinations of three dimensions (i.e., Dimension 1- range extent, Dimension 2-habitat specificity, and Dimension 3-local population size). Type (A) is common across three dimensions; Type (B) is rare in the Dimension 3; Type (C) is rare in the Dimension 1; Type (D) is rare in the Dimension 1 and 3; Type (E) is rare in the Dimension 2; Type (F) is rare in the Dimension 2 and 3; Type (G) is rare in the Dimension 1 and 2; and Type (H) is rare across three dimensions. The effects of species and basins were blocked because only models from the same dataset (species × basin) are comparable. Phylogenetic relationships among the species we studied might be another source of non-independence, so we used the family number (Nelson et al. 2006) as a surrogate of the phylogenetic eigenvector and treated it as a covariate in the ANCOVA (Diniz-Filho et al. 1998). We used the Box-Cox transformation (Box and Cox 1964) on the AUC to ensure that the linear model assumptions of normality of residuals and constant variance were valid.

We further selected four species (*Hypentelium nigricans*, *Cyprinella spiloptera*, *Nocomis platyrhynchus* and *Etheostoma osburni*) with rarity type A, C, E and F respectively in the New River to examine the effect of prevalence on SDM performance under training and validation.

The observed prevalence (i.e., the proportion of presences among all the observations in the raw data) of *E. osburni* and *C. spiloptera* were lower than *H.nigricans* and *N. platyrhynchus*. We applied a resampling procedure with following steps to obtain the mean AUC values over 100 logistic regressions for each species. 1) For *E. osburni* and *C. spiloptera*, we kept the total sample size (i.e., the sum of presence and absence records randomly sampled) at 100. For *N. platyrhynchus* and *H. nigricans*, we first set the total sample size at 300, and decreased to 100 to evaluate the effect of sample size, in addition to the effect of prevalence. 2) For each species, we varied prevalence between 0.1 and 0.9 by randomly sampling different ratios of presence and absence records without replacement. For example, we randomly sampled 10 observations from presences and 90 observations from absences to generate prevalence of 0.1, given a sample size of 100. 3) We built logistic regression for each sample and calculated the AUC in the fitting and 10-fold cross validation. 4) The steps 1-3 were repeated 100 times to obtain the mean AUC values for a species with each prevalence (0.1, 0.2, 0.3…0.9).

RESULTS

Summary of model performance

The choice of model and species' rarity designation were the factors that significantly affected the model performance in terms of validation AUC at significance level of 0.05, according to the ANCOVA (Table 2.3). Additionally, spatial autocorrelation affected model performance at significance level of 0.1.

The effect of model choice

The presence-absence model, Lasso logistic regression, outperformed the presence-only model, MaxEnt, in the 5-fold cross validation for the 76 study fish species (Table 2.3). In spite of the vast difference in training performance, validation AUC was not different between the two

presence-absence models, Lasso logistic model and BRT, according to the post hoc group comparisons in the Tukey's test (Tukey 1949) (Table A.2). The training AUC for the BRT models were all greater than 0.7 and over 65% of them were higher than 0.9, whereas the training AUC for the logistic models ranged from 0.55 to 0.95, with only about 15% higher than 0.9 (Table 2.1). The correlation between validation AUC of BRT and validation AUC of logistic models was very high, with Pearson's correlation over 0.90 (Figure 2.2). We focused on analyzing BRT models for brevity since the performance of the logistic model agreed in terms of the validation AUC. In addition, the BRT model provided a richer output for model interpretation, in the form of partial dependence plots and variable importance rankings.

The effect of species' rarity and prevalence

Model accuracy was slightly higher for rare species as defined by Pritt and Frimpong (2010). Particularly, species in the rarity Type B and Type D had AUC over 0.75 in the BRT cross validation, which outperformed most species in other rarity types (Table 2.1). Cross-validation AUC of models for species in the rarity B and C was significantly higher than AUC for species in the rarity A, according to the post hoc group comparisons (Table A.2). Rarity Type B, C and D are species with large geographic ranges but small local populations (Rabinowitz 1981, Pritt and Frimpong 2010).

The training AUC in the logistic model exhibited a U-shaped response to prevalence for both rare and common species (Figure 2.3). The total sample size (N) seemed to negatively affect model fitting since models with N = 100 had higher AUC than models with N = 300 in the fitting, for the two species examined with varying sample sizes. In contrast, the U-shaped response of AUC to prevalence disappeared in the 10-fold cross validation; and decreasing the total sample size for common species did not result in increased AUC in the cross validation.

The cross-validation AUC of the BRT models had a negative nonlinear relationship with the observed prevalence (i.e., the proportion of presences among all the observations) of the species, indicating that habitat suitability is easier to quantify when variance in occurrence is low (Figure A.2).

Spatial versus non-spatial models

The ANCOVA showed that incorporating spatial autocorrelation (at significance level of 0.1) improved model performance in terms of cross-validation AUC (Table 2.3). Specifically, model accuracy increased conspicuously in the spatial models for Yellow bullhead (*Ameiurus natalis*), Orangespotted sunfish (*Lepomis humilis*) and Longnose gar (*Lepisosteus osseus*) in the Brazos River, and Shorthead sculpin (*Cottus confusus*) in the Snake River. For instance, the Moran's I test (Moran 1950) on deviance residuals became non-significant (p-value >0.05) after accounting for spatial autocorrelation in the logistic models for Longnose gar (*L. osseus*) in the Brazos River.

Species-habitat relationships

We examined species-habitat relationship using measures of variable importance (or contribution) and partial dependence plots in the BRT models (Table A.3). In the non-spatial models, Base flow index (BFI), elevation (ELE), mean annual in-stream flow (MFU), 20-year average minimum January temperature (TMI), 20-year average maximum July temperature (TMA), percentage of agriculture in the watershed (D_AG), annual precipitation (PPT), human population density (POP), drainage area (DRA), and habitat condition index (HCI) were the 10 most important predictors among the 25 environmental variables examined. Generally, these variables related to fish occurrence non-linearly, including polynomial forms, and sudden change after thresholds were common (Figure 2.4). The hydrology-related variables (e.g., BFI, MFU and

PPT) were positively related to the occurrence of most fish species, but constant high flows or floods could be negative force for some species, particularly those living in steep mountain streams in the New River basin, such as Candy darter (*Etheostoma osburni*), Longnose dace (*Rhinichthys cataractae*) and Rosyface shiner (*Notropis rubellus*). Temperature, particularly extreme weather events in the winter and summer, were important factors constraining spatial distributions of most fish species, except for those living in the Brazos River basin. Majority of species responded negatively to habitat degradation, indicated by their associations with the habitat condition index. Nevertheless, tolerant and frequently introduced bait species such as Fathead minnow *(Pimephales promelas*) and Common shiner (*Luxilus cornutus*) appeared to be favored in habitats with intense human activities (e.g., high population density and high road density).

The 10 key predictor variables identified in the non-spatial BRT models remained in the spatial BRT models, but the most important predictor changed for 46 out of 86 models (Table A.4). The rank and percent contribution of the top three variables in the non-spatial models, base flow index (BFI), measures of temperature and elevation, dropped in 68%, 72% and 81% spatial models respectively (Figure A.2), suggesting that the importance of these variables may be inflated in the non-spatial models due to their built-in spatial dependence. Meanwhile, measures of local stream catchment disturbance, such as habitat condition index and land use type, gained more weights in the spatial models.

DISCUSSION

We have successfully demonstrated the utility of a high-resolution metacommunity database developed by integrating historical freshwater fish occurrences from state and national atlases and databases for modeling species distributions. We have also shown that at the highest

resolution, where such comprehensive datasets are most difficult to come by, presence-absence models outperform presence-only models in the critical step of model validation. Our results corroborate other studies that have previously suggested that the performance of a species distribution model depends on: 1) the data quality (Michener and Brunt 2000, Zuckerberg et al. 2011), 2) choice of statistical modeling technique (Guisan and Zimmermann 2000, Franklin and Miller 2009, Elith et al. 2011), 3) species' traits (Araújo and Luoto 2007, Pollock et al. 2014), and 4) incorporation of spatial autocorrelation (Borcard et al. 2004, Dormann et al. 2007, Bahn et al. 2008, Miller and Franklin 2010). We went further to show how these factors specifically affect models. The proposed framework of collating accumulated high-resolution species presence records into a metacommunity database, including inferred absence of species will serve as a comprehensive tool for understanding species-habitat relationships at multiple spatial scales and help improve conservation and management of taxa.

Presence observations were collated despite the different sampling techniques and crews, and more importantly, absences could be inferred from locations where historical presences have been recorded for other species, as long as there is no reason to conclude that sampling overwhelmingly targeted particular species. The approach used in inferring absences has a theoretical root in Bayesian reasoning (Soberón and Nakamura 2009) and these absences are presumed to be more accurate than the pseudo-absences that are randomly sampled from the background in the study area by default a presence-only model such as MaxEnt. The accuracy of inferred absences will be influenced by the resolution of species distribution atlases, GIS procedures (e.g., geo-referencing, snapping) and the accuracy of the habitat template to which the scanned presences are ultimately transposed. Including absences enable presence-absence models to accurately estimate the realized niche of a species (Soberón and Nakamura 2009) and

spatial autocorrelation can also be conveniently incorporated into these distribution models. Considering the financial cost, limited time, and the risk of sampling certain rare and vulnerable species to extinction, better utilization should be made of the data that have been gathered by researchers and government agencies through investments made over many decades. Such efforts would thus particularly facilitate delineating habitats for rare or endangered species, the conservation planning for which has been often constrained by data availability. As an illustration, relatively good model performance and accurate species-habitat relationship were obtained without new sampling for Candy darter (*Etheostoma osburni*) that is listed as near threatened on the IUCN red list (Dixon 1986).

We recommend the use of boosted regression tree models to select key environmental variables by measures of variable importance and evaluation of how a species responds to each environmental gradient by partial dependence curves. Partial dependence curves capture thresholds particularly well (Figure 2.4). Machine learning techniques developed in the last two decades have some attractive features, such as controlling multicollinearity (Breiman 2001, Friedman 2001) and being applicable for the case where the number of variables exceeds sample size (e.g., Huang et al. 2011). However, statistical machine learning techniques tend to over-fit data and produce complicated models with high-dimension interactions, making the model vulnerable in independent validation and prediction, as illustrated with Random Forest (Wenger and Olden 2012). Our results revealed that the BRT model, which have improvements over Random Forest, also tend to over-fit, particularly when the sample size was small. In this study, the BRT and logistic models did not differ significantly in AUC in the validation, although BRT outperformed in the training. For the Snake River Basin where the sample size was relatively small, the validation AUC of Lasso logistic model was even higher than the BRT models. Thus,

there is a tradeoff to make between potentially over-fitting a model and obtaining more versatile model outputs when sample size is small. The differing behavior of training and validation AUC observed in this study also demonstrates that only reporting model performance in the training or fitting could be misleading, particularly in studies comparing performance of different modeling approaches (e.g., Elith et al. 2006, Moisen et al. 2006, Peterson et al. 2007). While acknowledging that ecologists will have to continue to find ways to work efficiently with presence-only data, we also reinforce growing calls that presence-absence models should be used whenever absence records are available (Brotons et al. 2004, Peterson et al. 2007, Yackulic et al. 2012). Even the most powerful presence-only model, MaxEnt, lacks the ability to estimate species prevalence for accurate statistical inference (Elith et al. 2011), and to adequately evaluate model performance because no true absence are included. Our results show that it would be inefficient not to use the carefully inferred absence data and instead model distributions with a presence-only technique.

It is suggested to explicate spatial autocorrelation and association thereof with environmental predictors in modeling species distribution and assembly patterns (Diez and Pulliam 2007, Dormann et al. 2007). Incorporating spatial autocorrelation significantly improved model accuracy indicated by the ANCOVA in our study (Table 2.3), particularly for a few fish species in the Brazos River Basin (Table 2.1). Including environmental predictors (e.g., temperature, elevation, land use) that spatially auto-correlate may have already removed the spatial dependence in the residuals of the non-spatial model, so adding spatial eigenvectors from the PCNM would not improve the model performance in the New River, Illinois River, and Snake River basins. Theoretically, it is equivalent to the situation that adding covariates highly correlated with the covariates already in the model would not be beneficial. However, our results

showed that the suitable fish habitat delineated and predicted distribution of a species may change after spatializing the environmental variables (i.e., a process that detaches spatial information for the environmental variables), although the model performance in terms of AUC would not increase much. The "spatialization" technique utilized in this study essentially filtered the built-in spatial components in each predictor variables, so the variable contribution and rank, and species-habitat relationship are more robust in the spatial models.

Through a resampling procedure on the real data, we confirmed that the effect of prevalence on the model fitting could be confounded by the fact that the variance of the Bernoulli random variable is highest when $p = 0.5$ and lowest at the extremes. The fitting AUC exhibited a U-shaped response to the prevalence (Figure 2.3), corroborating observations based on simulated data (Peres-Neto and Cumming 2010, Santika 2011). The model performance measured by cross-validation AUC was not clearly affected by the prevalence compared to the consistent effect on training AUC, suggesting that cross-validation is essential especially when methods or species are being compared. In addition, we showed that decreasing the total sample size for common species resulted in increased AUC in the model fitting. This sample size effect may be the result of reduced variance in the response when sample size is small and analogous to the over-fitting in linear regression when the number of predictors is close to the sample size. Conclusively, this study provides support for both the ecological (habitat specificity) and statistical (variance of Bernoulli response) basis of rare species tending to have better model performance.

Our results corroborate previous studies that hydrology, climate and land form and cover are key factors that determine distribution of stream fish (e.g., Clark et al. 2001, Lyons et al. 2010). It was important initially to include predictors in various habitat categories (e.g.,

hydrology, stream geomorphology, climate, and anthropogenic impacts) since the biological and ecological traits for most rare non-game species are not well known. Using incomplete set of environmental variables would produce unreliable and misspecified models with the problem of lack of fit, which in turn either overestimate or underestimate species niche and distribution range. Our models demonstrate that none of the broad categories of habitat factors dominantly determined the distribution of these 76 fish species across the United States, and none should be excluded apriori in future species distribution models. Nutrient variables (Sum total of nitrogen and phosphorus) should be normalized by drainage area, but the use of land use/land cover variables and the habitat condition index capture aspects of the nutrient impact. Statistical techniques, such as tuning in the Lasso or ridge regression (Tibshirani 1996), importance ranking and built-in validation in machine-learning models (Breiman 2001, Friedman 2001), are available to fine-tune the set of predictor variables, so that over-parameterization and multicollinearity should not be a major concern.

ACKNOWLEDGEMENTS

This work was funded by the US Geological Survey Aquatic gap analysis program. We thank Caitlin Worsham, Dawn Mercer, Kaylie Fitzgerald, Liang Yu, Steve Floyd, and other students in EAF's lab at Virginia Tech for helping with processing maps to derive presence records.

REFERENCES

Anderson, P. K., A. A. Cunningham, N. G. Patel, and F. J. Morales. 2004. Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. Trends in Ecology & Evolution 19:535-544.

Angermeier, P. L., and I. Schlosser. 1995. Spatial variation in demographic processes of lotic fishes: conceptual models, empirical evidence, and implications for conservation. American Fisheries Symposium 17:392-401.

Araújo, M. B., and M. Luoto. 2007. The importance of biotic interactions for modelling species distributions under climate change. Global Ecology and Biogeography 16:743-753.

Bahn, V., W. B. Krohn, and R. J. O'Connor. 2008. Dispersal leads to spatial autocorrelation in species distributions: A simulation model. Ecological Modelling 213:285-292.

Bani, L., M. Baietto, L. Bottoni, and R. Massa. 2002. The use of focal species in designing a habitat network for a lowland area of Lombardy, Italy. Conservation Biology 16:826-831.

Bond, N., J. Thomson, P. Reich, and J. Stein. 2011. Using species distribution models to infer potential climate change-induced range shifts of freshwater fish in south-eastern Australia. Marine and Freshwater Research 62:1043-1061.

Borcard, D., and P. Legendre. 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. Ecological Modelling 153:51-68.

Borcard, D., P. Legendre, C. Avois-Jacquet, and H. Tuomisto. 2004. Dissecting the spatial structure of ecological data at multiple scales. Ecology 85:1826-1832.

Bouska, K, G. Whitledge and C. Lant. 2015. Development and evaluation of species distribution models for fourteen native central U.S. fish species. Hydrobiologia 747: 159-176.

Box, G. E. P. and D. R. Cox.1964. An analysis of transformations. Journal of the Royal Statistical Society (Series B) 26: 211-252.

Breiman, L. 2001. Random Forests. Machine Learning 45:5-32.

Brind'Amour, A., D. Boisclair, P. Legendre, and D. Borcard. 2005. Erratum: multiscale spatial distribution of a littoral fish community in relation to environmental variables. Limnology and Oceanography 50: 465–479.

Brotons, L., W. Thuiller, M. B. Araújo, and A. H. Hirzel. 2004. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. Ecography 27:437-448.

Brown, L. R., M. B. Gregory, and J. T. May. 2009. Relation of urbanization to stream fish assemblages and species traits in nine metropolitan areas of the United States. Urban Ecosystems 12:391-416.

Chu, C., N. E. Mandrak, and C. K. Minns. 2005. Potential impacts of climate change on the distributions of several common and rare freshwater fishes in Canada. Diversity and Distributions 11:299-310.

Chunco, A. J., S. Phimmachak, N. Sivongxay, and B. L Stuart. 2013. Predicting environmental suitability for a rare and threatened species (Lao Newt, *Laotriton laoensis*) using validated species distribution models. PLoS One 8:3.

Clark, M. E., K. A. Rose, D. A. Levine, and W. W. Hargrove. 2001. Predicting climate change effects on Appalachian trout: combining GIS and individual-based modeling. Ecological Applications 11:161-178.

Diez, J. M., and H. R. Pulliam. 2007. Hierarchical analysis of species distributions and abundance across environmental gradients. Ecology 88:3144-3152.

Diniz-Filho, J. A. F., C. E. R. De Sant'Ana, and L. M. Bini. 1998. An eigenvector method for estimating phylogenetic inertia. Evolution 52:1247–1262.

Donald, P. F. and R. J. Fuller. 1998. Ornithological atlas data: a review of uses and limitations. Bird Study 45: 129-145.

Dormann, C. F., J. M. McPherson, M. B. Araújo, R. Bivand, J. Bolliger, G. Carl, R. G. Davies, A. Hirzel, W. Jetz, W. Daniel Kissling, I. Kühn, R. Ohlemüller, P. R. Peres-Neto, B.

Reineking, B. Schröder, F. M. Schurr, and R. Wilson. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30:609-628.

Elith, J., C. H. Graham, R. P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, L. Jin, L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. McC. Overton, A. T. Peterson, and S. J. Phillips. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129-151.

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

Elith, J., S. J. Phillips, T. Hastie, M. Dudík, Y. E. Chee, and C. J. Yates. 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17:43-57.

EPA (Environmental Protection Agency), USGS (United States Geological Survey), Horizon Systems Corporations. 2010. NHDPlusV1 Data. Retrieved from http://www.horizon-systems.com/nhdplus/data.php on 25 April 2012.

EPA (Environmental Protection Agency), USGS (United States Geological Survey), Horizon Systems Corporations. 2012. NHDplusV2 Data. Retrieved from http://www.horizon-systems.com/nhdplus/NHDplusV2_data.php on 28 July 2013.

Esselman, P. C., and J. D. Allan. 2011. Application of species distribution models and conservation planning software to the design of a reserve network for the riverine fishes of northeastern Mesoamerica. Freshwater Biology 56:71-88.

Esselman, P. C., D. M. Infante, L. Wang, D. Wu, A. R. Cooper and W. W. Taylor. 2011. An index of cumulative disturbance to river fish habitats of the Conterminous United States from landscape anthropogenic activities. Ecological Restoration. 29: 133-151.

Froese, R. and D. Pauly. Editors. 2014. FishBase. World Wide Web electronic publication. www.fishbase.org, version (06/2014).

Franklin, J., and J. A. Miller. 2009. Mapping species distributions: spatial inference and prediction. Cambridge University Press, Cambridge, Massachusetts, USA.

Friedman, J., F. Jerome, H. Trevor, and T. Rob. 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33:1-22.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. The Annals of Statistics 29:1189-1232.

Friedman, J. H., and J. J. Meulman. 2003. Multiple additive regression trees with application in epidemiology. Statistics in Medicine 22:1365-1381.

Gilpin, M. E. and I. Hanski. 1991. Metapopulation dynamics: empirical and theoretical investigations. Academic Press, New York, USA.

Dixon, M. G. 1996. *Etheostoma osburni*. The IUCN red list of threatened species. Version 2014.2. <www.iucnredlist.org>. Downloaded on 02 July 2014.

Guisan, A., and N. E. Zimmermann. 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135:147-186.

Hanley, J. A., and B. J. McNeil. 1982. The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. Radiology 143: 29-36.

Herborg, L. M., N. E. Mandrak, B. C. Cudmore, and H. J. MacIsaac. 2007. Comparative distribution and invasion risk of snakehead (Channidae) and Asian carp (Cyprinidae) species in North America. Canadian Journal of Fisheries and Aquatic Sciences 64:1723-1735.

Hijmans, R. J., S. Phillips, J. Leathwick, and J. Elith. 2013. dismo: Species distribution modeling. R package version 0.9-3. http://CRAN.R-project.org/package=dismo.

Hirzel, A. H., P. Bertrand, P. A. Oggier, C. Yvon, C. Glenz, and R. Arlettaz. 2004. Ecological requirements of reintroduced species and the implications for release policy: the case of the bearded vulture. Journal of Applied Ecology 41:1103-1116.

Huang, J., Y. Cao, and K. S. Cummings. 2011. Assessing sampling adequacy of mussel diversity surveys in wadeable Illinois streams. Journal of the North American Benthological Society 30:923-934.

Jaynes, E. T. 1957. Information theory and statistical mechanics. Physical Review 106:620-630.

Kearney, M. 2006. Habitat, environment and niche: what are we modelling? Oikos 115: 186-191.

Kelly, M., and R. K. Meentemeyer. 2002. Landscape dynamics of the spread of sudden oak death. Photogrammetric Engineering and Remote Sensing 68:1001-1009.

Lee, D. S., S. P. Platania, and G. H. Burgess. 1980. Atlas of North American freshwater fishes. North Carolina State Museum of Natural History, Raleigh, North Carolina, USA.

Leibold, M. A., M. Holyoak, N. Mouquet, P. Amarasekare, J. M. Chase, M. F. Hoopes, R. D. Holt, J. B. Shurin, R. Law, D. Tilman, M. Loreau, and A. Gonzalez. 2004. The metacommunity concept: a framework for multi-scale community ecology. Ecology Letters 7:601-613.

Legendre, P., D. Borcard, F. G. Blanchet and S. Dray. 2012. PCNM: MEM spatial eigenfunction and principal coordinate analyses. R package version 2.1-2/r106. http://R-Forge.R-project.org/projects/sedar/.

Lyons, J., J. S. Stewart, and M. Mitro. 2010. Predicted effects of climate warming on the distribution of 50 stream fishes in Wisconsin, U.S.A. Journal of Fish Biology 77:1867-1898.

MacKenzie, D. I., and W. L. Kendall. 2002. How should detection probability be incorporated into estimates of relative abundance? Ecology 83:2387-2393.

McKay, L., T. Bondelid, T. Dewald, J. Johnston, R. Moore, and A. Rea. 2012. NHDPlus Version 2: User Guide.

Michener, W. K., and J. W. Brunt. 2000. Ecological data: design, management, and processing. Blackwell Science, Malden, Massachusetts, USA.

Miller, J., and J. Franklin. 2010. Incorporating spatial autocorrelation in species distribution models. Pages 685-702 in M. M. Fischer and A. Getis, editors. Handbook of applied spatial analysis. Springer, Berlin, Germany.

Moran, P. A. P. 1950. Notes on continuous stochastic phenomena. Biometrika 37: 17-23.

Moisen, G. G., E. A. Freeman, J. A. Blackard, T. S. Frescino, N. E. Zimmermann, and T. C. Edwards Jr. 2006. Predicting tree species presence and basal area in Utah: A comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. Ecological Modelling 199:176-187.

Mugodo, J., M. Kennard, P. Liston, S. Nichols, S. Linke, R. Norris, and M. Lintermans. 2006. Local stream habitat variables predicted from catchment scale characteristics are useful for predicting fish distribution. Hydrobiologia 572:59-70.

National Fish Habitat Board. 2012. National Fish Habitat Action Plan (NFHAP), 2nd Edition. National Fish Habitat Partnership (NFHP). Association of Fish and Wildlife Agencies, Washington, DC, USA.

Nelson, J. S. 2006. Fishes of the world, 4th edition. Wiley, New York, USA.

Niu, S. Q., M. P. Franczyk, and J. H. Knouft. 2012. Regional species richness, hydrological characteristics and the local species richness of assemblages of North American stream fishes. Freshwater Biology 57:2367-2377.

Pacifici, K., R. M. Dorazio, and M. J. Conroy. 2012. A two-phase sampling design for increasing detections of rare species in occupancy surveys. Methods in Ecology and Evolution 3:721-730.

Pearce, J., and D. Lindenmayer. 1998. Bioclimatic analysis to enhance reintroduction biology of the endangered helmeted honeyeater (*Lichenostomus melanops cassidix*) in Southeastern Australia. Restoration Ecology 6:238-243.

Peres-Neto, P. R. and G. S. Cumming. 2010. A multi-scale framework for the analysis of fish metacommunities. Gido, G.B. and D.A. Jackson (eds.) Community Ecology of Stream Fishes: Concepts, Approaches, and Techniques, American Fisheries Society Symposium.

Peterson, A. T., M. Pape, and M. Eaton. 2007. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. Ecography 30:550-560.

Perry, A. L., P. J. Low and J. R. Ellis and J. D. Reynolds. 2005. Climate change and distribution shifts in marine fishes. Science 308: 1912-1915.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190:231-259.

Phillips, S. J., M. Dud, and R. E. Schapire. 2004. A maximum entropy approach to species distribution modeling. Page 83 Proceedings of the twenty-first international conference on Machine learning. ACM, Banff, Alberta, Canada.

Pollock, L. J., R. Tingley, W. K. Morris, N. Golding, R. B. O'Hara, K. M. Parris, P. A. Vesk, and M. A. McCarthy. 2014. Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). Methods in Ecology and Evolution 5:397-406.

Pritt, J. J., and E. A. Frimpong. 2010. Quantitative determination of rarity of freshwater fishes and implications for imperiled-species designations. Conservation Biology 24:1249-1258.

Pritt, J. J., and E. A. Frimpong. 2014. The effect of sampling intensity on patterns of rarity and community assessment metrics in stream fish samples. Ecological indicators 39:169-178.

PRISM Climate Group, Oregon State University, http://prism.oregonstate.edu, created 4 Feb 2004.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rabinowitz, D. 1981. Seven forms of rarity. Pages 205-217 in H. Synge, editor. The biological aspects of rare plant conservation. John Wiley and Sons, Chichester, UK.

Santika, T. 2011. Assessing the effect of prevalence on the predictive performance of species distribution models using simulated data. Global Ecology and Biogeography 20: 181-192.

Soberón, J., and M. Nakamura. 2009. Niches and distributional areas: concepts, methods, and assumptions. Proceedings of the National Academy of Sciences of the United States of America 106:19644-19650.

Swets, J. 1988. Measuring the accuracy of diagnostic systems. Science 240: 1285-1293.

Tibshirani, R. 1996. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58:267-288.

Tukey, J. 1949. Comparing Individual Means in the Analysis of Variance. Biometrics 5: 99-114.

US Geological Survey. 1986. Land use and land cover digital data from 1:250,000 and 1:100,000 scale maps. Data user's guide 4. Reston, Virginia, USA.

Wenger, S. J., and J. D. Olden. 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. Methods in Ecology and Evolution 3:260-267.

Wildt, A. R., and O. Ahtola. 1978. Analysis of covariance. Sage Publications, Beverly Hills, California, USA.

Wilson, D. S. 1992. Complex interactions in metacommunities, with implications for biodiversity and higher levels of selection. Ecology 73:1984-2000.

Yackulic, C. B., R. B. Chandler, E. F. Zipkin, J. A. Royle, J. D. Nichols, E. H Campbell Grant and S. Veran. 2012. Presence-only modelling using MAXENT: when can we trust the inferences? Methods in Ecology and Evolution 4:236-243.

Zarkami, R., R. Sadeghi, and P. Goethals. 2012. Use of fish distribution modelling for river management. Ecological Modelling 230:44-49.

Zimmermann, N. E., T. C. Edwards, C. H. Graham, P. B. Pearman, and J. C. Svenning. 2010. New trends in species distribution modelling. Ecography 33:985-989.

Zuckerberg, B., F. Huettmann, and J. Frair. 2011. Proper data management as a scientific foundation for reliable species distribution modeling. Pages 45-70 in C. A. Drew, Y. F. Wiersma, and F. Huettmann, editors. Predictive species and habitat modeling in landscape ecology. Springer, New York, USA.

Table 2.1. Summary of performance (in terms of AUC) of logistic models with Lasso regularization (LM) and boosted regression tree (BRT) models in the training process (_train) and cross validation (_cv) for the 76 fish species in the four selected river basins (BR-Brazos River, IL-Illinois River, NR-New River, SN-Snake River). "BRT_s_cv" and "LM_s_cv" means that spatial autocorrelation was incorporated in the models. The prevalence was calculated as the proportion of presence in the total records (presence and inferred absence). The common name and family of each fish species are given in the Table A.1. Total number of study units (N) was 293, 852, 575, and 195 in Brazos River, Illinois River, New River and Snake River basin respectively. The number of presences and absences for a species in a basin could be calculated as N × Prevalence and N × (1 - Prevalence) respectively. For example, the number of presences of *Acrocheilus alutaceus* in the Snake River basin is 195 × 0.077 = 15.

Species	Basin	Rarity	Prevalence	AUC					
				BRT_train	BRT_cv	BRT_s_cv	LM_train	LM_cv	LM_s_cv
Acrocheilus alutaceus	SN	A	0.077	0.945	0.692	0.728	0.889	0.816	0.847
Ameiurus natalis	BR	A	0.099	0.968	0.658	0.909	0.765	0.710	0.910
Aphredoderus sayanus	IL	A	0.033	0.972	0.672	0.737	0.773	0.730	0.655
Campostoma anomalum	NR	A	0.466	0.998	0.721	0.644	0.689	0.710	0.681
Campostoma anomalum	BR	A	0.222	0.981	0.894	0.728	0.883	0.836	0.741
Catostomus columbianus	SN	A	0.185	0.967	0.705	0.662	0.823	0.737	0.706
Catostomus commersonii	NR	A	0.383	0.746	0.587	0.565	0.653	0.619	0.573
Catostomus commersonii	IL	A	0.276	0.765	0.600	0.594	0.598	0.559	0.582
Cottus bairdii	NR	A	0.195	0.866	0.621	0.654	0.678	0.615	0.655
Cottus bairdii	SN	A	0.318	0.938	0.679	0.735	0.803	0.719	0.741
Cottus kanawhae	NR	A	0.047	0.999	0.969	0.934	0.500	0.967	0.944
Cottus confusus	SN	A	0.051	0.992	0.648	0.883	0.840	0.703	0.748
Cyprinella galactura	NR	A	0.045	0.998	0.882	0.882	0.933	0.897	0.917
Cyprinella lutrensis	IL	A	0.256	0.838	0.705	0.696	0.731	0.715	0.716
Cyprinella venusta	BR	A	0.102	0.939	0.834	0.928	0.826	0.804	0.935
Dorosoma petenense	BR	A	0.102	0.964	0.852	0.913	0.932	0.845	0.927
Etheostoma blennioides	NR	A	0.270	0.890	0.792	0.717	0.725	0.730	0.717
Etheostoma caeruleum	NR	A	0.042	0.997	0.939	0.958	0.500	0.973	0.957

Etheostoma exile	IL	A	0.021	0.996	0.896	0.857	0.500	0.909	0.842
Etheostoma nigrum	IL	A	0.287	0.798	0.621	0.576	0.644	0.639	0.610
Fundulus notatus	IL	A	0.115	0.825	0.639	0.646	0.712	0.674	0.684
Gambusia affinis	BR	A	0.454	0.912	0.650	0.623	0.701	0.572	0.593
Hypentelium nigricans	NR	A	0.456	0.737	0.561	0.546	0.623	0.548	0.569
Ictalurus punctatus	BR	A	0.294	0.925	0.745	0.679	0.711	0.721	0.702
Lepisosteus osseus	BR	A	0.102	0.985	0.792	0.939	0.910	0.775	0.931
Lepomis humilis	BR	A	0.372	0.870	0.659	0.838	0.625	0.634	0.831
Lepomis megalotis	BR	A	0.174	0.889	0.622	0.724	0.618	0.572	0.760
Luxilus chrysocephalus	IL	A	0.234	0.801	0.697	0.667	0.690	0.684	0.669
Luxilus chrysocephalus	NR	A	0.056	0.998	0.954	0.935	0.500	0.953	0.940
Luxilus cornutus	IL	A	0.068	0.974	0.905	0.895	0.918	0.908	0.897
Menidia beryllina	BR	A	0.143	0.923	0.701	0.676	0.708	0.746	0.714
Nocomis biguttatus	IL	A	0.206	0.782	0.639	0.615	0.649	0.651	0.649
Nocomis leptocephalus	NR	A	0.188	0.862	0.632	0.711	0.639	0.585	0.725
Notropis atherinoides	IL	A	0.101	0.918	0.784	0.764	0.812	0.775	0.762
Notropis dorsalis	IL	A	0.279	0.863	0.642	0.604	0.706	0.664	0.651
Notropis hudsonius	IL	A	0.062	0.972	0.821	0.814	0.500	0.826	0.831
Notropis rubellus	NR	A	0.268	0.822	0.648	0.621	0.692	0.624	0.650
Notropis stramineus	IL	A	0.253	0.752	0.584	0.583	0.620	0.571	0.582
Notropis volucellus	NR	A	0.167	0.868	0.725	0.719	0.771	0.773	0.722
Noturus gyrinus	BR	A	0.113	0.950	0.781	0.847	0.835	0.827	0.887
Pimephales notatus	NR	A	0.301	0.934	0.745	0.773	0.725	0.719	0.788
Pimephales promelas	BR	A	0.106	0.948	0.668	0.675	0.757	0.740	0.786
Pimephales vigilax	BR	A	0.628	0.903	0.685	0.617	0.670	0.647	0.580
Prosopium williamsoni	SN	A	0.097	0.903	0.591	0.563	0.733	0.783	0.578
Ptychocheilus oregonensis	SN	A	0.103	0.985	0.839	0.822	0.894	0.909	0.914
Rhinichthys cataractae	SN	A	0.179	0.887	0.599	0.525	0.836	0.661	0.538
Rhinichthys cataractae	NR	A	0.242	0.754	0.578	0.574	0.632	0.604	0.596

Richardsonius balteatus	SN	A	0.231	0.843	0.569	0.600	0.720	0.627	0.634
Amia calva	IL	B	0.032	0.977	0.831	0.806	0.500	0.778	0.824
Etheostoma microperca	IL	B	0.036	0.982	0.880	0.898	0.500	0.868	0.883
Lythrurus ardens	NR	B	0.083	0.919	0.758	0.774	0.842	0.714	0.785
Notropis buccatus	IL	B	0.044	0.983	0.848	0.887	0.500	0.799	0.888
Notropis buccatus	NR	B	0.054	1.000	0.967	0.903	0.500	0.971	0.951
Opsopoeodus emiliae	IL	B	0.029	0.976	0.790	0.715	0.850	0.838	0.746
Opsopoeodus emiliae	BR	B	0.126	0.972	0.880	0.749	0.942	0.909	0.828
Campostoma oligolepis	IL	C	0.021	0.998	0.818	0.833	0.862	0.728	0.789
Carpiodes velifer	IL	C	0.051	0.946	0.727	0.712	0.749	0.714	0.694
Cottus beldingii	SN	C	0.087	0.993	0.719	0.643	0.769	0.815	0.793
Cottus rhotheus	SN	C	0.041	0.998	0.664	0.750	0.956	0.817	0.810
Cyprinella spiloptera	IL	C	0.091	0.935	0.717	0.761	0.807	0.753	0.734
Cyprinella spiloptera	NR	C	0.155	0.926	0.793	0.817	0.803	0.789	0.809
Etheostoma chlorosoma	IL	C	0.046	0.958	0.830	0.788	0.826	0.804	0.825
Etheostoma chlorosoma	BR	C	0.092	0.998	0.946	0.868	0.966	0.936	0.860
Etheostoma spectabile	IL	C	0.134	0.898	0.677	0.636	0.705	0.648	0.652
Etheostoma spectabile	BR	C	0.113	0.982	0.892	0.858	0.909	0.892	0.876
Hybognathus nuchalis	IL	C	0.067	0.889	0.750	0.742	0.762	0.759	0.750
Ictiobus bubalus	IL	C	0.044	0.956	0.837	0.842	0.887	0.867	0.819
Notropis buchanani	BR	C	0.119	0.929	0.765	0.868	0.791	0.774	0.899
Percina phoxocephala	IL	C	0.097	0.928	0.743	0.677	0.801	0.747	0.640
Chrosomus erythrogaster	IL	C	0.062	0.945	0.744	0.697	0.813	0.737	0.732
Etheostoma asprigene	IL	D	0.052	0.970	0.823	0.785	0.854	0.838	0.825
Etheostoma gracile	BR	D	0.177	0.981	0.892	0.724	0.928	0.916	0.692
Percina sciera	BR	D	0.160	0.940	0.782	0.861	0.763	0.736	0.832
Catostomus ardens	SN	E	0.062	0.999	0.846	0.908	0.922	0.921	0.954
Luxilus cerasinus	NR	E	0.059	0.987	0.751	0.771	0.876	0.728	0.785
Nocomis platyrhynchus	NR	E	0.289	0.817	0.668	0.600	0.700	0.625	0.604

Notropis rubricroceus	NR	E	0.033	0.973	0.798	0.772	0.895	0.771	0.893
Percina oxyrhynchus	NR	E	0.111	0.888	0.718	0.752	0.789	0.760	0.749
Percina roanoka	NR	E	0.064	0.991	0.844	0.831	0.730	0.838	0.791
Chrosomus oreas	NR	E	0.191	0.895	0.678	0.776	0.655	0.659	0.725
Etheostoma kanawhae	NR	F	0.099	0.945	0.747	0.812	0.770	0.744	0.785
Etheostoma osburni	NR	F	0.111	0.980	0.869	0.835	0.908	0.887	0.858
Notropis scabriceps	NR	F	0.174	0.877	0.695	0.666	0.731	0.647	0.647
Luxilus albeolus	NR	G	0.151	0.857	0.667	0.727	0.707	0.702	0.745
Exoglossum laurae	NR	H	0.191	0.913	0.665	0.704	0.771	0.698	0.672
Phenacobius teretulus	NR	H	0.113	0.894	0.750	0.774	0.764	0.759	0.782

Notes: The rarity classification of the selected species were obtained from Pritt and Frimpong's (2010) implementation of Rabinowitz (1981). Rarity types are eight combinations of three dimensions (i.e., Dimension 1- range extent, Dimension 2-habitat specificity, and Dimension 3-local population size) in the rarity classification framework (Rabinowitz 1981). Type (A) is common across three dimensions; Type (B) is rare in the Dimension 3; Type (C) is rare in the Dimension 1; Type (D) is rare in the Dimension 1 and 3; Type (E) is rare in the Dimension 2; Type (F) is rare in the Dimension 2 and 3; Type (G) is rare in the Dimension 1 and 2; and Type (H) is rare across three dimensions.

Table 2.2 A. The sources and descriptions of environmental variables used to develop species distribution models for the 76 native stream fish species in the United States. Data are from NHDplusV1 (National Hydrography Dataset plus Version 1; EPA, USGS, Horizon Systems Corporations 2010) and NHDplusV2 (National Hydrography Dataset plus Version 2; EPA, USGS, Horizon Systems Corporations 2012), NFHAP (National Fish Habitat Board 2012), USGS-LCI (US Geological Survey-Land Cover Institute; USGS 1986), and PRISM (PRISM climate group 2004).

Variable	Type	Source	Description
COMID	/	NHDplusV2	Common identifier of an NHD flow line
SINU	Stream morphology	NHDplusV2	Sinuosity. Reach length divided by straight line length (straight line from beginning node to end node of reach)
ELE	Geology	NHDplusV2	Mean elevation in meters
SLP	Geology	NHDplusV2	Mean slope in degrees
RDX	Disturbance	NHDplusV2	Number of road-stream crossings per inter-confluence catchment
BFI	Hydrology	NHDplusV2	The ratio of base flow (i.e., the component of streamflow attributed to ground-water discharge) to total flow, expressed as a percentage.
SO	Stream morphology	NHDplusV2	Stream order (Strahler 1952)
DRA	Stream morphology	NHDplusV2	Total area of catchment (Square meters)
MFU	Hydrology	NHDplusV2	Mean Annual Flow in cubic feet per second (cfs) at bottom of flowline as computed by Unit Runoff Method
MVU	Hydrology	NHDplusV2	Mean Annual Velocity (fps) at bottom of flowline as computed by Jobson Method (1996)
HCI	Disturbance	NFHAP	An index of cumulative disturbance of catchments of inter-confluence stream segments calculated based on 15 disturbance variables (Esselman et al. 2011). The influence of each distribution variable was weighted by the results of multiple linear regression of all variables against a commonly used biological indicator of habitat condition (i.e., percent intolerant fishes at a site).

NT	Water chemistry	NHDplusV1	Sum total of Nitrogen in the catchment in kilograms
PT	Water chemistry	NHDplusV1	Sum total of Phosphorus in the catchment in kilograms
POP	Disturbance	NHDplusV1	Human population density (Persons per square kilometer multiplied by 10)
TMI	Climate	PRISM	20-Year (1961-1980) average annual minimum temperature in Celsius multiplied by 100 for each NHDPlus catchment
TMA	Climate	PRISM	20-Year (1961-1980) average annual maximum temperature in Celsius multiplied by 100 for each NHDPlus catchment
TM	Climate	PRISM	20-Year (1961-1980) average temperature in Celsius multiplied by 100 for each NHDPlus catchment
PPT	Climate	PRISM	20-year (1961-1980) average annual precipitation in millimeters multiplied by 100 (Millimeters multiplied by 100)
C_UB	Land use/land cover	USGS-LCI	percentage of urban in the inter-confluence catchment
C_AG	Land use/land cover	USGS-LCI	percentage of agriculture in the inter-confluence catchment
C_FR	Land use/land cover	USGS-LCI	percentage of forest in the inter-confluence catchment
C_WT	Land use/land cover	USGS-LCI	percentage of water in the inter-confluence catchment
D_AG	Land use/land cover	USGS-LCI	percentage of agriculture in the watershed
D_FR	Land use/land cover	USGS-LCI	percentage of forest in the watershed
D_WT	Land use/land cover	USGS-LCI	percentage of water in the watershed
D_UB	Land use/land cover	USGS-LCI	percentage of urban in the watershed

Table 2.2 B. The minimum (min) and maximum (max) values of predictor variables in each river basin (BR-Brazos River, IL-Illinois River, NR-New River, and SN-Snake River). The descriptions of variables were listed in Table 2.2 A.

Variables	BR		IL		NR		SN	
	Min	max	min	max	min	max	min	max
SINU	1.0	4.8	1.0	3.4	1.0	5.8	1.0	2.2
ELE	4.5	963.0	127.2	294.1	254.3	1364.2	295.0	2644.8
SLP	0.1	7.3	0.0	10.7	1.0	25.7	0.1	33.7
RDX	0	46	0	33	0	25	0	9
BFI	3.0	38.6	14.0	66.1	32.3	67.0	44.8	85.9
SO	1	7	1	8	1	8	1	8
DRA	0.3	109842.8	0.8	67896.5	0.3	17993.3	0.1	208887.9
MFU	0.1	7575.4	0.2	18057.8	0.2	9978.1	0.0	46669.9
MVU	0.4	3.1	0.4	2.6	0.6	3.6	0.4	3.9
HCI	1.0	4.3	0.0	4.0	1.4	4.4	0.0	4.9
NT	0.9	375350.2	2.3	1824836.4	0.0	319763.2	0.7	9960081.0
PT	0.2	68493.7	0.5	366853.1	0.0	82509.1	0.2	2459485.0
POP	0.0	1869.0	0.0	4428.6	0.0	2457.4	0.0	1176.5
TMI	-3.7	6.2	-12.5	-7.2	-10.3	-5.1	-15.8	-2.5
TMA	33.2	36.8	27.3	31.5	24.2	29.4	21.3	35.7
TM	15.6	20.4	8.0	12.6	7.1	12.2	0.7	12.0
PPT	460.5	1381.9	820.9	1036.8	887.7	1556.6	205.8	1766.1
C_UB	0.0	99.0	0.0	100.0	0.0	95.5	0.0	98.5
C_AG	0.0	100.0	0.0	100.0	0.0	100.0	0.0	100.0
C_FR	0.0	100.0	0.0	100.0	0.0	100.0	0.0	100.0
C_WT	0.0	53.1	0.0	95.7	0.0	63.1	0.0	65.7
D_AG	0.4	98.5	0.0	100.0	0.0	82.9	0.0	100.0
D_FR	0.0	79.4	0.0	69.2	13.5	99.4	0.0	100.0
D_WT	0.0	24.0	0.0	31.6	0.0	8.1	0.0	100.0
D_UB	0.0	63.1	0.0	100.0	0.0	25.2	0.0	47.1

Table 2.3. The Analysis of covariance (ANCOVA; Wildt and Ahtola 1977) for evaluating the effect of model types, incorporation of spatial autocorrelation, species' rarity type, and data resolution on the performance of species distribution models in terms of the area under the Receiver-Operating-Characteristic (ROC) curve (AUC). Degree of freedom (D.F), mean square (M.S.), F statistic and *p*-value are listed in this table.

Source	D.F.	M.S.	F	*p*-value
Treatment factors				
Model type	2	0.016	86.291	< 0.001
Spatial	1	0.00079	3.954	0.0504
Rarity	7	0.000936	5.012	< 0.001
Resolution	1	0.000366	1.957	0.163
Block factors and Covariate				
Basin	3	0.0004	2.153	0.093
Species	75	0.00083	4.463	0.035
Family number (covariate)	1	0.000017	0.090	0.764
Residuals	420	0.000187		

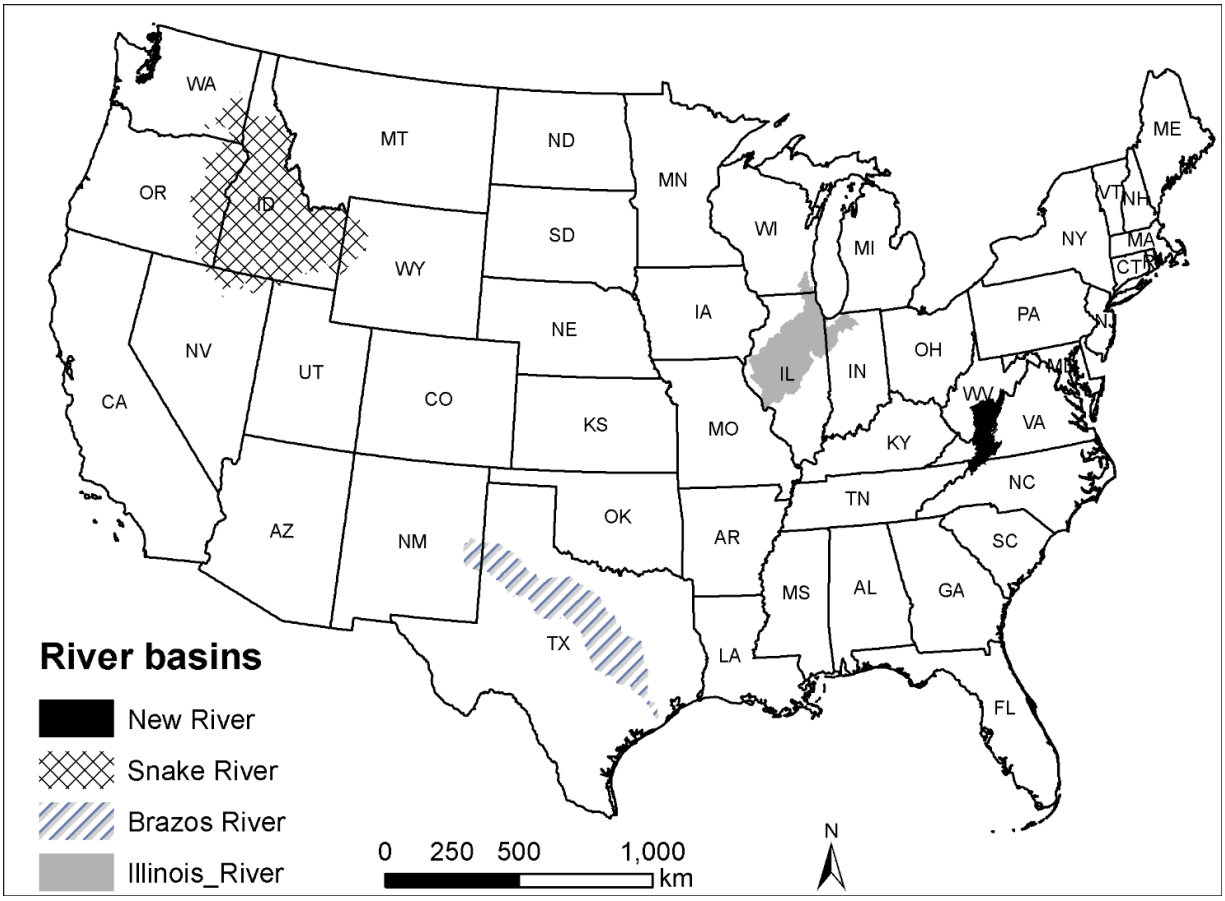


Figure 2.1. A map showing the distribution of 4 river basins (i.e., New River, Illinois River, Brazos River, and Snake River) selected for this study in the contiguous United States. We can see that all these 4 rivers pass through multiple states. Fish presence data are sufficient in these 4 basins in the *IchthyMaps* database for developing and validating species distribution models.

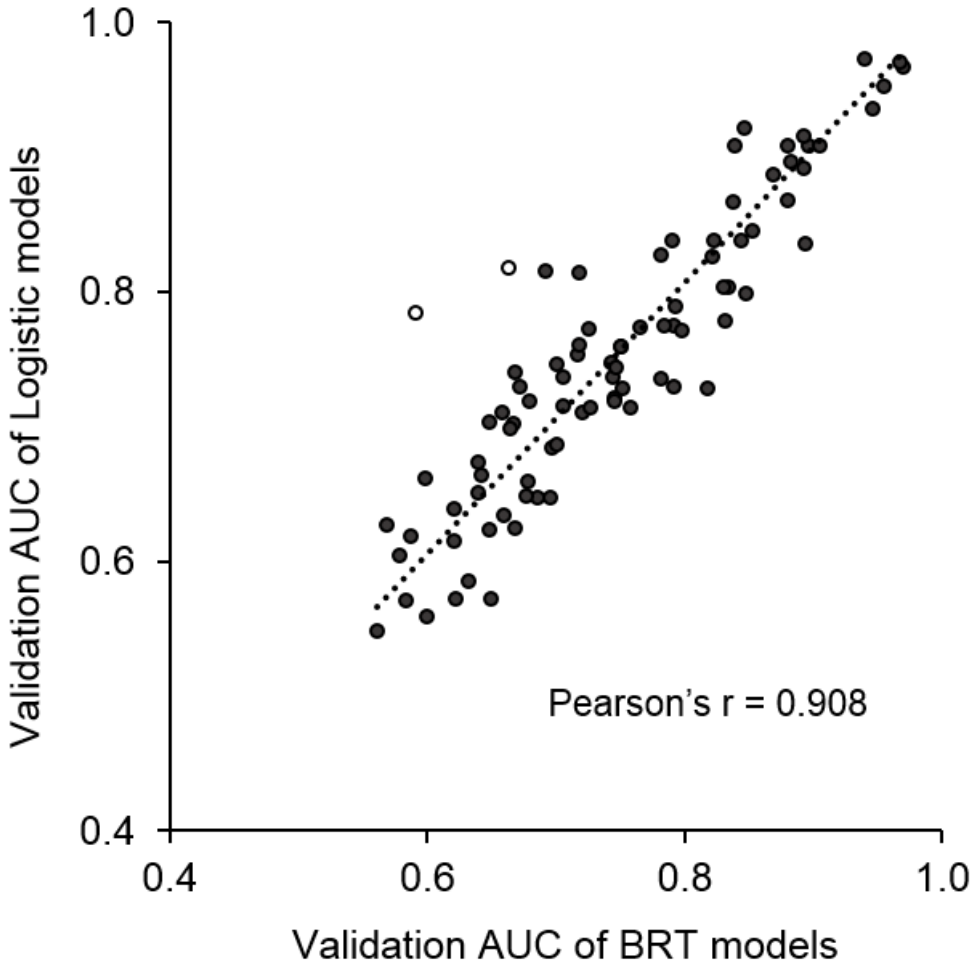


Figure 2.2. Comparing the performance of Lasso logistic regression model and boosted regression tree (BRT) models in terms of the area under the Receiver-Operating-Characteristic (ROC) curve in the 5-fold cross validation for 76 species in the 4 selected river basins (i.e., New River, Illinois River, Brazos River and Snake River). The results from the two set of models were generally in agreement, with Pearson's r over 0.9. For fish species Mountain whitefish, *Prosopium williamsoni* and Torrent sculpin, *Cottus rhotheus* (marked as circles) in the Snake River where occurrence data was relatively sparse, the Lasso logistic models outperformed the BRT models.

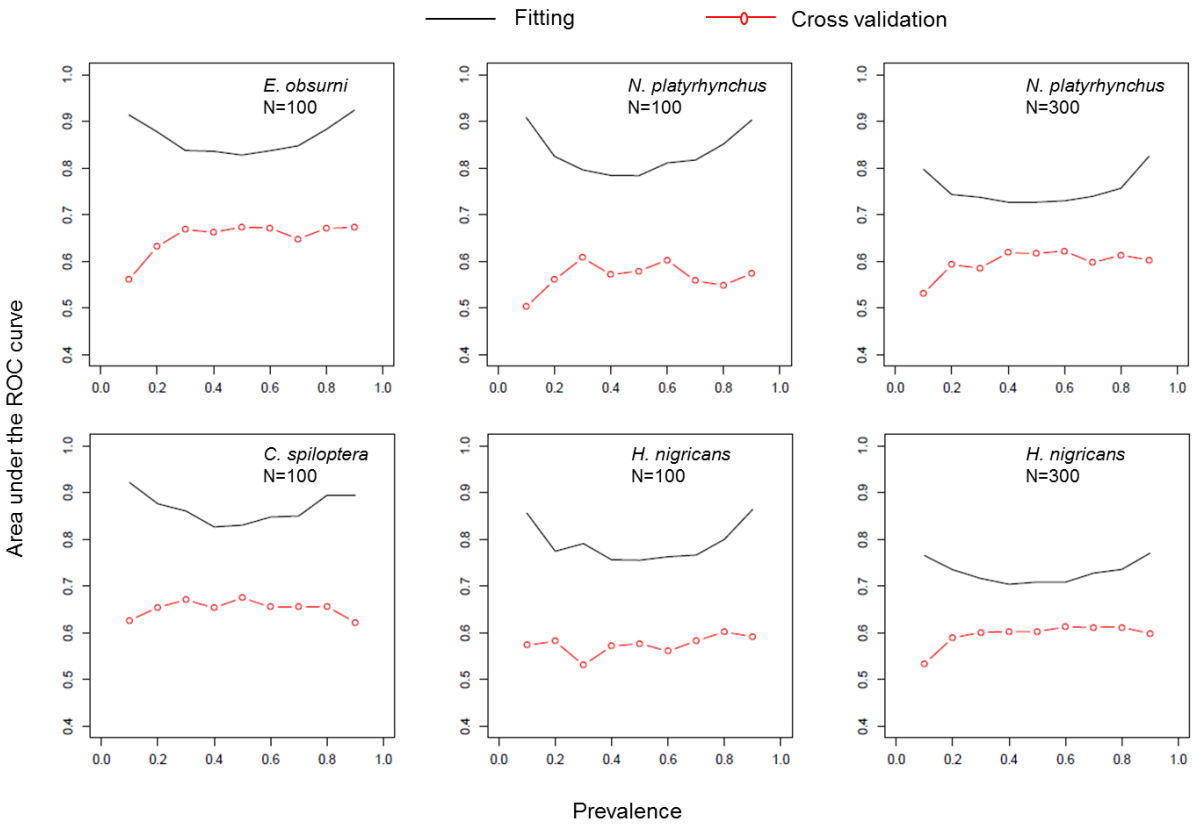


Figure 2.3. The effect of prevalence (i.e., the proportion of presences among all the observations) on the performance of species distribution models. The total sample size (N) for the two rare species, Candy darter (*Etheostoma osburni*) and Spotfin shiner (*Cyprinella spiloptera*), was set at 100; while N was decreased from 300 to 100 for the two common species, Bigmouth chub (*Nocomis platyrhynchus*) and Northern hog sucker (*Hypentelium nigricans*), to evaluate the effect of sample size.

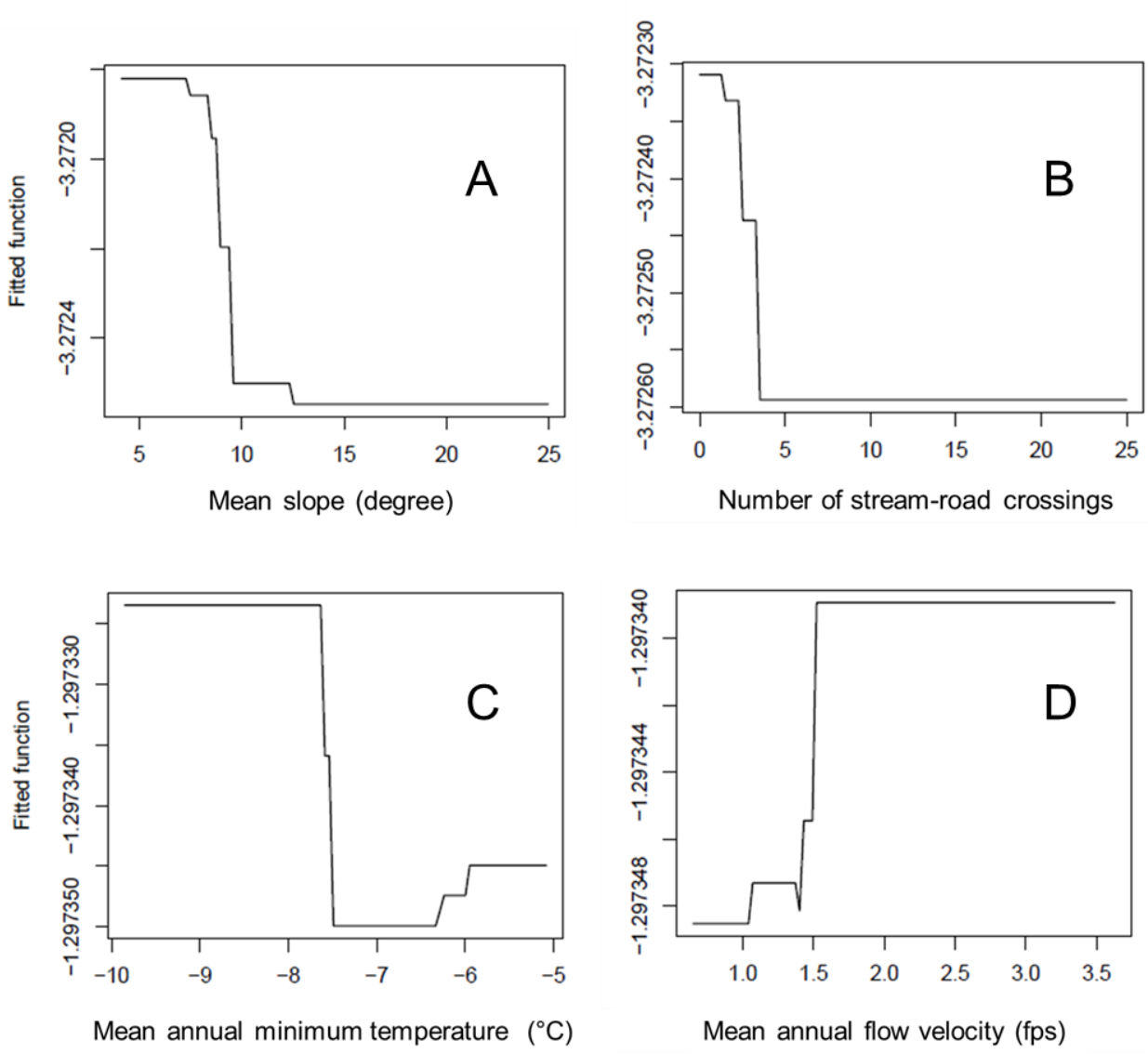


Figure 2.4. Examples of using partial dependence curves to capture ecological thresholds of spatial distribution of fish species in the New River. For example, the thresholds of mean slope (degree) in the watershed and number of stream-road crossings were identified for Rainbow darter (*Etheostoma caeruleum*) in the panel A and B. The thresholds of 20-year (1961-1980) average annual minimum temperature and mean annual flow velocity were identified for Mountain redbelly dace (*Chrosomus oreas*) in the panel C and D.

APPENDIX A SUPPLEMENTARY INFORMATION

The development of *IchthyMaps* database for advancing metacommunity ecology and freshwater fish conservation

The species presence and absence data used to develop species distribution models in this study were derived from the *IchthyMaps* database. This database, as a product of Aquatic Gap Analysis Program, contains 606,550 fish presence records in 224,305 NHDplusV2 (National Hydrography Dataset plus version 2) inter-confluence stream segments. We developed the *IchthyMaps* database by collating and synthesizing species occurrence records from atlases of freshwater fishes throughout the conterminous United States. We determined the general time frame when the occurrence records were obtained by referring to the books' prefaces or contacting the authors, although other information (e.g., method of collection, and sampling effort) was usually not specified for individual occurrence records in these atlases. Majority of the fish distribution atlases collected were published during 1970 to 1990 and presence records were sampled during 1950 to 1990 by electro-fishing. Therefore, we considered occurrence records up to 1990 as a practical cutoff for historical distributions, with the assumption that overall habitat and fish community of each stream segment did not change dramatically during this era compared to changes that might have happened since 1990 or could be of particular research and conservation interest in the future.

The work flow of deriving fish occurrences from atlases is described in five steps: 1) scanning individual atlases of species occurrences, 2) geo-rectifying and geo-referencing digitized atlases, 3) extracting occurrence records, 4) spatially joining the occurrence records to networks at different scales, and 5) integrating presence records to form the metacommunity database.

Step 1: We removed the maps from each book for all the states and regions where atlases were of usable detail and quality, for efficient creation of undistorted, high-resolution maps. For example, in this study, the atlases used to derive species occurrence records included: Atlas of North American Freshwater Fishes [1], Fishes of Montana [2], the Fishes of Illinois [3], Fishes of Idaho [4], Handbook of darters [5], the freshwater fishes of North Carolina [6], Freshwater fishes of Virginia [7], and the Fishes of West Virginia [8]. Each batch (book) of species occurrence atlases was digitized into raster layers through scanning and stored separately. Most printed maps of fish species presence used in this project were displayed on a base stream network map of a resolution that was very close to the NHDPlusV2.

Step 2: The digitized species occurrence atlases were then geo-rectified and geo-referenced to have the same coordinate and projection system as the NHDPlusV2 base map in ArcInfo (version 10). Geo-referencing many thousands of raster maps could be very time-consuming due to base maps of different sources, maps showing only sections of the country, and varied types of distortion in the scanning process. Therefore, we developed book-by-book templates with optimal coordinate and projection systems that greatly speeded up this process.

After geo-referencing the points on the maps to stream segments, the segments became representatives of locations where we consider fish sampling to have previously occurred.

Step 3: Presences records of each species, usually marked as dots or circles on the maps, were extracted by Feature Analyst, an automated feature extraction extension for ArcInfo (Textron Systems, Providence, RI). Feature Analyst allows rapid and accurate digitizing of target vector features from raster layers (e.g., occurrence points from geo-referenced maps in this case). We began by creating a shapefile of point feature class for each geo-referenced atlas, and providing Feature Analyst with a set of training features (e.g., dots, circles, triangles) matching the symbols for species occurrence, depending on symbols used in the original atlases. After setting up the learning parameters (e.g., searching extent, tolerance of similarity), feature analyst automatically searched symbols similar to those in the training set of the targeted species occurrence map. After running Feature Analyst, a technician inspected and adjusted the output point shapefile by adding the neglected points and removing false-positive errors. Multiple training and learning were iterated when multiple symbols corresponding to different time frames of data collection were used in a single map or when dense occurrence points existed. A point shapefile of occurrences for each species from each book was ultimately stored in a folder of map sources.

Step 4: Once shapefiles had been created for all scanned atlases, we used python scripts (Python 2.7) to populate the scientific name of each species and add latitude and longitude coordinates for each point in the attribute table of the species occurrence point shapefile. The updated attribute tables containing species name and latitude and longitude coordinates were integrated by merging all rows of records for each species. It turned out that integrating different GIS layers and all records for a single species through processing attribute tables was much more efficient than merging thousands of shapefiles in ArcGIS. We imported the ".DBF" files (i.e., attribute table) associated with the presence point shapefiles into the R program iteratively to join the imported tables by row in a 'for' loop. As a result, all the presence records were integrated into a table with 3 columns (species' scientific name, latitude and longitude of presence). We named this table as all-presence tables because it synthesized all presence records of fish species we collated from the historical atlases. We spatially joined presence records in this table to the NHDPlusV2 inter-confluence stream segments, HUC (hydrologic unit code) 8-digit and 12-digit watersheds in ArcGIS, so potential users can conveniently query the database at different scales.

Step 5: To construct a metacommunity matrix for a region, one need firstly querying the presence records of the region from the *IchthyMaps* database (for example, Matrix A in Figure 1.2). Those sampled segments with at least one presence of a non-game species recorded and no presence record of a focal species were designated as absences for that species. Game fish which are mostly sampled in targeted species surveys were excluded from inferring the absence of other species. Inferring absence from metacommunity database assumes that sampling was sufficient to detect all species occurred at each unit. The created metacommunity matrix is essentially a presence-absence matrix (Matrix B in Figure 1.2) with study units (segment or watershed) as the row names and species as the column names. Each row of the presence-absence matrix

represents co-occurrences of species belonging to a regional pool, and a column lists the presence and absence of a species at the study units (Figure 1.2). Creating presence-absence matrix from all-presence table can be done with the R “xtabs” function.

References in the Appendix

1. Lee DS, S. P. Platania, G. H. Burgess (1980) Atlas of North American Freshwater Fishes: North Carolina State Museum of Natural History.
2. Brown CJD (1971) Fishes of Montana: Montana State University.
3. Smith PW (1979) The Fishes of Illinois: University of Illinois.
4. Simpson JC, Wallace RL (1982) Fishes of Idaho: University Press of Idaho.
5. Page LM (1983) Handbook of darters: Tfh Publications Incorporated.
6. Menhinick EF (1991) The Freshwater Fishes of North Carolina. Raleigh, N.C.; North Carolina Wildlife Resources Commission ; Distributed by Larkin Distributors.
7. Jenkins RE, Burkhead NM (1994) Freshwater Fishes of Virginia: American Fisheries Society.
8. Stauffer JR, Jr., Boltz JM, White LR (1995) The Fishes of West Virginia. Proceedings of the Academy of Natural Sciences of Philadelphia 146: 1-389.

Table A.1. A table listing the common name and family of fish species modeled in this study.

Species	Family	Common name
Acrocheilus alutaceus	Cyprinidae	Chiselmouth
Ameiurus natalis	Ictaluridae	Yellow bullhead
Amia calva	Amiidae	Bowfin
Aphredoderus sayanus	Aphredoderidae	Pirate perch
Campostoma anomalum	Cyprinidae	Central stoneroller
Campostoma oligolepis	Cyprinidae	Largescale stoneroller
Carpiodes velifer	Catostomidae	Highfin carpsucker
Catostomus ardens	Catostomidae	Utah sucker
Catostomus columbianus	Catostomidae	Bridgelip sucker
Catostomus commersonii	Catostomidae	White sucker
Chrosomus erythrogaster	Cyprinidae	Southern redbelly dace
Chrosomus oreas	Cyprinidae	Mountain redbelly dace
Cottus bairdii	Cottidae	Mottled sculpin
Cottus beldingii	Cottidae	Paiute sculpin
Cottus kanawhae	Cottidae	Banded sculpin
Cottus rhotheus	Cottidae	Torrent sculpin
Cottus confusus	Cottidae	Shorthead sculpin
Cyprinella galactura	Cyprinidae	Whitetail shiner
Cyprinella lutrensis	Cyprinidae	Red shiner
Cyprinella spiloptera	Cyprinidae	Spotfin shiner
Cyprinella venusta	Cyprinidae	Blacktail shiner
Dorosoma petenense	Clupeidae	Threadfin shad
Etheostoma asprigene	Percidae	Mud darter
Etheostoma blennioides	Percidae	Greenside darter
Etheostoma caeruleum	Percidae	Rainbow darter
Etheostoma chlorosoma	Percidae	Bluntnose darter
Etheostoma exile	Percidae	Iowa darter
Etheostoma gracile	Percidae	Slough darter
Etheostoma kanawhae	Percidae	Kanawha darter
Etheostoma microperca	Percidae	Least darter
Etheostoma nigrum	Percidae	Johnny darter
Etheostoma osburni	Percidae	Candy darter
Etheostoma spectabile	Percidae	Orangethroat darter
Exoglossum laurae	Cyprinidae	Tonguetied minnow
Fundulus notatus	Fundulidae	Blackstripe topminnow
Gambusia affinis	Poeciliidae	Mosquitofish
Hybognathus nuchalis	Cyprinidae	Mississippi silvery minnow
Hypentelium nigricans	Catostomidae	Northern hog sucker

Ictalurus punctatus	Ictaluridae	Channel catfish
Ictiobus bubalus	Catostomidae	Smallmouth buffalo
Lepisosteus osseus	Lepisosteidae	Longnose gar
Lepomis humilis	Centrarchidae	Orangespotted sunfish
Lepomis megalotis	Centrarchidae	Longear sunfish
Luxilus albeolus	Cyprinidae	White shiner
Luxilus cerasinus	Cyprinidae	Crescent shiner
Luxilus chrysocephalus	Cyprinidae	Striped shiner
Luxilus cornutus	Cyprinidae	Common shiner
Lythrurus ardens	Cyprinidae	Rosefin shiner
Menidia beryllina	Atherinopsidae	Inland silverside
Nocomis biguttatus	Cyprinidae	Hornyhead chub
Nocomis leptocephalus	Cyprinidae	Bluehead chub
Nocomis platyrhynchus	Cyprinidae	Bigmouth chub
Notropis atherinoides	Cyprinidae	Emerald shiner
Notropis buccatus	Cyprinidae	Silverjaw minnow
Notropis buchanani	Cyprinidae	Ghost shiner
Notropis dorsalis	Cyprinidae	Bigmouth shiner
Notropis hudsonius	Cyprinidae	Spottail shiner
Notropis rubellus	Cyprinidae	Rosyface shiner
Notropis rubricroceus	Cyprinidae	Saffron shiner
Notropis scabriceps	Cyprinidae	New River shiner
Notropis stramineus	Cyprinidae	Sand shiner
Notropis volucellus	Cyprinidae	Mimic shiner
Noturus gyrinus	Ictaluridae	Tadpole madtom
Opsopoeodus emiliae	Cyprinidae	Pugnose minnow
Percina oxyrhynchus	Percidae	Sharpnose darter
Percina phoxocephala	Percidae	Slenderhead darter
Percina roanoka	Percidae	Roanoke darter
Percina sciera	Percidae	Dusky darter
Phenacobius teretulus	Cyprinidae	Kanawha minnow
Pimephales notatus	Cyprinidae	Bluntnose minnow
Pimephales promelas	Cyprinidae	Fathead minnow
Pimephales vigilax	Cyprinidae	Bullhead minnow
Prosopium williamsoni	Salmonidae	Mountain whitefish
Ptychocheilus oregonensis	Cyprinidae	Northern pikeminnow
Rhinichthys cataractae	Cyprinidae	Longnose dace
Richardsonius balteatus	Cyprinidae	Redside shiner

Figure A.1. An example of Receiver-Operating-Characteristic (ROC) curve (darker curved line) for logistic model for New River shiner. Each point on the ROC curve correspond to the sensitivity and specificity given a discrimination threshold on the predicted probability of species presence. AUC is calculated by integrating the area under the ROC curve. Note that the X-axis starts from 1 and ends at 0. The AUC value in this example is 0.731.

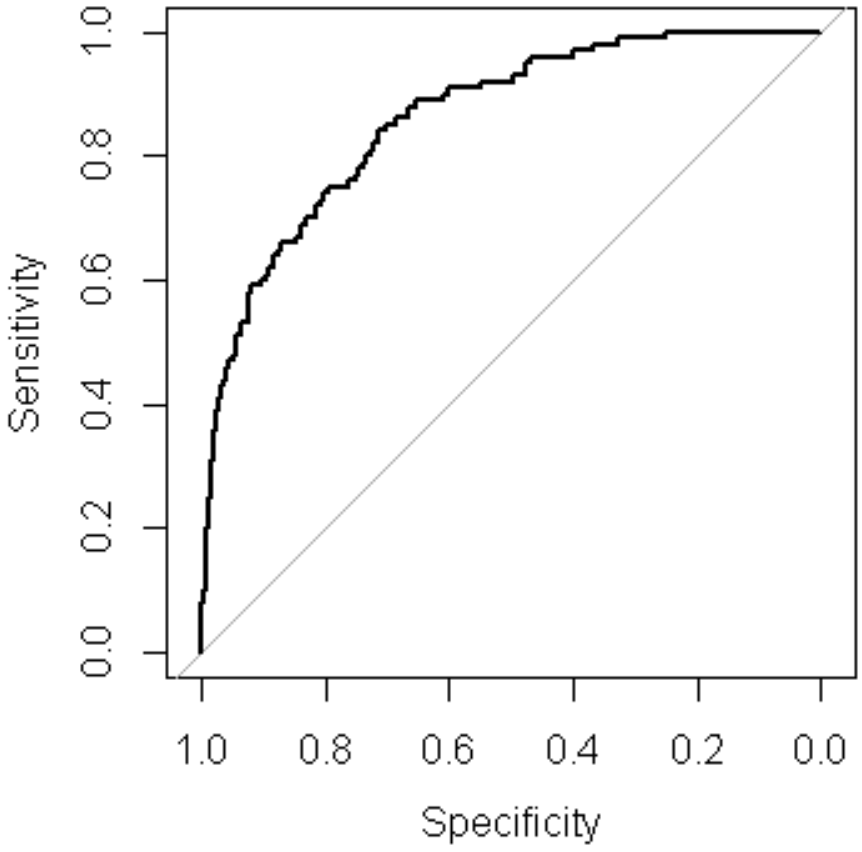


Table A.2. A table summarizing the Tukey's test (Tukey 1949) after the analysis of variance that evaluate the effects on the performance (AUC) of species distribution models. The three model types compared are logistic model (LM), boosted regression trees (BRT), and MaxEnt models. The descriptions of the rarity types A-H are provided in Table 2.1.

Treatments compared	Difference	Lower bound	Upper bound	*p*-value
Model types				
LM-BRT	-0.010	-0.034	0.014	0.579
MaxEnt-BRT	-0.136	-0.165	-0.107	< 0.001
MaxEnt-LM	-0.126	-0.154	-0.098	< 0.001
Incorporation of spatial autocorrelation				
Yes-No	0.020	0.001	0.038	0.037
Data resolution				
NHD-HUC	0.017	-0.005	0.038	0.123
Rarity types				
B-A	0.092	0.039	0.146	< 0.001
C-A	0.038	-0.002	0.079	0.074
D-A	0.031	-0.043	0.106	0.905
E-A	0.021	-0.028	0.069	0.901
F-A	0.026	-0.043	0.095	0.945
G-A	-0.009	-0.126	0.108	1.000
H-A	-0.010	-0.079	0.059	1.000
C-B	-0.054	-0.116	0.008	0.140
D-B	-0.061	-0.149	0.027	0.408
E-B	-0.072	-0.139	-0.004	0.029
F-B	-0.066	-0.150	0.017	0.239
G-B	-0.101	-0.227	0.024	0.219
H-B	-0.102	-0.185	-0.018	0.005
D-C	-0.007	-0.088	0.073	1.000
E-C	-0.018	-0.075	0.040	0.982
F-C	-0.012	-0.088	0.064	1.000
G-C	-0.048	-0.169	0.073	0.932
H-C	-0.048	-0.124	0.028	0.529
E-D	-0.011	-0.096	0.074	1.000
F-D	-0.005	-0.103	0.093	1.000
G-D	-0.041	-0.177	0.096	0.985
H-D	-0.041	-0.139	0.057	0.909
F-E	0.006	-0.075	0.086	1.000
G-E	-0.030	-0.154	0.094	0.996
H-E	-0.030	-0.111	0.050	0.946
G-F	-0.035	-0.169	0.098	0.993
H-F	-0.036	-0.130	0.058	0.943
H-G	0.000	-0.134	0.133	1.000

Figure A.2. A figure showing the correlation of the model performance of boosted regression tree models in terms of AUC (the area under the ROC curve) and the observed prevalence of stream fish species in the four selected basins (i.e., New River, Illinois River, Brazos River, and Snake River). The observed prevalence is the proportion of presence observations in all observations. This nonlinear negative correlation suggests that the habitat requirements and spatial distributions of more common species tend to be difficult to model.

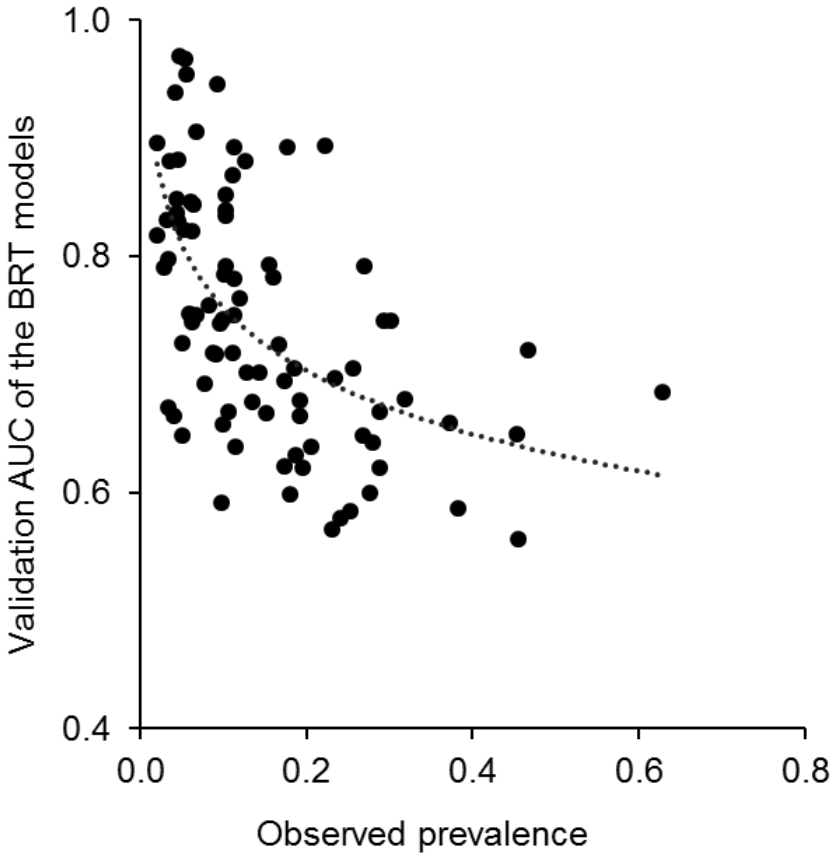


Table A.3. Summary on the key habitat factors for each of the 76 stream fish species in four river basins (i.e., BR-Brazos River, IL-Illinois River, NR-New River, SN-Snake River) in the non-spatial boosted regression tree (BRT) models. The number in the bracket was the measure of variable importance or percentage of contribution by the BRT models (Hijmans et al. 2011). The response of the each species on important habitat factors were evaluated using the partial dependence plots (Hijmans et al. 2011). We use "⬏" to represent strong positive relationship, and "↴" for negative relationship, and "~" for complicated non-linear relationships (e.g., polynomial, hinge). For instance, "TM (39.3) ⬏" means that annual mean temperature (TM) was the most important and positive environmental variable that contributed 39.3% in the BRT model for Chiselmouth (*Acrocheilus alutaceus*) in the Snake River basin. The descriptions of environmental predictors are listed in the Table 2.2.

Species	Basin	Key predictors				
		V1	V2	V3	V4	V5
Acrocheilus alutaceus	SN	TM (39.3) ⬏	TMI (9.6) ~	PPT (8.1) ~	C_WT (7.8) ~	BFI (7.5) ↴
Ameiurus natalis	BR	PPT (10.5) ⬏	HCI (9.2) ~	BFI (8.8) ⬏	SINU (7.6) ↴	DRA (7.3) ~
Aphredoderus sayanus	IL	TMI (16) ⬏	C_FR (8.7) ~	SLP (8.2) ~	SINU (7.2) ~	PPT (6.9) ⬏
Campostoma anomalum	NR	BFI (68.6) ~	TMI (7) ↴	C_FR (4.6) ↴	SLP (3.1) ⬏	ELE (2.2) ↴
Campostoma anomalum	BR	BFI (19.3) ⬏	PPT (17.2) ~	ELE (13.6) ~	SLP (9.3) ⬏	DRA (7.1) ~
Catostomus columbianus	SN	BFI (21.9) ↴	SLP (12.6) ↴	C_AG (8.3) ~	SO (6.5) ↴	MFU (6.3) ~
Catostomus commersonii	NR	BFI (40) ~	NT (11.3) ~	HCI (8.8) ↴	PT (8.2) ~	TMI (4.7) ~
Catostomus commersonii	IL	TMI (14.1) ~	TM (9.1) ⬏	MVU (7.1) ~	DRA (6.4) ~	SINU (6.3) ⬏
Cottus bairdii	NR	SLP (9.5) ↴	C_WT (9.3) ~	MVU (8.7) ~	NT (8.4) ~	DRA (6.9) ~
Cottus bairdii	SN	BFI (21.1) ⬏	PPT (16.2) ⬏	D_FR (10.1) ⬏	HCI (7) ~	TMI (6.4) ↴
Cottus kanawhae	NR	TMI (39) ↴	TM (20.9) ↴	PPT (14.6) ↴	ELE (5.2) ↴	TMA (4.8) ⬏
Cottus confusus	SN	TMA (24.7) ~	TMI (17.2) ↴	D_FR (11.4) ~	C_FR (7) ⬏	HCI (5.9) ~
Cyprinella galactura	NR	TMI (15.4) ⬏	PPT (12.8) ~	DRA (11.8) ⬏	SINU (10) ⬏	BFI (7.8) ↴
Cyprinella lutrensis	IL	TMA (35.7) ⬏	TM (20.8) ⬏	POP (5.9) ~	BFI (4) ↴	SO (3.6) ⬏
Cyprinella venusta	BR	BFI (35.1) ⬏	PPT (18.1) ~	SINU (5.4) ↴	MVU (4.8) ⬏	DRA (4.7) ~
Dorosoma petenense	BR	DRA (27.2) ⬏	MFU (23.3) ⬏	BFI (9.7) ↴	C_WT (9.3) ⬏	C_AG (3.2) ~
Etheostoma blennioides	NR	DRA (31.2) ~	BFI (12.6) ↴	MFU (12.3) ~	TMI (10.3) ↴	D_FR (4.2) ⬏

Etheostoma caeruleum	NR	SLP (33.7) ⮥	ELE (21) ↴	BFI (10.5) ↴	PPT (5.7) ↴	HCI (5.6) ↴
Etheostoma exile	IL	BFI (20.5) ⮥	C_WT (17.8) ~	TMI (9.3) ↴	C_AG (8.6) ↴	D_AG (6.1) ↴
Etheostoma nigrum	IL	ELE (21) ⮥	NT (11.5) ~	PT (10.2) ~	BFI (7.1) ⮥	POP (5.1) ~
Fundulus notatus	IL	SLP (18.4) ~	PT (15.3) ~	NT (13.8) ~	TMI (7.2) ⮥	ELE (6.8) ↴
Gambusia affinis	BR	C_FR (15.7) ~	SLP (10.5) ⮥	C_AG (9.2) ~	MVU (9) ⮥	NT (6) ~
Hypentelium nigricans	NR	BFI (44.6) ↴	PT (8.3) ~	POP (7.3) ~	DRA (5.4) ~	C_FR (4.8) ⮥
Ictalurus punctatus	BR	MFU (19.9) ⮥	DRA (13.3) ⮥	MVU (8.7) ⮥	PT (6.6) ⮥	ELE (6.1) ~
Lepisosteus osseus	BR	D_AG (13.2) ~	MFU (11.4) ~	HCI (10.6) ↴	POP (9) ~	DRA (8.7) ↴
Lepomis humilis	BR	ELE (33.3) ~	BFI (11.7) ↴	C_WT (7.7) ~	D_FR (6.3) ⮥	MFU (4.3) ⮥
Lepomis megalotis	BR	POP (19.5) ~	MVU (10.9) ~	D_FR (8.9) ⮥	TM (7.9) ⮥	SLP (7.6) ~
Luxilus chrysocephalus	IL	D_FR (23.5) ~	BFI (19.9) ~	HCI (12.1) ↴	TMA (9) ~	TM (4.7) ⮥
Luxilus chrysocephalus	NR	TMI (52.9) ↴	TMA (9.7) ⮥	DRA (6.8) ⮥	MFU (6.6) ⮥	C_UB (5.7) ~
Luxilus cornutus	IL	TM (41.1) ↴	TMI (17.3) ↴	POP (6.9) ⮥	PPT (4.1) ↴	D_AG (3.7) ⮥
Menidia beryllina	BR	DRA (18.6) ⮥	RDX (16.3) ~	C_WT (11) ⮥	POP (9.5) ~	MVU (9) ⮥
Nocomis biguttatus	IL	D_AG (23) ~	SLP (9.2) ~	PT (9.2) ~	BFI (8.5) ~	NT (6.5) ~
Nocomis leptocephalus	NR	BFI (15.8) ⮥	SLP (12.9) ~	PPT (12.5) ↴	ELE (7) ↴	TM (6.8) ⮥
Notropis atherinoides	IL	ELE (36.5) ↴	TMA (10.3) ↴	D_AG (8.1) ↴	C_AG (4.7) ↴	POP (4.6) ⮥
Notropis dorsalis	IL	D_AG (11.2) ⮥	TM (9.7) ⮥	PPT (7.7) ↴	MFU (7) ~	TMA (6.2) ⮥
Notropis hudsonius	IL	ELE (22.4) ↴	DRA (9.7) ⮥	C_WT (8.9) ~	TMI (6.7) ↴	POP (5.5) ~
Notropis rubellus	NR	BFI (19.8) ↴	MFU (15.2) ~	TMI (12.3) ↴	TM (12.1) ~	PPT (8) ⮥
Notropis stramineus	IL	MFU (13.4) ~	PT (11.9) ~	POP (10.6) ~	NT (9.4) ~	ELE (8.4) ↴
Notropis volucellus	NR	MFU (18.9) ⮥	DRA (15.5) ⮥	BFI (14.1) ↴	MVU (8.2) ⮥	TMI (7.6) ↴
Noturus gyrinus	BR	ELE (24) ↴	MVU (7.7) ⮥	TM (7.4) ⮥	TMI (6.5) ⮥	D_AG (5.8) ⮥
Pimephales notatus	NR	BFI (34.4) ~	MFU (14.2) ~	SLP (7.6) ⮥	HCI (5.6) ↴	SINU (5.3) ⮥
Pimephales promelas	BR	RDX (13.5) ⮥	POP (11.8) ~	C_UB (8.8) ⮥	TMI (8.5) ↴	PPT (8.5) ⮥
Pimephales vigilax	BR	MFU (14.1) ~	TMA (13.9) ⮥	DRA (11) ~	BFI (8.6) ↴	ELE (6.9) ↴
Prosopium williamsoni	SN	C_UB (48.3) ~	RDX (12.8) ↴	TMA (7) ↴	C_AG (5.6) ↴	BFI (3.8) ⮥
Ptychocheilus oregonensis	SN	BFI (53.8) ~	ELE (9.2) ↴	SLP (6.1) ↴	C_AG (5.5) ⮥	POP (3.8) ~

Rhinichthys cataractae	SN	SINU (42.4) ⬏	C_AG (9.9) ↷	BFI (9.7) ⬏	DRA (7.9) ~	RDX (4.8) ⬏
Rhinichthys cataractae	NR	BFI (46.1) ~	ELE (7.1) ↷	C_AG (6.1) ⬏	RDX (6.1) ~	MVU (5.7) ⬏
Richardsonius balteatus	SN	C_FR (16.6) ~	SLP (16.1) ↷	TMA (10.8) ~	NT (8.9) ~	D_FR (8.3) ↷
Amia calva	IL	ELE (17.9) ~	TM (17.8) ↷	C_UB (10.8) ~	C_WT (9.7) ~	BFI (6.3) ~
Etheostoma microperca	IL	BFI (42.5) ⬏	PPT (8.2) ⬏	TMA (8.1) ↷	NT (7.6) ~	PT (5.9) ~
Lythrurus ardens	NR	PPT (17.8) ↷	ELE (12.1) ↷	PT (12) ~	BFI (11.7) ↷	NT (7.5) ~
Notropis buccatus	IL	ELE (17.2) ↷	TMI (15.3) ⬏	NT (12.3) ~	TM (11.5) ↷	TMA (8.4) ↷
Notropis buccatus	NR	BFI (50.9) ~	SLP (13) ⬏	C_UB (6.8) ~	TM (5.7) ⬏	DRA (3.6) ⬏
Opsopoeodus emiliae	IL	MFU (18.6) ⬏	ELE (12.9) ↷	C_WT (10.3) ~	TMA (6.9) ~	TM (5.5) ~
Opsopoeodus emiliae	BR	ELE (43.1) ~	POP (7.5) ~	PPT (7.4) ⬏	SLP (5.7) ⬏	HCI (5.1) ↷
Campostoma oligolepis	IL	ELE (19.9) ~	TMI (11.7) ↷	NT (10.4) ~	BFI (8.4) ⬏	D_AG (8.3) ⬏
Carpiodes velifer	IL	D_FR (16) ~	TMA (14.5) ⬏	MVU (11.6) ⬏	C_UB (7.2) ⬏	TMI (5.1) ⬏
Cottus beldingii	SN	PPT (30.3) ⬏	SLP (10.6) ⬏	D_FR (7.8) ↷	BFI (7.6) ~	MVU (6.4) ~
Cottus rhotheus	SN	SLP (46.8) ⬏	C_FR (22.5) ↷	TMA (6) ↷	NT (3.7) ~	HCI (3.1) ~
Cyprinella spiloptera	IL	TMA (13.4) ↷	BFI (10.4) ⬏	ELE (8.9) ↷	POP (8) ~	MVU (7.2) ⬏
Cyprinella spiloptera	NR	MFU (18.5) ~	BFI (14.2) ↷	DRA (13.1) ⬏	PPT (10.8) ↷	ELE (6.8) ↷
Etheostoma chlorosoma	IL	ELE (42.1) ↷	D_AG (9.4) ↷	TMI (9.2) ⬏	NT (7) ~	BFI (4.3) ~
Etheostoma chlorosoma	BR	PPT (46) ⬏	TMI (13.2) ⬏	SLP (5.7) ⬏	C_AG (5.2) ↷	C_FR (4.8) ↷
Etheostoma spectabile	IL	BFI (13.7) ~	PPT (10.5) ⬏	TMI (10) ↷	ELE (8.4) ⬏	SLP (7.9) ⬏
Etheostoma spectabile	BR	BFI (43.4) ⬏	TMA (17.1) ⬏	SLP (8.7) ⬏	C_FR (6.4) ~	SINU (4.5) ↷
Hybognathus nuchalis	IL	ELE (19.6) ↷	D_FR (13.4) ⬏	TMA (11) ⬏	SINU (9.5) ~	TM (8.5) ⬏
Ictiobus bubalus	IL	ELE (50.1) ↷	C_AG (9.1) ↷	DRA (6.2) ~	D_AG (4.9) ↷	HCI (3.1) ⬏
Notropis buchanani	BR	MFU (37.6) ⬏	POP (7) ~	MVU (6.8) ~	TM (4.9) ~	DRA (4.8) ⬏
Percina phoxocephala	IL	MFU (12.6) ~	PT (12.4) ~	SINU (9.2) ⬏	TMA (8.8) ⬏	DRA (7.6) ~
Chrosomus erythrogaster	IL	TM (20.1) ↷	TMI (11.8) ↷	BFI (8.9) ↷	C_FR (6.7) ⬏	TMA (6.6) ⬏
Etheostoma asprigene	IL	ELE (35.1) ↷	POP (10.7) ~	TMA (7) ⬏	C_FR (5.4) ~	D_AG (4.7) ↷
Etheostoma gracile	BR	PPT (32.9) ⬏	TMI (15.6) ⬏	C_AG (8.3) ↷	ELE (4.7) ~	DRA (4.7) ⬏
Percina sciera	BR	MFU (23.3) ~	PPT (12.8) ⬏	C_AG (7.7) ⬏	SLP (7.2) ⬏	ELE (6.7) ↷

Catostomus ardens	SN	BFI (31.9) ~	ELE (14.4) ~	C_AG (10.4) ⬏	POP (8.9) ~	D_AG (5) ⬏
Luxilus cerasinus	NR	ELE (12.4) ~	TMA (11.2) ⬏	SINU (10.9) ⬏	PPT (7.4) ~	MFU (7) ~
Nocomis platyrhynchus	NR	BFI (32.8) ~	DRA (21.1) ~	NT (5.3) ~	MFU (4.9) ~	MVU (3.3) ⬎
Notropis rubricroceus	NR	C_FR (27.6) ~	POP (17) ~	BFI (15.4) ⬏	PPT (8.6) ⬎	NT (4.6) ~
Percina oxyrhynchus	NR	MVU (21.9) ⬏	MFU (20.8) ⬏	DRA (13.3) ⬏	POP (7.7) ~	ELE (7.3) ⬎
Percina roanoka	NR	DRA (18.3) ⬏	MFU (12) ⬏	TMA (11.9) ⬎	BFI (7.9) ⬏	POP (7.4) ~
Chrosomus oreas	NR	D_AG (14.7) ⬏	SINU (10.6) ⬏	PPT (9.1) ⬎	TM (8.5) ⬏	BFI (8.3) ⬏
Etheostoma kanawhae	NR	SINU (18.9) ⬏	BFI (17.2) ⬏	POP (10.4) ~	C_AG (9.9) ~	PPT (9.5) ⬏
Etheostoma osburni	NR	BFI (27.5) ⬎	TMI (19.1) ⬎	D_FR (12.8) ~	PPT (6.2) ~	C_UB (4.1) ⬏
Notropis scabriceps	NR	TMI (25.3) ⬎	DRA (13.3) ~	BFI (13.1) ⬏	PPT (7.4) ⬏	PT (5.4) ~
Luxilus albeolus	NR	ELE (23) ⬎	BFI (13) ⬎	TMA (8.4) ⬏	PPT (6.6) ⬎	SINU (6.4) ⬏
Exoglossum laurae	NR	MVU (13.1) ⬎	POP (8.5) ~	TMI (8.3) ⬎	SINU (7.9) ~	PPT (6.9) ⬎
Phenacobius teretulus	NR	BFI (37.6) ⬏	DRA (17) ~	C_FR (8) ⬎	MVU (4.8) ~	NT (4.8) ~

Table A.4. Summary on the key environmental factors for each of the 76 stream fish species in four river basins (i.e., BR-Brazos River, IL-Illinois River, NR-New River, SN-Snake River) in the spatial boosted regression tree (BRT) models. In the spatial models, The principal coordinate analysis of neighbor matrices (PCNM; Borcard and Legendre 2002) was used to spatialize environmental variables. The number in the bracket is the measure of variable importance or percentage of contribution to the BRT models (Hijmans et al. 2011). The response of the each species on important environmental factors were evaluated using partial dependence plots (Hijmans et al. 2011). We use "⬏" to represent strong positive relationship, and "↴" for negative relationship, and "~" for complicated non-linear relationships (e.g., polynomial, hinge). The descriptions of environmental predictors are listed in the Table 2.2 A.

Species	Basin	Key predictors				
		V1	V2	V3	V4	V5
Acrocheilus alutaceus	SN	TM (19) ⬏	TMI (18.2) ⬏	ELE (11.2) ↴	D_UB (11.2) ↴	POP (9.5) ↴
Ameiurus natalis	BR	C_FR (22.6) ↴	ELE (21.5) ↴	PPT (7.3) ⬏	BFI (7.2) ⬏	DRA (7.1) ~
Aphredoderus sayanus	IL	TMI (34.6) ⬏	PPT (11.2) ⬏	MVU (9.6) ↴	C_WA (4.4) ~	PT (4) ⬏
Campostoma anomalum	NR	BFI (60) ↴	PPT (5) ↴	D_AG (4.3) ~	POP (3.2) ⬏	TMI (3) ↴
Campostoma anomalum	BR	BFI (31.6) ⬏	SLP (10) ⬏	PPT (9.4) ~	SO (8.7) ↴	RDX (5.5) ⬏
Catostomus columbianus	SN	C_GR (27.5) ⬏	SO (11.2) ↴	DRA (10.1) ↴	TMI (9.1) ⬏	BFI (8.3) ↴
Catostomus commersonii	NR	HCI (24.5) ↴	BFI (23.1) ↴	MVU (10.3) ↴	TMA (8.3) ~	ELE (4.8) ↴
Catostomus commersonii	IL	SINU (12.8) ⬏	HCI (9.1) ~	C_FR (7.7) ↴	ELE (7.3) ~	D_FR (6.7) ↴
Cottus bairdii	NR	C_WA (15.2) ↴	BFI (10.3) ⬏	PT (10) ⬏	TMI (7.6) ↴	C_FR (6.7) ~
Cottus bairdii	SN	C_GR (17.7) ↴	SLP (12.1) ⬏	BFI (9.7) ⬏	TM (9.6) ~	TMI (8.2) ↴
Cottus kanawhae	NR	TM (22.4) ↴	TMI (17.9) ↴	PPT (17) ↴	POP (12.7) ⬏	C_WA (7.6) ~
Cottus confusus	SN	D_AG (30.3) ↴	HCI (20.3) ⬏	TM (12.6) ↴	MVU (10.1) ⬏	C_GR (9.7) ~
Cyprinella galactura	NR	SINU (22.3) ⬏	PPT (19.9) ↴	TMI (11.4) ⬏	TM (8.2) ~	C_AG (6.7) ↴
Cyprinella lutrensis	IL	TMA (26) ⬏	TM (22.6) ⬏	SO (4.9) ⬏	POP (4.4) ~	BFI (4.4) ↴
Cyprinella venusta	BR	BFI (32.1) ⬏	PPT (12) ~	C_UB (5) ~	C_WA (4.8) ⬏	DRA (4) ⬏
Dorosoma petenense	BR	C_WA (17.3) ⬏	MFU (16.2) ⬏	DRA (11.8) ⬏	BFI (8.5) ↴	HCI (6.7) ↴
Etheostoma blennioides	NR	BFI (32.4) ↴	MVU (10.4) ⬏	ELE (8.3) ↴	NT (8) ↴	C_WA (5.1) ⬏

Etheostoma caeruleum	NR	SLP (19.3) ⬏	MFU (17.4) ⬏	TMA (9.6) ⬏	ELE (8.8) ↴	PPT (8.4) ↴
Etheostoma exile	IL	BFI (21.5) ⬏	D_AG (15) ↴	POP (11.2) ⬏	PPT (10.5) ↴	RDX (8.5) ⬏
Etheostoma nigrum	IL	ELE (29.2) ⬏	NT (16.5) ~	PT (11.2) ~	BFI (9.1) ⬏	POP (5.1) ~
Fundulus notatus	IL	SLP (24.7) ↴	TMA (12.4) ⬏	TMI (11.2) ⬏	D_FR (8.8) ↴	SINU (6.3) ⬏
Gambusia affinis	BR	C_FR (19.3) ↴	C_AG (18.1) ~	MVU (14) ⬏	ELE (7.9) ⬏	NT (7.3) ⬏
Hypentelium nigricans	NR	BFI (31) ↴	NT (14.8) ⬏	ELE (12.6) ↴	MVU (8) ⬏	PT (5.9) ⬏
Ictalurus punctatus	BR	SLP (14.7) ↴	C_AG (9.1) ⬏	PPT (8.1) ~	SO (7.6) ⬏	MVU (6.7) ~
LepiSOsteus osseus	BR	POP (17.4) ⬏	SLP (13) ↴	C_WA (10.1) ⬏	ELE (8.7) ↴	HCI (8.6) ↴
Lepomis humilis	BR	C_AG (23.2) ↴	ELE (15.7) ~	C_WA (12.1) ⬏	BFI (9.1) ↴	RDX (7.5) ↴
Lepomis megalotis	BR	PPT (14.1) ⬏	D_FR (13.2) ~	C_WA (9.5) ⬏	BFI (9.5) ⬏	TMI (8.8) ⬏
Luxilus chrysocephalus	IL	HCI (12.7) ↴	BFI (11.5) ~	D_FR (8.7) ↴	D_AG (8.6) ⬏	TMA (7.8) ~
Luxilus chrysocephalus	NR	TMI (40.9) ↴	TMA (9.7) ⬏	NT (6.9) ⬏	BFI (6.2) ↴	POP (6.1) ⬏
Luxilus cornutus	IL	TM (38.4) ↴	TMI (14.8) ↴	SINU (7.2) ↴	C_WA (5) ~	D_AG (4.5) ⬏
Menidia beryllina	BR	SO (21.3) ⬏	C_WA (19) ⬏	MVU (15.1) ⬏	MFU (5.6) ⬏	DRA (5.2) ⬏
Nocomis biguttatus	IL	D_FR (15.2) ↴	D_AG (10.5) ⬏	HCI (9.3) ↴	MVU (9.2) ⬏	SO (8.7) ↴
Nocomis leptocephalus	NR	BFI (17.2) ⬏	D_FR (10.2) ↴	SO (7.9) ~	MVU (7.9) ⬏	RDX (4.7) ~
Notropis atherinoides	IL	ELE (29.5) ↴	D_AG (13.3) ↴	MVU (5.5) ⬏	C_AG (5) ↴	TMA (4.6) ↴
Notropis dorsalis	IL	TM (14.4) ⬏	POP (9.3) ~	TMA (7.6) ⬏	PPT (6.9) ~	D_AG (6.5) ⬏
Notropis hudsonius	IL	ELE (12.5) ↴	DRA (10.6) ⬏	POP (7.3) ~	C_AG (7.1) ↴	RDX (6.1) ~
Notropis rubellus	NR	BFI (30.1) ↴	POP (8.9) ↴	MVU (8.3) ⬏	TMI (7.1) ↴	PPT (6.5) ~
Notropis stramineus	IL	C_UB (22.9) ↴	D_AG (19.6) ⬏	HCI (9.5) ⬏	C_FR (8.7) ↴	SINU (5.5) ⬏
Notropis volucellus	NR	ELE (20.9) ↴	C_WA (17.2) ⬏	POP (11) ↴	TMA (6.1) ~	MFU (5.2) ⬏
Noturus gyrinus	BR	ELE (27.1) ↴	TMI (14.8) ⬏	BFI (13.6) ↴	TM (7) ⬏	D_FR (6.9) ↴
Pimephales notatus	NR	BFI (50) ↴	ELE (8.8) ↴	PPT (6.2) ↴	TMI (3.7) ↴	POP (3.5) ⬏
Pimephales promelas	BR	TMA (19.5) ⬏	DRA (13.6) ↴	RDX (12.2) ⬏	BFI (6.2) ↴	POP (6.2) ⬏
Pimephales vigilax	BR	MVU (14.9) ⬏	TMA (9) ⬏	SINU (8.8) ⬏	MFU (7.4) ⬏	HCI (6.8) ~
Prosopium williamsoni	SN	D_AG (20.8) ↴	C_UB (15.5) ⬏	SINU (10.6) ⬏	C_FR (10.4) ↴	MVU (7.2) ⬏
Ptychocheilus oregonensis	SN	BFI (45.4) ↴	C_AG (9.6) ⬏	TMI (6.1) ⬏	TM (5.2) ⬏	MVU (5) ↴

Rhinichthys cataractae	SN	SINU (39.2) ⬏	C_AG (9.2) ↷	BFI (9.8) ⬏	DRA (6.9) ~	RDX (5.8) ⬏
Rhinichthys cataractae	NR	BFI (32.7) ↷	RDX (20) ↷	D_AG (8) ~	TM (5.2) ↷	MVU (4.6) ⬏
Richardsonius balteatus	SN	DRA (17.6) ⬏	C_GR (10.5) ⬏	HCI (9.3) ↷	BFI (9) ↷	C_FR (7) ↷
Amia calva	IL	TMA (35) ~	TM (13) ~	HCI (11) ⬏	BFI (8.1) ⬏	ELE (6.6) ~
Etheostoma microperca	IL	BFI (20) ⬏	NT (13.3) ⬏	RDX (13) ⬏	TMA (7.7) ↷	HCI (6.3) ↷
Lythrurus ardens	NR	PPT (18.1) ↷	SINU (11.5) ⬏	BFI (9.9) ~	TMA (8.1) ~	ELE (8) ↷
Notropis buccatus	IL	ELE (19.2) ↷	TMI (10.8) ~	POP (9.2) ~	NT (9) ⬏	TMA (5.9) ↷
Notropis buccatus	NR	BFI (25.7) ↷	ELE (12.8) ↷	SLP (8.5) ~	C_AG (8.2) ↷	C_FR (6.7) ⬏
Opsopoeodus emiliae	IL	ELE (25.1) ~	BFI (9.9) ⬏	MVU (9) ⬏	DRA (7.1) ⬏	POP (6.7) ~
Opsopoeodus emiliae	BR	TMI (30.9) ⬏	TM (19.2) ⬏	HCI (8.5) ↷	D_AG (6.5) ~	ELE (6.2) ↷
Campostoma oligolepis	IL	TMI (16.1) ~	TM (11.3) ~	SINU (8.5) ⬏	NT (7.3) ⬏	POP (6.8) ~
Carpiodes velifer	IL	TMI (17.6) ~	TMA (16.8) ⬏	SINU (11.1) ↷	C_UB (8) ⬏	MVU (7.7) ⬏
Cottus beldingii	SN	PPT (30.5) ⬏	SLP (18.7) ⬏	D_UB (10.1) ~	RDX (9.2) ↷	MFU (3.9) ⬏
Cottus rhotheus	SN	C_UB (36.4) ↷	D_AG (18.1) ↷	DRA (12.3) ↷	PPT (9.8) ~	POP (5.9) ↷
Cyprinella spiloptera	IL	TM (10.6) ↷	TMA (9.9) ↷	SINU (9.3) ~	MVU (7.3) ⬏	PPT (7.3) ↷
Cyprinella spiloptera	NR	ELE (17.4) ↷	MFU (16.8) ⬏	RDX (8.1) ⬏	POP (7.7) ↷	BFI (7.3) ~
Etheostoma chlorosoma	IL	C_WA (14.8) ⬏	ELE (14.3) ↷	MVU (8.7) ↷	D_AG (7.4) ↷	TM (7.3) ↷
Etheostoma chlorosoma	BR	PPT (43.4) ⬏	C_AG (6.8) ↷	TMA (6.4) ↷	D_FR (6.1) ⬏	POP (4.8) ↷
Etheostoma spectabile	IL	MVU (9.4) ↷	BFI (9) ↷	POP (8.7) ~	SLP (7.6) ⬏	TMI (6.3) ↷
Etheostoma spectabile	BR	BFI (42.5) ⬏	TMA (7.7) ⬏	SLP (7.5) ⬏	C_UB (5.9) ⬏	PPT (5.2) ~
Hybognathus nuchalis	IL	ELE (17) ↷	TMA (12.2) ⬏	TM (11.7) ⬏	DRA (10.2) ⬏	MVU (8.7) ⬏
Ictiobus bubalus	IL	ELE (29.8) ↷	C_WA (10.8) ⬏	DRA (6.2) ⬏	NT (6.1) ⬏	MFU (5.6) ⬏
Notropis buchanani	BR	MVU (15.5) ~	MFU (15.4) ⬏	C_WA (10.6) ↷	BFI (8.8) ↷	SO (6.3) ↷
Percina phoxocephala	IL	TMA (19.3) ⬏	SINU (9.2) ⬏	SLP (8.3) ~	C_FR (7) ↷	D_AG (6.6) ⬏
Chrosomus erythrogaster	IL	TMI (14) ↷	NT (11.3) ↷	RDX (10) ⬏	TM (9.1) ↷	TMA (7.9) ⬏
Etheostoma asprigene	IL	ELE (33.7) ↷	MVU (7.5) ⬏	C_WA (5.4) ~	D_FR (4.5) ~	D_AG (4.5) ↷
Etheostoma gracile	BR	PPT (39.7) ⬏	TMI (16.6) ⬏	D_AG (6.3) ~	POP (5.3) ↷	BFI (4.5) ↷
Percina sciera	BR	C_FR (11.5) ~	MVU (11.4) ⬏	C_UB (10) ~	SINU (9.1) ⬏	PPT (7.9) ⬏

Catostomus ardens	SN	C_AG (33.8) ↗	BFI (27.7) ~	POP (8.3) ↗	C_UB (6.5) ↗	TMI (6.3) ↴
Luxilus cerasinus	NR	BFI (17.2) ~	TMA (13.2) ~	D_AG (10.7) ↗	TM (9.7) ↴	C_UB (8.9) ↴
Nocomis platyrhynchus	NR	BFI (17.1) ↴	MVU (10) ~	TMI (6.4) ↴	PPT (6.3) ~	SO (6.1) ↗
Notropis rubricroceus	NR	SINU (22.1) ↴	BFI (14) ↗	D_AG (8.6) ↗	C_UB (6.8) ↴	RDX (6.6) ↴
Percina oxyrhynchus	NR	MVU (20.6) ↗	MFU (11.8) ↗	ELE (7) ↴	SINU (5.8) ↴	DRA (5) ↗
Percina roanoka	NR	MVU (15.5) ↗	DRA (10.2) ↗	BFI (9.9) ↗	ELE (9.5) ~	TMI (6) ↗
Chrosomus oreas	NR	D_FR (12.1) ↴	TMI (11.1) ~	C_FR (8.3) ↴	D_AG (7.2) ↗	SO (6.5) ~
Etheostoma kanawhae	NR	BFI (28.3) ↗	SO (12) ~	SLP (6.9) ~	SINU (6.3) ↗	MVU (5.8) ↴
Etheostoma osburni	NR	TMI (17.3) ↴	D_FR (9.7) ↗	HCI (8.9) ↗	BFI (7.3) ↴	TM (6.5) ↴
Notropis scabriceps	NR	TMI (41.4) ↴	BFI (21.1) ↗	PPT (4.9) ↗	TM (4.4) ↴	SLP (4) ↗
Luxilus albeolus	NR	TMA (15.8) ↗	MFU (14.8) ↗	TM (12.4) ↗	DRA (9.6) ↗	C_UB (6.2) ~
Exoglossum laurae	NR	TM (13.3) ↴	C_WA (9.6) ↴	TMA (7.3) ~	NT (6.6) ↗	TMI (6.3) ↴
Phenacobius teretulus	NR	BFI (35.1) ↗	C_AG (10.2) ↗	TMI (5.5) ↗	C_FR (5.1) ~	TM (5) ↴

Chapter 3: Limited transferability of stream–fish distribution models among river basins: reasons and implications

Jian Huang, Emmanuel A. Frimpong*

Department of Fish and Wildlife Conservation, Virginia Polytechnic Institute and State University, 100 Cheatham Hall, Blacksburg, VA 24061, USA

* Corresponding author; E-mail: frimp@vt.edu; Tel.: +1-540-231-6880; Fax: +1-540-231- 7580

Abstract: Spatial transferance of species distribution models is often applied in the study of land use and climate change impacts, spread of invasive species, and conservation planning. However, model transferability and risk of error are rarely evaluated prior to predicting species distribution to different regions or time frames. In this study, we developed distribution models for 21 fish species and made predictions of occurrence of these species in another river basin to assess model transferability and to evaluate the effect of habitat heterogeneity and model types on transferability. In addition to internal and external evaluation of model performance based on the area under the receiver–operating–characteristic curve, we assessed the cross–basin consistency of variable selections and fish–habitat relationships. The transferability of all three models (logistic regression, boosted regression trees, and MaxEnt model) was limited in terms of both prediction accuracy and cross–basin consistence of fish–habitat relationships for over 70% of the species examined. Model transferability could be enhanced by 1) using simple but robust algorithms such as logistic regression under Lasso regularization, 2) including only direct habitat features with a sound ecological basis (e.g., temperature and hydrology), 3) incorporating spatial

autocorrelation in model training, and 4) matching the range and location of the habitat predictors between the model region and prediction region.

INTRODUCTION

Spatial distribution of a species is dynamic, determined by abiotic factors and biological interactions at different spatial and temporal scales (Maurer and Taper 2002, Franklin and Miller 2009). Earlier species distribution models (SDMs) have primarily sought to understand correlative species–habitat relationships that underpin the dynamic process of distribution, and to present distribution patterns with maps (Mac Nally 2000). Advances in geographic information systems, remote sensing and monitoring techniques have made massive volumes of environmental data available, while species occurrence data (presence and absence) remain a constraint in species distribution models. Both the processes of collecting biological samples and integrating data from different sources from a large geographic extent can be very time– and labor–consuming (Frimpong, Huang, Liang, and Ostroff, manuscript in review; available on request). This shortage in data of biological responses (e.g., occurrence, abundance, density) and the need for present conservation decisions to be extended into the future has necessitated other major usage of species distribution models– making spatial and temporal predictions. Model predictions are made to delineate suitable habitats for natural resource conservation (Fielding and Haworth 1995, Guay et al. 2003, Murray et al. 2011, Martin et al. 2012), to forecast potential range of exotic species (Jones 2012, Wang and Jackson 2014), or to project species distribution in future climate scenarios (Chu et al. 2005, Lyons et al. 2010, Comte and Grenouillet 2013, Schibalski et al. 2014).

Depending on the similarity of environmental space in the training and prediction dataset, model predictions are classified into two groups, interpolation and extrapolation (Wiener 1949). Interpolations are applied when there is a need to predict species distribution within the range of environments sampled, while extrapolations predict species distribution in ranges of environments that do not fully overlap the sampled range (Elith and Leathwick 2009). Transferring models in this study refers to predicting species distribution into new geographic regions, which could be considered as a special case of extrapolation. Transferring models to a different region or future scenarios entail more risk of error, compared to interpolation (Peterson et al. 2007). The predictions could be unreliable if the environmental gradients and species–habitat relationships vary spatially or temporally. There is increasing agreement among species distribution modelers on the need to evaluate model transferability or generality, namely the ability, as measured by accuracy and precision, to predict species distribution in a different region or time frame (Randin et al. 2006, Elith and Leathwick 2009, Wenger and Olden 2012). With more of such evaluations, researchers will become aware of the risks in model transfer and take remedial actions in the case of limited transferability.

Four criteria have been applied to evaluate transferability of SDMs: 1) The model based on training data fit well internally (e.g., Randin et al. 2006, Schibalski et al. 2014). 2) The fitted model performs well in the cross validation. Researchers usually use random 3–10 fold cross validation, but Wenger and Olden (2012) recommended the use of non–random cross validation when no independent data is available for external evaluation. In the non–random cross validation, the data are stratified geographically into fixed subsets representing different sub–regions rather than random samples. Random samples always yield unbiased estimate of the overall species–habitat relationship but obscure spatial heterogeneity. 3) The fitted model

performs well in external evaluation, namely predicting accurately with spatially or temporally independent data (Randin et al. 2006, Murray et al. 2011, Wang and Jackson 2014); 4) A stricter criterion for successful model transferability will require that predictor selection, and species–habitat relationships described are consistent over space or time (Schibalski et al. 2014). We suggest using criterion #2, #3, and #4, because #1 can be an artifact of variable selections and species prevalence in the training data (Huang and Frimpong, manuscript under review; available on request from EAF).

Most studies on SDMs transferability use the area under the receiver–operating–characteristic curve (AUC) and its varieties in both the internal and external evaluations (e.g., Randin et al. 2006, Peterson et al. 2007, Murray et al. 2011, Tuanmu et al. 2011, Wenger and Olden 2012, Wang and Jackson 2014). The good features of AUC include 1) it measures model performance over the entire range of error costs (Hanley and McNeil 1982), 2) it does not require arbitrary or *a prior* discrimination thresholds, 3) it is invariant to *a priori* probability distributions of the responses (Bradley 1997); 4) it can be conveniently calculated in nearly all types of models, and 5) it is statistically more discriminating and more consistent than threshold–dependent measures such as overall correct classification rate (Ling et al 2003). When it comes to classifying binary response, researchers in fields of machine–learning, biogeography, medical and psychology favor to use AUC as a single–value criterion on model performance. It is worth noting that AUC measures discrimination power rather than the goodness of fit, and it could not substitute sensitivity and specificity when the costs of omission error and commission error are unequal (Lobo et al. 2008).

Transferring models over space or time is a widely recognized challenge because species distributions can be affected by abiotic and biotic factors which are usually difficult to analyze

without adequate experiment designs (Randin et al. 2006, Peterson et al. 2007, Elith and Leathwich 2009, Barbosa et al. 2009). Many researchers attribute limited transferability to the heterogeneity of habitats among regions (Brown and Lomolino 1998, Ervin and Holly 2011). Habitat heterogeneity increases with spatial or temporal extents in both terrestrial and aquatic ecosystems (Frimpong et al. 2005). Factors such as climate, landform, and geology are widely used in SDMs, but these factors may represent latitudinal or coast–inland gradients, causing problems in predictions particularly at large scales (Murray et al. 2011). Another constraint on the transferability of distribution models is the mismatches in the range and location of the environmental gradients in the training and prediction dataset (Jackson et al. 2001, Randin et al. 2006, Murray et al. 2011). Models based on a part of the species' range may yield only regionally applicable habitat relationships, so they have limited predictive power in extrapolations (Segurado and Araújo 2004, Arntzen 2006). Additionally, different habitat features or their interactions may constrain the spatial distribution of a species in different regions (Ervin and Holly 2011). For example, the spatial distribution of a trout species may be controlled by temperature in pristine streams, but its distribution could be more affected by water quality in a disturbed area.

Biotic factors affecting transferability of SDMs include 1) plasticity of the focal species' ecological traits (Bulluck et al. 2006, Randin et al. 2006), 2) the equilibrium status of dispersal (Elith and Leathwick 2009), and 3) the prevalence of the species (Barbosa et al. 2009, Wang and Jackson 2014). The species–habitat relationships described in one region could not be adequately generalized to another region if the species' ecological traits are too plastic, leading to failures in model transfer. Oppositely, good model transferability tends to be obtained for rare species with high habitat specificity (e.g., Murray et al. 2011, Tuanmu et al. 2011, Martin et al. 2012).

Products of most species distribution models are estimated optimal habitats without accounting for the connectivity of the study regions or the equilibrium status of spread; thus, poor transferability may occur for taxa whose dispersal into optimal habitats are obstructed by physical barriers. This may partially explain why the model transferability of plants (e.g., Strauss and Biedermann 2007) and terrestrial animals (e.g., Martin et al. 2012, Tuanmu et al. 2011) are usually higher than the model transferability of stream fish species (e.g., Wenger and Olden 2012) that are often constrained by watershed boundaries.

Additionally, model transferability can be influenced by the choice of modeling approaches (Meynard and Quinn 2007, Peterson et al. 2007), the quality of data used for model fitting and prediction (Barbosa et al. 2009, Wang and Jackson 2014) and variable selections (Martin et al. 2012). A few studies (e.g., Meynard and Quinn 2007, Wenger and Olden 2012) have shown that generalized linear models (GLM) have more robust transferability than machine learning models, suggesting that over–fitting may limit generality in the more complicated machine–learning models. Other studies that assessed model transferability (e.g., Randin et al. 2006, Barbosa et al. 2009, Wang and Jackson 2014) found that models based on data of larger sample size tend to be more transferable than those based on limited samples.

In this study, we evaluate the spatial transferability of species distribution models for 21 fish species in five river basins in the eastern and central United States to test the hypothesis that transferability of SDMs would be affected by model type and habitat heterogeneity. We predicted that good transferability would be obtained in models with simple structure and when the environmental gradients in the model region and prediction region are similar, although such pattern may be confounded by other factors (e.g., data quality, plasticity of species traits) as reviewed previously. We compared the transferability of logistic models, boosted regression

trees, and MaxEnt models, which are three widely used modeling techniques. In addition to internal and external evaluation based on AUC, we assessed the cross–basin consistency of variable selections in the boosted regression trees, and compared the fish–habitat relationships described in different basins using partial dependence plots. We expected high transferability when the variable selection, variable importance ranks, and the species–habitat relationships are consistent between the training region and prediction region.

METHODS

Study basins and species

The 5 study basins are New River, Roanoke River, Illinois River, Brazos River, and Colorado (Texas) River (Figure 3.1). These basins were chosen because they contained sufficient data on species occurrence in the *IchthyMaps* historical fish distribution database used for this study (Frimpong, Huang, Liang, and Ostroff, manuscript under review; available on request from EAF). The Brazos River and Colorado River are adjacent and in the same HUC-2 region (Texas region). The New River and Roanoke River are adjacent but in two different HUC-2 regions, HUC-2 region 5 (Ohio region) and region 3 (South Atlantic North region) respectively. The Illinois River is far away from the other four basins. The overall environmental conditions of Brazos River and Colorado River are similar, while the New River and Illinois River differ the most in terms of climate, landscape, and hydrology (Table 3.1). Including both adjacent and separated basins and paired basins with varied similarity in environmental conditions allow conclusions that are more general.

We used three criteria to select the study species: 1) the candidate species needs to be distributed over multiple study river basins so that we can assess across–basin model transferability; 2) non-native species were excluded since their dispersal may not be in

equilibrium; and 3) the selected species should have at least 30 occurrence records (a sufficient size needed for reliable inference) in each of the model basins. In total, 21 fish species satisfying these criteria were selected. The presence and absence of these 21 species were derived from the *IchthyMaps* database that contains fish occurrence records sampled primarily during 1950 to 1990 in the United States.

Developing species distribution models

We selected to compare the transferability of logistic regression under the Lasso regularization (Tibshirani 1996), boosted regression trees (BRT) model (Friedman 2001), and maximum entropy presence–only model (MaxEnt; Phillips et al. 2006). Logistic regression is a generalized linear model that is suited to binary responses. The Lasso (least absolute shrinkage and selection operator) adds a penalty term (absolute sum of coefficients) to the negative log–likelihood in generalized linear models (Friedman et al. 2010). Increasing the penalty in Lasso will force more model coefficients to be zero in the optimization of the constrained likelihood function, so that Lasso could mitigate the effect of multicollinearity and control model complexity (Tibshirani 1996). The regularization parameter of the absolute shrinkage is usually tuned in the cross validations. For example, we chose the value of regularization parameter that maximized the AUC in the 5–fold cross validation. When the regularization parameter is set to 0, Lasso would not alter the logistic regression at all. We implemented the Lasso–version logistic models in the R program (R core team 2014) with the package ‘glmnet’ (Friedman et al. 2010). The second modeling technique, boosted regression tree (BRT), was developed by Friedman (2001) and introduced to studies of species distribution models by Elith et al. (2006). Boosting algorithm ensembles individual simple classifiers (e.g., small classification or regression trees) by iteratively weighting individual trees and observations. A classifier with low misclassification

rate and misclassified observations in the cross validation are assigned more weight. The number of classifiers was determined through minimizing a loss function and final predictions were made by summing weighted classifiers. The third method, MaxEnt, searches the spatial species distribution that maximizes entropy under the constraints of the values of abiotic features (Phillips et al. 2006). MaxEnt models are essentially fitted by maximizing the likelihood of a statistical substitute of entropy under the Lasso regularization (Phillips et al. 2006). We used the absences inferred from the *IchthyMaps* database instead of pseudo absences randomly drawn from the backgrounds, which are expected to reduce the false absence rate in the model. The boosted regression trees and MaxEnt models were batched for the 21 fish species in the R package 'dismo' (Hijmans et al. 2013). The predictor variables considered in these models are in the categories of climate, landscape, geology, hydrology, stream morphology, disturbance and water chemistry. Twenty–three variables were kept after removing some highly correlated variables based on the correlation matrix. Information (e.g., source, description, and range) on the predictor variables are listed in the Table 3.1. Habitat condition index measured the cumulative disturbance of catchment of inter-confluence stream segments based on 15 disturbance variables such as land use composition, human population, dam density, road density, and point-source pollution (Esselman et al. 2011). The influence of each disturbance variable was weighted by the results of multiple linear regression of all variables against a commonly used biological indicators of habitat condition (i.e., percent intolerant fishes at a site). The most heavily weighted disturbance include urban lands, point-source pollution, pasture lands, and dam densities.

Internal evaluation

The three models were validated internally in the New River and Brazos River basins by 5–fold cross validation with the criterion of the area under the receiver–operating–characteristic (ROC) curve (AUC). In the 5–fold cross validation, the whole data are randomly split into 5 subsets of equal size, and each subset (1/5 of whole data) is treated as testing data once while the rest (4/5 of whole data) is used as training data. Five measures of AUC were obtained when predicting each testing data by the model based on training data, and the mean of these 5 measures of AUC is reported as the AUC in the internal evaluation.

Evaluation of transferability

The three models were evaluated externally in the Roanoke River, Colorado River, and Illinois River basins, a key process to assess spatial transferability of SDMs for stream fish. There are four sets of model transferance for each modeling approach: from New River to Roanoke River (NR–RR) and to Illinois River (NR–IR), from Brazos River to Colorado River (BR–CR) and to Illinois River (BR–IR). The transferability AUC was calculated with the R package pROC (Robin et al. 2011) based on the observed presence/absence and predictions from the models built in the New River and Brazos River (Table 3.2).

When interpreting AUC values, it is common (e.g., Araújo et al. 2005, Randin et al. 2006) to use the classification proposed by Swets (1988): excellent $AUC > 0.90$, good $0.80 < AUC < 0.90$, fair $0.70 < AUC < 0.80$, poor $0.60 < AUC < 0.70$; fail $0.50 < AUC < 0.60$. This classification was originally developed to evaluate the performance of models when diagnosing systems internally (i.e., model fitting), but it would seem too demanding in the case of external evaluation, considering the heterogeneity in habitat and uncertainty in species–habitat associations. Our observation is consistent with the previous studies on model transferability: predictive power is usually highest in the model fitting, followed by in cross validation, and then

in external evaluations of independent datasets (Olden et al. 2002, Strauss and Biedermann 2007, Murray et al. 2011, Heinänen et al. 2012, Wenger and Olden 2012, Wang and Jackson 2014). Summarizing these studies, we suggest using AUC < 0.6 as an index of limited or poor transferability, and reclassifying the AUC values in the external evaluations or model transfers, for example, excellent AUC > 0.80, good 0.70 < AUC < 0.80, fair 0.60 < AUC < 0.70, limited 0.50 < AUC < 0.60, poor 0.40 < AUC < 0.50.

If poor among–basins transferability occurred (AUC < 0.5), we developed spatial models with the principal coordinate analysis of neighbor matrices (PCNM; Borcard and Legendre 2002) to examine whether incorporating spatial autocorrelation could enhance transferability. We used PCNM packages (Legendre et al. 2012) in R program to extract mutually orthogonal eigenvectors from the Euclidean distance matrix among sampled stream segments in the model basin. These mutually orthogonal eigenvectors (i.e., spatial eigenvectors) associated with large standardized Moran's *I* (Moran 1950) values (empirically > 1.96) are kept to form the spatial matrix. We developed multivariate regression models with environmental variables as responses and spatial matrix as predictors, to filter the spatial structure in the environmental variables (Brind'Amour et al. 2005). We then used the predicted environmental variables as predictors in the spatial species distribution models. The cross validation AUC and transferability AUC were calculated for the spatial models, and compared with the non–spatial models.

We ranked the predictor variables by measures of relative contribution to each study fish species in the boosted regression trees. The BRT calculates the contribution of a variable as a function of the number of times the variable is selected for partitioning, weighted by the squared improvement contributed by each partition in all trees (Friedman and Meulman 2003, Hijmans et al. 2013). The contributions of variables are scaled to make the sum of relative contributions

equal 100% in the R package ‘dismo’ (Hijmans et al. 2013). To measure the consistency of variable selection among basins, we counted the number of variables ranked in the top 10 in both of the paired basins (Table B.1).

This measure of consistency in variable selection may be biased because a variable, despite its high rank of relative importance in two basins, may affect the species’ distribution in different ways. Therefore, we built partial dependence plots to visualize the response of a species to each predictor variable after marginalizing the effects of all other predictors in the boosted regression trees. Partial dependence plot is widely used in machine learning techniques, including random forest (Breiman 2001), boosted regression trees (Friedman 2001), and MaxEnt (Phillips et al. 2006).

RESULTS

The three modeling approaches (GLM, BRT, and MaxEnt) had moderate to good performance in the 5–fold cross validations, with AUC ranging from 0.558 to 0.842 (Table 3.2). The AUC in the 5–fold cross validation differed significantly (at alpha = 0.05 level) among three model approaches according to the Friedman Rank Sum Test (*p*–value = 0.0065). The models overall showed low transferability among river basins (Table 3.2). From internal cross–validation to external evaluation, the mean AUC dropped 16%, 20% and 18% respectively for GLM, BRT, and MaxEnt (Table 3.2). The Friedman Rank Sum Test showed that there was no significant difference in the performance of three models in the external evaluations (*p*–value = 0.4317). If transferability AUC = 0.6 was used as the threshold to discriminate good and limited transferability, only 24% of logistic models were transferable and the percentages were even lower for BRT (12%) and MaxEnt model (16%).

Model transferability depended on the prediction region. Relatively high transferability occurred when predicting species distribution in Colorado River basin based on Brazos River models, likely because these two river basins are adjacent (Figure 3.1) and the range and location of their environmental gradients match well (Table 3.1). The models built in the New River basin did not predict well the spatial distribution of species in the Roanoke River and Illinois River basins, although the New River models had high AUC in the cross validation (Table 3.2). These observations imply that model transferability depends more on where the model predicts to rather than on how well the model fit internally. Poor transferability (AUC < 0.5) occurred for 5 fish species (*Cyprinella spiloptera, Nocomis leptocephalus, Campostoma anomalum, Percina roanoka,* and *Chrosomus oreas*) in the New River, Illinois River and Roanoke River. After the incorporation of spatial autocorrelation in the model basin, the transferability AUC increased by 0.01 ~ 0.08 for all three model types (Table B.2). This suggests that transferability of moderately performing models could also be improved by incorporating spatial autocorrelation.

Variable selections and fish–habitat relationships differed among basins. Variables such as temperature and hydrological conditions (particularly base flow index) were more often ranked as important predictors in the boosted regression trees (Table B.1). Habitat condition index was important predictor for all species but *Nocomis leptocephalus* (Bluehead chub) and *Macrhybopsis hyostoma* (Shoal Chub). Few species (*e.g., Notropis rubellus* and *Luxilus chrysocephalus*) had consistent relationships with habitat condition index among different basins; *Notropis buccatus*, *Pimephales notatus,* and *Etheostoma caeruleum* showed distinct responses to habitat disturbance by basin. Comparing the ten key predictors in each pair of basins, we found that the number of predictors ranked top ten or above in both of the paired basins varied from 5 to 9, with an average of 6.3 (Table B.1). For example, BFI, PPT, ELE,

TMA, SINU, MFU and TM are seven key predictors in common for *Semotilus atromaculatus* in the New River and Roanoke River (Table 3.1, Table B.1). The partial dependence plots showed that a predictor variable might be fitted with different functions in a pair of BRT models built in two river basins for the same species (Figure 3.2). We give examples on three general cases. Mostly, the response functions for a variable were consistent among basins as illustrated in Figure 3.2 panel A. In this example, *Nocomis leptocephalus* (Bluehead chub) responded positively to mean annual flow velocity (feet/second) in the New River and Roanoke River basins similarly. High model transferability relied on this kind of consistent species–habitat relationship among river basins. In the second case, the response functions may be different when the range of predictor variable mismatched significantly in the two model basins, corresponding to the effect of variable range illustrated in Figure 3.3 panels A and C. *Campostoma anomalum* (Central stoneroller) showed a strong negative relationship with mean slope (range from 3 to 26) in the New River but this relationship disappeared in the Brazos River where the mean slope ranged from 1 to 7 (Figure 3.2 panel B). In the third case, different or even opposite species–habitat relationships can be observed in two river basins when the location of the variables mismatch, corresponding to the effect of variable location illustrated in the Figure 3.3 panels B and D. Neither the New River model nor the Roanoke River model captured the true effect of elevation on the distribution of *Chrosomus oreas*, since the elevation in the Roanoke River ranges from 0 to 800 meters which is generally lower than New River (Figure 3.2 panel C). After we combined the data of New River and Roanoke River basins, the occurrence of *Chrosomus oreas* (Mountain redbelly dace) and stream elevation showed a unimodal relationship and the optimal elevation for this species appeared to be between 550 to 1000 meters (Figure B.1).

DISCUSSION

Transferability of species distribution models was limited (AUC between 0.5 and 0.6) for majority of stream fish species modeled in this study, corroborating the study of Wenger and Olden (2012) who assessed model transferability for two stream fish species. Making prediction to unsampled regions or future scenarios is more difficult than simple interpolations in the sampled regions, owing to spatial habitat heterogeneity and differing ecological relationships (Peterson et al. 2007, Barbosa et al. 2009, Elith and Leathwick 2009). It is widely recognized that performance in model fitting tends to be higher than validation, and in–sample validation tends to be higher than out–sample prediction (Wang and Jackson 2014). These results reveal the challenge in the transferance of models and suggest the necessity of model transferability evaluations in the study of evaluating climate–change impacts, invasive species assessments, and species reintroduction programs.

Comparing with other studies, we found that model transferability may be taxa–specific. Good transferability are more often reported in studies on plants (Thomas and Bovee 1993, Strauss and Biedermann 2007), birds (Heinänen et al. 2012), and terrestrial manmals, such as Iberian desman (Barbosa et al. 2009), giant panda (Tuanmu et al. 2011), and Brown bear (Martin et al. 2012). The movement of stream fish are usually constrained by watershed boundaries or human–made barriers such as impoundments, road culverts, localized segments with poor water quality and large patches of disturbed landscapes between otherwise suitable stream segments. Constrained by natural or human–made barriers, most fish species may not be occupying all optimal habitats delineated by species distribution models. For example, the few Asian carp species could not easily enter the Great Lakes after they invaded the Mississippi River system because of the electrical barriers set up in the Chicago area waterways (Wittmann et al. 2014).

Trees and terrestrial animals, compared to stream fish, have moderate to good ability to spread to unoccupied suitable habitats. Reasonably good transferability has also been observed for fish in lakes (Wang and Jackson 2014) and coastal systems (Sundblad et al. 2009) that are relatively free of obstacles to movement. Incorporating variables of dispersal and landscape permeability may improve the transferance of SDMs (Midgley et al. 2006, Schurr et al. 2007, Elith and Leathwick 2009). From an optimistic perspective, the unoccupied good habitats can be suitable locations for reintroduction of species of concern and other conservation initiatives, although they would cause high misclassification rate in the assessment of model transferability (Martin et al. 2012).

The choice of modeling approach can also influence the spatial transferability of species distribution models. Our results showed that logistic model with Lasso regularization outperformed complex machine–learning models that may involve high–order functions or interaction terms for good fitting internally, which is consistent with few other studies (Meynard and Quinn 2007, Barbosa et al. 2009, Wenger and Olden 2012). Generalized linear models have clear structure and their complexity can be explicitly controlled by choosing the types and forms of predictor variables. Algorithms based on classification and regression trees are not feasible for extrapolation–prediction beyond the range of predictors (Loh et al. 2007). However, machine–learning models are very useful for examining species–habitat relationships within the sampled region, owing to their ability to account for non–linear patterns and interactions of multiple factors. When the environmental conditions of the model region and prediction region are similar, a case equivalent to interpolation, tree–based algorithms may have good performance in model transferance. In the study of Wang and Jackson (2014), species distribution models for spiny water flea were transferred to adjacent regions with similar environments to the model

regions, so it would not be surprising that random forest models could slightly outperform logistic regression.

We recommend the use of spatial models with simple structures when model transfers are needed. Spatial eigenvector mapping approaches (Griffith 2000, Borcard and Legendre 2002) could extract the spatial eigenvectors based on the among-sites distance, and then filter the spatial structure in the environmental predictors, allowing predicting species distribution purely based on ecological components. The increase in transferability AUC in all three modelling approaches suggest the use of spatial models in future studies of among-region model transferance. GLM is not immune to over–fitting (Wenger et al. 2011), so we applied the Lasso regularization in the logistic model with only first–order terms to control the over–fitting and potential multicollinearity. Different from the ridge regularization, the Lasso regularization can force the estimated coefficients of some predictors to zero when the penalty on model complexity is high, essentially allowing an automatic variable selection. Wenger and Olden (2012) summarized the strategies of controlling over–fitting and model complexity in machine learning models. In addition, increasing the regularization values in MaxEnt models, thus controlling the complexity can produce more transferable models (Heinänen et al. 2012). More efforts are needed to assess data quality, particularly sample size, the range and position of predictor variables, in both the model region and prediction region before transference of species distribution models. In this study, we verified the concept of effect of variable range (Jackson et al. 2001) and effect of variable location by the partial dependence plots in paired boosted regression trees. Inconsistent or even contradictory results can be observed if the ranges or the locations of the predictors mismatch between training and prediction datasets. Relatively good performance was achieved in predicting fish distributions in the Colorado River (Texas) with

Brazos River models likely because environmental conditions are also similar (Table 3.1). In contrast, the New River and Roanoke River, although adjacent, are different in physiography (e.g., mean evaluation, temperature and landform), and not in the same large water systems, leading to low model transferability among those basins. One remedy to poor transferability is to let the model region have large extent or wide range of environmental gradients, meaning the ranges of predictors in the prediction region should be largely within the model ranges. This essentially makes the prediction similar to internal evaluation so good performance could be expected. Broadening the extent of training models, however, requires that the data constraints that have necessitated model extrapolations first need to be overcome. The statistical support behind making training data ranges encompass prediction data ranges is that making predictions near the core of the modeling space is more stable than predictions outside of modeling space, or far away from the center of training data (Schabenberger and Pierce 2002). Alternatively, extending the sample size in the training data, would improve model transferability (Strauss and Bierdermann 2007, Barbosa et al. 2009, Wang and Jackson 2014).

This study confirmed that predictor variables with direct effects are more transferable than indirect variables (Sundblad et al. 2009, Murray et al. 2011, Wenger and Olden 2012). We found that variables such as temperature and hydrological conditions were more often ranked as important predictors in the boosted regression trees (Table B.1), likely because these variables directly affect fish's survival, growth, and production. In contrast, measures of anthropological disturbances (e.g., road crossing, human population size, habitat condition index) are less transferable, and the fish–habitat relationships described by partial dependence plots for these variables were at times inconsistent among paired basins, although their ranges matched (Table 3.1 and Table 3.2). Habitat condition index that measures habitat disturbance was important

predictor for over 80% species modeled in this study, but few species (e.g., *Notropis buccatus*, *Pimephales notatus,* and *Etheostoma caeruleum*) showed distinct responses to habitat disturbance by basin. Large–scale landscape variables are easy to obtain but they are usually region–specific, making them difficult to transfer among regions. Agreeing with (Sundblad et al. 2009, Murray et al. 2011, Wenger and Olden 2012), we recommend to include fine–scale variables with a sound ecological basis in the extrapolation of species distribution models. Two approaches are promising to help gather fine–scale direct variables more efficiently. First, regression models using larger–scale predictors can be used to predict fine–scale direct variables, such as hydrology (Carlisle et al. 2010, Segura et al. 2014), water temperature (Morse 1970, McKenna et al. 2010) of streams for a given time frame. Second, there is an increased use of advanced techniques such as side–scan sonar (Kaeser and Litts 2010) and remote sensing (Feurer et al. 2008, Tang et al. 2009) to create high–resolution and spatially–continuous in–stream habitat across broad aquatic landscapes. Additional variable selection among direct variables is necessary to control model complexity and multicollinearity in model transference.

ACKNOWLEDGEMENTS

This work was funded by the US Geological Survey Aquatic Gap Analysis Program.

REFERENCES

Araújo, M. B., et al. 2005. Validation of species–climate impact models under climate change. – Glob. Chang. Biol. 11: 1504–1513.

Arntzen, J. 2006. From descriptive to predictive distribution models: a working example with Iberian amphibians and reptiles. – Front. Ecol. 3: 8.

Barbosa, A. M., R. Real, and J. Mario Vargas. 2009. Transferability of environmental favourability models in geographic space: The case of the Iberian desman (*Galemys pyrenaicus*) in Portugal and Spain. – Ecol. Model. 220: 747–754.

Borcard, D., and P. Legendre. 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. – Ecol. Model. 153: 51–68.

Bradley, A.P. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. – Pattern Recognit. 30: 1145–1159.

Brind'amour, A., D. Boisclair, S. Dray, and P. Legendre. 2011. Relationships between species feeding traits and environmental conditions in fish communities: a three-matrix approach. – Ecol. Appl. 21: 363–377.

Breiman, L. 2001. Random Forests. – Mach. Learn. 45: 5–32.

Brown, J. H. and M.V. Lomolino. 1998. Biogeography (Second ed.). – Sinauer Associates.

Bulluck, L., E. Fleishman, C. Betrus, and R. Blair. 2006. Spatial and temporal variations in species occurrence rate affect the accuracy of occurrence models. – Glob. Ecol. Biogeogr. 15: 27–38.

Carlisle, D. M., J. Falcone, D. M. Wolock, M. R. Meador, and R. H. Norris. 2010. Predicting the natural flow regime: models for assessing hydrological alteration in streams. – River Res. Appl. 26: 118–136.

Chu, C., N. E. Mandrak, and C. K. Minns. 2005. Potential impacts of climate change on the distributions of several common and rare freshwater fishes in Canada. – Divers. Distrib. 11: 299–310.

Comte, L., and G. Grenouillet. 2013. Do stream fish track climate change? Assessing distribution shifts in recent decades. – Ecography 36: 1236–1246.

Elith, J., C. H. Graham, R. P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, L. Jin, L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. McC. Overton, A. T. Peterson, and S. J. Phillips. 2006. Novel methods improve prediction of species' distributions from occurrence data. – Ecography 29: 129–151.

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

EPA (Environmental Protection Agency), USGS (United States Geological Survey), Horizon Systems Corporations [Internet]. 2010. NHDPlusV1 Data. [cited 2013 July 29]. Available from: http://www.horizon–systems.com/nhdplus/data.php

EPA (Environmental Protection Agency), USGS (United States Geological Survey), Horizon Systems Corporations [Internet]. 2012. NHDplusV2 Data. [cited 2013 July 29]. Available from: http://www.horizon–systems.com/nhdplus/NHDplusV2_data.php

Ervin, G. N., and D. C. Holly. 2011. Examining local transferability of predictive species distribution models for invasive plants: an example with cogongrass (*Imperata cylindrica*). – Invas. Plant. Sci. Mana. 4: 390–401.

Esselman, P. C., D. M. Infante, L. Wang, D. Wu, A. R. Cooper and W. W. Taylor. 2011. An index of cumulative disturbance to river fish habitats of the Conterminous United States from landscape anthropogenic activities. – Ecol. Restoration. 29: 133–151.

Feurer, D., J.-S. Bailly, C. Puech, Y. Le Coarer, and A. A. Viau. 2008. Very–high–resolution mapping of river–immersed topography by remote sensing. – Prog. Phys. Geogr. 32: 403–419.

Fielding, A. H., and P. F. Haworth. 1995. Testing the generality of bird–habitat models. – Conserv. Biol. 9: 1466–1481.

Franklin, J., and J. A. Miller. 2009. Mapping species distributions: spatial inference and prediction. – Cambridge University Press.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. – Ann. Stat. 29: 1189–1232.

Friedman, J. H., and J. J. Meulman. 2003. Multiple additive regression trees with application in epidemiology. – Stat. Med. 22: 1365–1381.

Friedman, J., F. Jerome, H. Trevor, and T. Rob. 2010. Regularization paths for generalized linear models via coordinate descent. – J. Stat. Softw. 33: 1–22.

Frimpong, E. A, T. M Sutton, B. A. Engel, and T. P. Simon. 2005. Spatial-scale effects on relative importance of physical habitat predictors of stream health. – Environ. Manage. 36: 899–917.

Griffith, D. A. 2000. A linear regression solution to the spatial autocorrelation problem. – J. Geogr. Syst. 2: 141–156.

Guay, J. C., D. Boisclair, M. Leclerc, and M. Lapointe. 2003. Assessment of the transferability of biological habitat models for Atlantic salmon parr (*Salmo salar*). – Can. J. Fish. Aquat. Sci. 60: 1398–1408.

Hanley, J. A. and B. J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristics curve. – Radiology 143: 29–36.

Heinänen, S., J. Erola, and M. von Numers. 2012. High resolution species distribution models of two nesting water bird species: a study of transferability and predictive performance. – Landsc. Ecol. 27: 545–555.

Hijmans, R. J., S. Phillips, J. Leathwick, and J. Elith. 2013. dismo: Species distribution modeling. R package version 0.9-3. http://CRAN.R–project.org/package=dismo.

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. – Can. J. Fish. Aquat. Sci. 58: 157–170.

Jones, C. C. 2012. Challenges in predicting the future distributions of invasive plant species. – For. Ecol. Manage. 284: 69–77.

Kaeser, A. J., and T. L. Litts. 2010. A novel technique for mapping habitat in navigable streams using low–cost side scan sonar. – Fisheries 35: 163–174.

Loh, W.-Y., C.-W. Chen, and W. Zheng. 2007. Extrapolation errors in linear model trees. – ACM. T. Knowl. Discov. D. 1: 1–17.

Legendre, P., D. Borcard, F. G. Blanchet and S. Dray. 2012. PCNM: MEM spatial eigenfunction and principal coordinate analyses. R package version 2.1-2/r106. http://R-Forge.R-project.org/projects/sedar/.

Lyons, J., J. S. Stewart, and M. Mitro. 2010. Predicted effects of climate warming on the distribution of 50 stream fishes in Wisconsin, U.S.A. – J. Fish Biol. 77: 1867–1898.

Mac Nally, R. 2000. Regression and model–building in conservation biology, biogeography and ecology: The distinction between - and reconciliation of - ‘predictive’ and ‘explanatory’ models. – Biodivers. Conserv. 9: 655–671.

Martin, J., E. Revilla, P.-Y. Quenette, J. Naves, D. Allainé, and J. E. Swenson. 2012. Brown bear habitat suitability in the Pyrenees: transferability across sites and linking scales to make the most of scarce data. – J. Appl. Ecol. 49: 621–631.

Maurer, B. A., and M. L. Taper. 2002. Connecting geographical distributions with population processes. – Ecol. Lett. 5: 223–231.

McKenna, J. E., R. S. Butryn, and R. P. McDonald. 2010. Summer stream water temperature models for great lakes streams: New York. – T. Am. Fish.Soc. 139: 1399–1414.

Meynard, C. N., and J. F. Quinn. 2007. Predicting species distributions: a critical comparison of the most common statistical models using artificial species. – J. Biogeogr. 34: 1455–1469.

Midgley, G. F., G. O. Hughes, W. Thuiller, and A. G. Rebelo. 2006. Migration rate limitations on climate change-induced range shifts in Cape Proteaceae. – Divers. Distrib. 12: 555–562.

Moran, P. A. P. 1950. Notes on Continuous Stochastic Phenomena. – Biometrika 37: 17–23.

Morse, W. L. 1970. Stream temperature prediction model. – Water Resour. Res. 6: 290–302.

Murray, J. V., S. Low Choy, C. A. McAlpine, H. P. Possingham, and A. W. Goldizen. 2011. Evaluating model transferability for a threatened species to adjacent areas: Implications for rock–wallaby conservation. – Austral Ecol. 36: 76–89.

Ling, C. X., J. Huang, H. Zhang. 2003. AUC: A better measure than accuracy in comparing learning algorithms. – In: Springer, Verlag Berlin, pp. 329–341.

Lobo, J. M., A. Jiménez-Valverde, and R. Real. 2008. AUC: a misleading measure of the performance of predictive distribution models. – Glob. Ecol. Biogeogr. 17: 145–151.

Peterson, A. T., M. Pape, and M. Eaton. 2007. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. – Ecography 30: 550–560.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. – Ecol. Model. 190: 231–259.

Price, C. V., N. Nakagaki, K. J. Hitt, and R. C. Clawges [Internet]. 2006. Enhanced Historical Land-Use and Land-Cover Data Sets of the U.S. Geological Survey, USGS Digital Data Series 240. [cited 2014 May 31]. Available from: http://pubs.usgs.gov/ds/2006/240

PRISM Climate Group, Oregon State University [Internet]. 2004. PRISM Climate Data. [cited 2014 May 1]. Available from: http://prism.oregonstate.edu.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available http://www.R–project.org/.

Randin, C. F., T. Dirnböck, S. Dullinger, N. E. Zimmermann, M. Zappa, and A. Guisan. 2006. Are niche-based species distribution models transferable in space? – J. Biogeogr. 33: 1689–1703.

Robin, X., N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez and M, Müller. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. – BMC Bioinformatics 12: 77.

Schabenberger, O., and F. J. Pierce. 2002. Contemporary statistical models for the plant and soil sciences. – CRC Press.

Schibalski, A., A. Lehtonen, and B. Schröder. 2014. Climate change shifts environmental space and limits transferability of treeline models. – Ecography 37: 321–335.

Schurr, F. M., G. F. Midgley, A. G. Rebelo, G. Reeves, P. Poschlod, and S. I. Higgins. 2007. Colonization and persistence ability explain the extent to which plant species fill their potential range. – Glob. Ecol. Biogeogr. 16: 449–459.

Segura, C., P. Caldwell, G. Sun, S. McNulty, and Y. Zhang. 2014. A model to predict stream water temperature across the conterminous USA. – Hydrol. Process. (Online version of record published before inclusion in an issue).

Segurado, P. and M.B. Araújo. 2004. An evaluation of methods for modelling species distributions. – J. Biogeogr. 31: 1555–1568.

Strauss, B., and R. Biedermann. 2007. Evaluating temporal and spatial generality: How valid are species–habitat relationship models? – Ecol. Model. 204: 104–114.

Sundblad, G., M. Härmä, A. Lappalainen, L. Urho, and U. Bergström. 2009. Transferability of predictive fish distribution models in two coastal systems. – Estuar. Coast. Shelf Sci. 83: 90–96.

Swets, J. 1988. Measuring the accuracy of diagnostic systems. – Science 240: 1285–1293.

Tang, Q., H. Gao, H. Lu, and D. P. Lettenmaier. 2009. Remote sensing: hydrology. – Prog. Phys. Geogr. 33: 490–509.

Thomas, J. A., and K. D. Bovee. 1993. Application and testing of a procedure to evaluate transferability of habitat suitability criteria. – Regulated Rivers: Res. Manag. 8: 285–294.

Tibshirani, R. 1996. Regression shrinkage and selection via the Lasso. – J. R. Stat. Soc. Series B (Methodological) 58: 267–288.

Tuanmu, M.–N., A. Viña, G. J. Roloff, W. Liu, Z. Ouyang, H. Zhang, and J. Liu. 2011. Temporal transferability of wildlife habitat models: implications for habitat monitoring. – J. Biogeogr. 38: 1510–1523.

Wang, L., and D. Jackson. 2014. Shaping up model transferability and generality of species distribution modeling for predicting invasions: implications from a study on *Bythotrephes longimanus*. – Biol. Invasions 16: 2079–2103.

Wenger, S. J., D. J. Isaak, C. H. Luce, H. M. Neville, K. D. Fausch, J. B. Dunham, D. C. Dauwalter, M. K. Young, M. M. Elsner, B. E. Rieman, A. F. Hamlet, and J. E. Williams. 2011. Flow regime, temperature, and biotic interactions drive differential declines of trout species under climate change. – Natl. Acad. Sci. U.S.A. 108: 14175–14180.

Wenger, S. J., and J. D. Olden. 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. – Methods Ecol. Evol. 3: 260–267.

Wiener, N. 1949. Extrapolation, interpolation, and smoothing of stationary time series, with engineering applications. – Technology Press of the Massachusetts Institute of Technology.

Wittmann, M. E., M. C. Roger, J.D. Rothlisberger, and D. M. Lodge. 2014. Using structured expert judgment to assess invasive species prevention: Asian carp and the Mississippi-great lakes hydrologic connection. – Environ. Sci. Technol. 48: 2150–2156.

Table 3.1. The descriptions of predictor variables used to develop species distribution models for the 21 native stream fish species in the United States. Data are from NHDplusV2 (National Hydrography Dataset Plus Version 2; EPA, USGS, Horizon Systems Corporations 2012), and U.S. Geological Survey (Price et al. 2006), NFHAP (National Fish Habitat Board 2012), NHDplusV1 (National Hydrography Dataset Plus Version 1; EPA, USGS, Horizon Systems Corporations 2010), and PRISM (PRISM climate group 2004). The variables were measured in the unit of NHD inter–confluence stream segment.

		Range of variable				
Variable	Descriptions	Brazos River	Colorado River	Illinois River	New River	Roanoke River
BFI	The ratio of base flow (i.e., the component of stream flow attributed to ground–water discharge) to total flow (%).	3~38.6	6.3~59.3	14~66.1	33~67	29.7~63.9
ELE	Mean elevation (meter)	4.5~963	0.4~1256.3	128.2~294.1	360.4~1364.2	0.3~900
MFU	Mean annual flow in cubic feet per second (cfs) at bottom of flowline as computed by Unit Runoff method.	0.1~7575.4	0~4457	0.2~18057.8	0.2~9978.1	0.1~9149.9
MVU	Mean annual velocity (fps) at bottom of flowline as computed by Jobson (1996).	0.4~3.1	0.4~4.2	0.5~2.6	0.7~3.6	0.4~2.3
RDX	Number of road–stream crossing.	0~46	0~88	0~33	0~25	0~15
SINU	Sinuosity–reach length divided by length of straight line between two nodes.	1~4.8	1~6.1	1~3.4	1~5.8	1~6.2
SLP	Mean slope (degrees)	0.1~7.3	0~11.6	0~10.7	4.1~25	0~27.9
SO	Stream order (Strahler 1952)	1~7	1~7	1~8	1~8	1~7
C_AG[1]	Percentage of agriculture	0~100	0~92.6	0~100	0~100	0~100
C_FR[1]	Percentage of forest	0~100	0~97	0~100	0~100	0~100
C_UB[1]	Percentage of urban	0~99	0~96.4	0~100	0~95.5	0~100

HCI[2]	An index of cumulative disturbance for a catchment.*	1~4.3	1~5	0~4	1.4~4.4	0~4.2
NT[3]	Sum total of nitrogen input to the land surface of the watershed (kg/year)	0.9~375350.2	0.5~2440140.5	2.3~1824836.4	0.1~319763.2	0.7~54132
POP[3]	Human population density (persons per km^2)	0~186.9	0~293.4	0~442.8	0~127.3	0~127.2
PT[3]	Sum total of phosphorus input to the land surface of the watershed (kg/year).	0.2~68493.7	0.1~297743.5	0.5~366853.1	0~82509.1	0.2~16161.9
PPT[4]	20–year (1961–1980) average annual precipitation (mm).	460.5~1381.9	385.9~1328.2	820.9~1036.8	887.7~1556.6	934.6~1581.2
TM[4]	20–Year (1961–1980) average temperature (°C).	15.6~20.4	14.3~21.1	8~12.6	7.9~12.2	10~16
TMA[4]	20–Year (1961–1980) average annual maximum temperature (°C).	33.2~36.8	32.3~38.1	27.3~31.5	24.5~29.4	24.8~31
TMI[4]	20–Year (1961–1980) average annual minimum temperature (°C).	–3.7~ 6.2	–4.0~ 7.2	–12.5~ –7.2	–9.9~ –5.1	–6.9~ –1.8

Note: the number 1–4 by the variables denote the sources: null–NHDplusV2; 1–USGS; 2–NFHP; 3– NHDplusV1; 4– PRISM. *An index of cumulative disturbance of catchments of inter–confluence stream segments calculated based on 15 disturbance variables (Esselman et al. 2011). The influence of each distribution variable was weighted by the results of multiple linear regression of all variables against a commonly used biological indicator of habitat condition (i.e., percent intolerant fishes at a site). A HCI of 0 indicates pristine condition, while higher scores indicate severer disturbance.

Table 3.2. Summary of the performance of boosted regression tree (BRT), logistic models (GLM), and MaxEnt models in terms of area under the receiver–operator characteristic curve (AUC) in the 5–fold cross–validation and among–basins extrapolation for the 21 fish species in 5 river basins of the United States (i.e., BR–Brazos River, CR–Colorado River, IR–Illinois River, NR–New River, and RR–Roanoke River).

		Basin		Cross validation (5–fold)			Prediction		
Species name	Common name	Model	Prediction	GLM	BRT	MaxEnt	GLM	BRT	MaxEnt
Lepisosteus osseus	Longnose gar	BR	CR	0.634	0.569	0.616	0.571	0.454	0.618
Ameiurus natalis	Yellow bullhead	BR	CR	0.618	0.653	0.654	0.586	0.488	0.540
Aphredoderus sayanus	Pirate perch	BR	CR	0.620	0.693	0.592	0.874	0.580	0.770
Menidia beryllina	Inland silverside	BR	CR	0.701	0.564	0.558	0.656	0.643	0.633
Etheostoma gracile	Slough darter	BR	CR	0.620	0.612	0.587	0.656	0.676	0.715
Macrhybopsis hyostoma	Shoal chub	BR	CR	0.603	0.604	0.602	0.594	0.467	0.631
Lepisosteus osseus	Longnose gar	BR	IR	0.633	0.659	0.588	0.571	0.501	0.598
Ameiurus natalis	Yellow bullhead	BR	IR	0.618	0.653	0.654	0.508	0.516	0.497
Lepomis humilis	Orangespotted sunfish	BR	IR	0.616	0.566	0.652	0.530	0.533	0.491
Rhinichthys atratulus	Blacknose dace	NR	IR	0.742	0.700	0.707	0.510	0.565	0.453
Notropis rubellus	Rosyface shiner	NR	IR	0.602	0.602	0.559	0.552	0.595	0.567
Notropis buccatus	Silverjaw minnow	NR	IR	0.753	0.772	0.727	0.491	0.610	0.413
Cyprinella spiloptera	Spotfin shiner	NR	IR	0.603	0.570	0.584	0.510	0.474	0.473
Luxilus chrysocephalus	Striped shiner	NR	IR	0.793	0.739	0.690	0.526	0.479	0.516
Etheostoma caeruleum	Rainbow darter	NR	IR	0.605	0.574	0.648	0.572	0.488	0.460
Semotilus atromaculatus	Creek chub	NR	RR	0.842	0.802	0.821	0.741	0.556	0.508
Rhinichthys atratulus	Blacknose dace	NR	RR	0.742	0.700	0.707	0.644	0.469	0.448
Nocomis leptocephalus	Bluehead chub	NR	RR	0.616	0.559	0.576	0.477	0.481	0.489
Notropis blennius	River shiner	NR	RR	0.727	0.682	0.680	0.523	0.478	0.564
Campostoma anomalum	Central stoneroller	NR	RR	0.836	0.809	0.807	0.505	0.471	0.500
Pimephales notatus	Bluntnose minnow	NR	RR	0.641	0.578	0.558	0.598	0.516	0.481
Cyprinella spiloptera	Spotfin shiner	NR	RR	0.615	0.618	0.594	0.791	0.593	0.463
Percina roanoka	Roanoke darter	NR	RR	0.726	0.609	0.688	0.420	0.453	0.441
Etheostoma flabellare	Fantail darter	NR	RR	0.843	0.799	0.809	0.518	0.524	0.522
Chrosomus oreas	Mountain redbelly dace	NR	RR	0.630	0.727	0.579	0.453	0.483	0.518

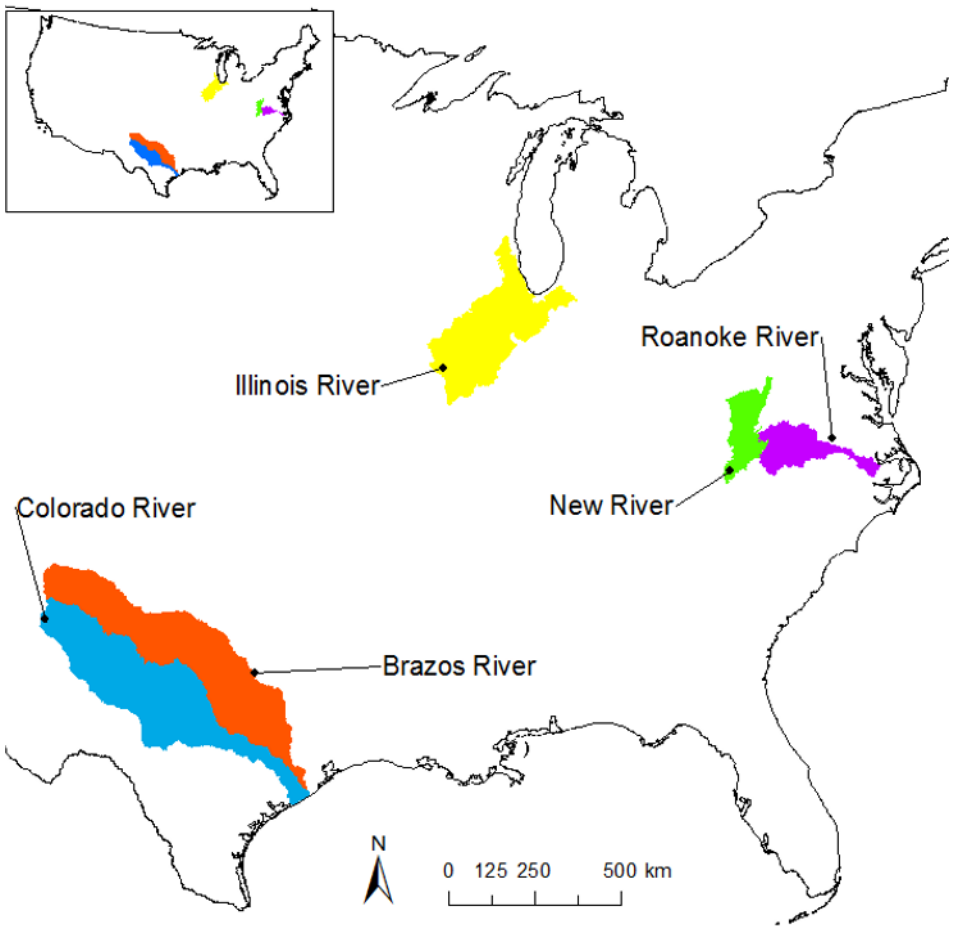
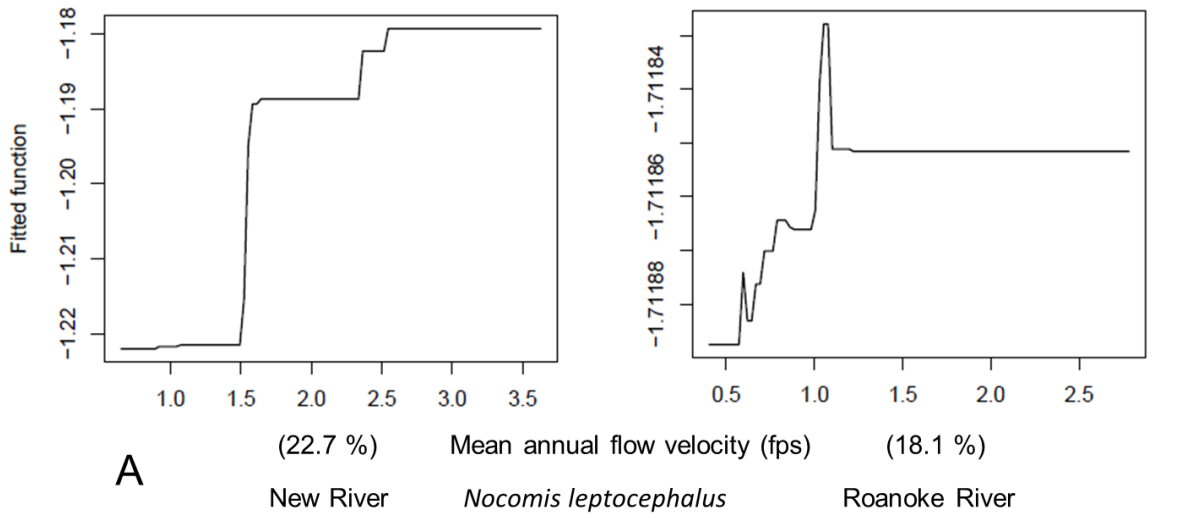


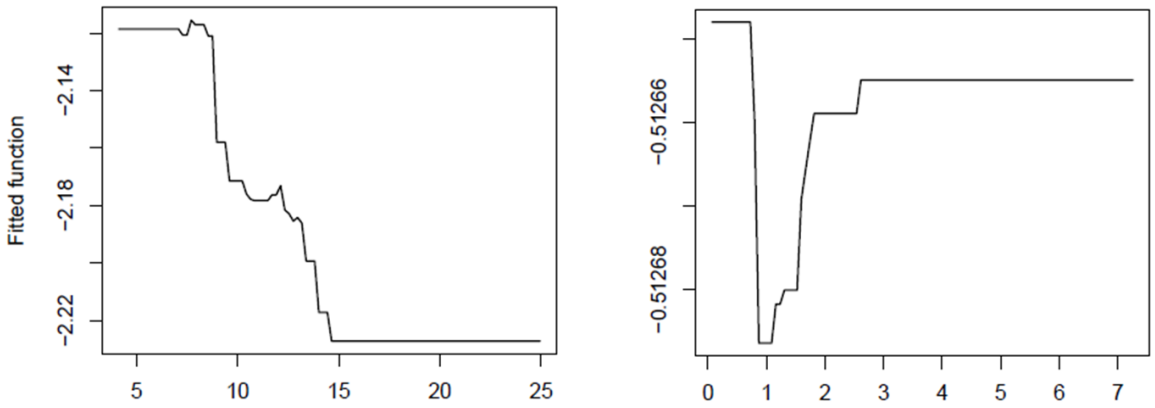
Figure 3.1. A map showing the five study river basins in the eastern United States. The species distribution models were developed and cross–validated based on data in the New River and Brazos River. Model transference were made from New River to Illinois River and Roanoke River, and from Brazos River to Illinois River and Colorado River in Texas.



A

(22.7 %) Mean annual flow velocity (fps) (18.1 %)

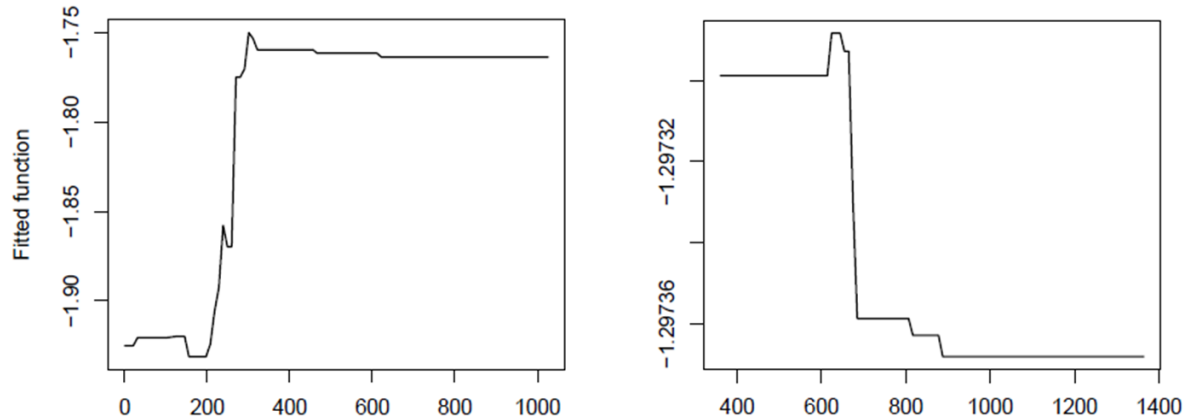
New River *Nocomis leptocephalus* Roanoke River



B

(5.6 %) Mean slope (degree) (7.0 %)

New River *Campostoma anomalum* Brazos River



C

(18.9 %) Elevation (m) (10.3 %)

Roanoke River *Chrosomus oreas* New River

Figure 3.2. Sample partial dependence plots for fish species occurring in multiple basins. Panel A shows that *Nocomis leptocephalus* (Bluehead chub) responded to mean annual flow velocity (feet/second) in a similar way among basins. In the panel B, *Campostoma anomalum* (Central stoneroller) in the New River showed a strong negative relationship with mean slope (range from 3 to 26), but this relationship did not hold in the Brazos River where the mean slope ranges from 1 to 7. This example illustrates that species–habitat relationship is affected by the range of the variable in the sample. Panel C is an example illustrating the effect of variable location: the probability of occurrence of *Chrosomus oreas* (Mountain redbelly dace) increased with elevation in the Roanoke River but decreased with elevation in the New River. The mismatches in the range and location of the predictor variable shown in the panel B and C are possible causes of limited model transferability among regions. Note: The percentage values in the brackets are the relative importance of the predictor variable in the boosted regression tree models (Friedman et al. 2001), and the scale of fitted function are specific to the model from which the partial dependence plot is generated.

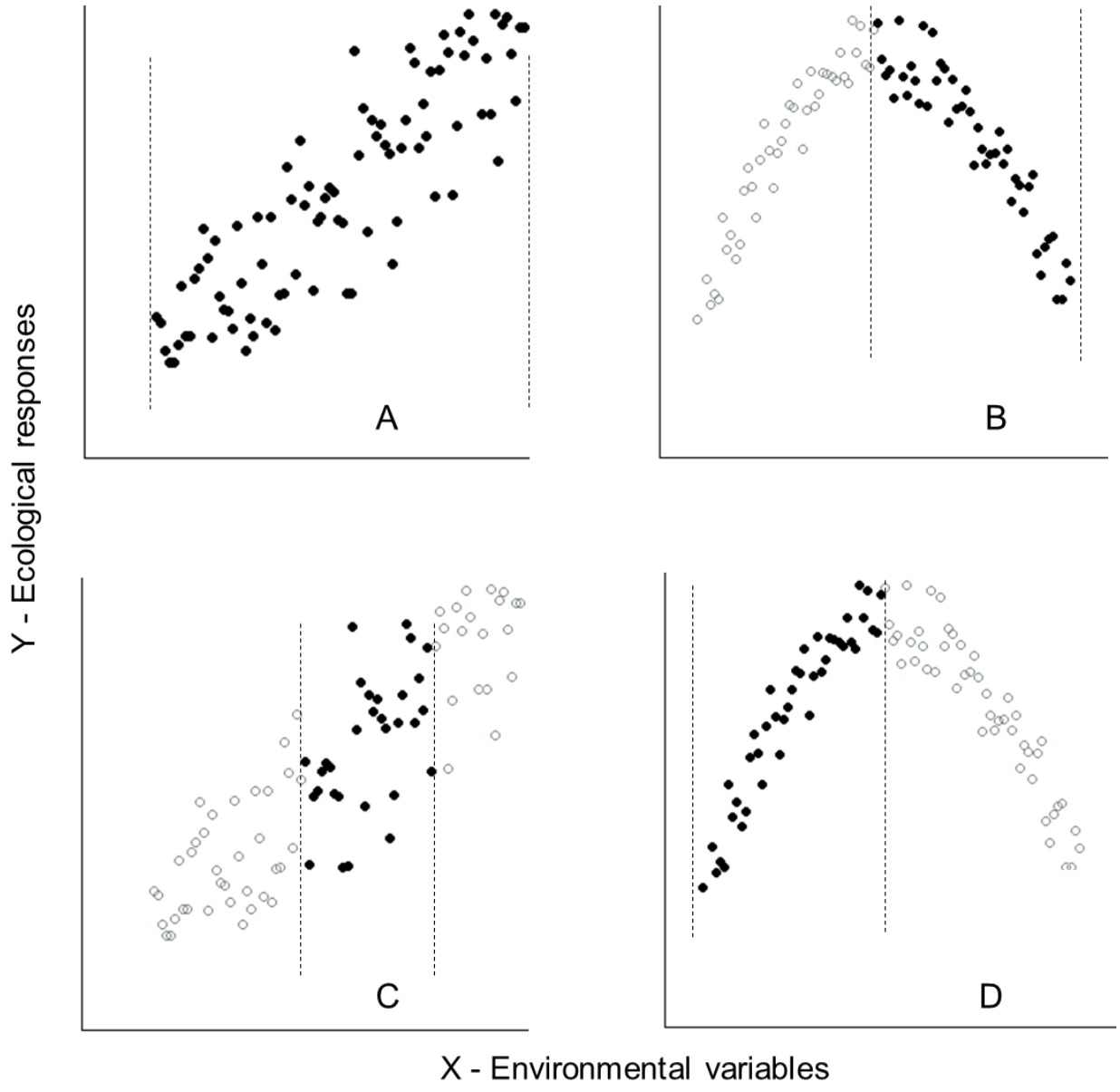


Figure 3.3. Effects of the range (panel A and C) and location (panel B and D) of predictor variables on the relationship of habitat (X) and ecological response (Y) described by models. Comparing panel A and C, we may conclude a strong X–Y relationship if the X's entire range is covered in the study, while we may not conclude so if X's range is restricted (Jackson et al. 2001). We made panel C based on the Figure 2 of Jackson et al. (2001). The X–Y relationships may also be inconsistent or even contradictory among studies or regions or time frames when the locations of X mismatch heavily. We see a strong negative X–Y relationship when the right half of the X is covered in the panel B, and a strong positive X–Y relationship when the left half of the X is covered in the panel D. Neither X–Y relationship described in panel B and panel D is appropriate when we transfer (extrapolate) models or generalize the relationship to a large scale.

APPENDIX B SUPPLEMENTARY INFORMATION

Table B.1. Summary of the key habitat features for each of the 21 stream fish species in five river basins (BR–Brazos River, CR–Colorado River, IL–Illinois River, NR–New River, and RR–Roanoke River) in the boosted regression tree (BRT) models. Predictor variables V1–V10 are the ten most important variables for each species in a given river basin. The response of the each species to important variables were evaluated using partial dependence plots (Hijmans et al. 2013). We use "↗" to represent strong positive relationship, "↘" for strong negative relationship, and "~" for complicated non–linear relationships. For instance, "BFI ↘" means that base flow index (BFI) was the most important variable in the BRT model for *Lepisosteus osseus* (Longnose gar) in the Brazos River basin. The descriptions of predictor variables are listed in the Table 3.1.

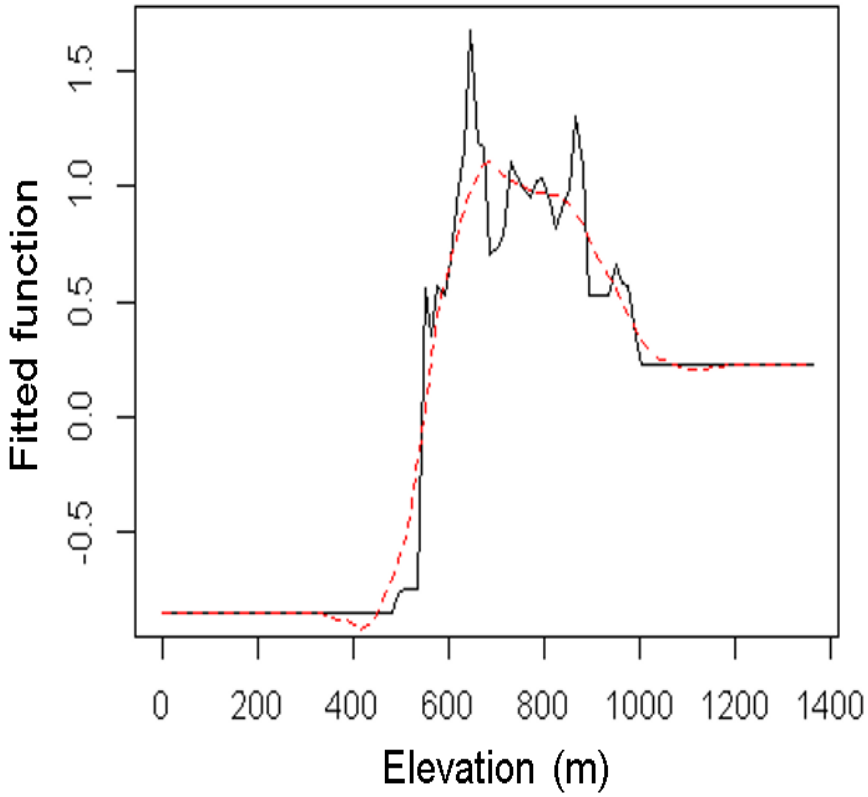
		Key environmental predictors									
Species names	Basin	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
Lepisosteus osseus	BR	BFI ↘	TMI ~	TMA ~	TM ↗	SINU ~	MFU ~	ELE ~	POP ~	MVU ~	C_UB ~
	IR	BFI ~	HCI ~	NT ~	TMI ↗	C_AG ↗	MVU ↗	TM ↘	SINU ↗	PPT ↗	SO ↗
	CR	MVU ↗	C_AG ↗	BFI ↗	MFU ↗	TMA ↘	PPT ↘	ELE ~	NT ~	PT ~	SLP ~
Semotilus atromaculatus	NR	BFI ↗	PPT ↗	ELE ↗	HCI ↗	TMI ↗	SINU ~	TMA ↘	MFU ~	MVU ~	TM ↗
	RR	PPT ↗	TM ↗	BFI ~	C_AG ↘	NT ↗	TMA ~	ELE ↗	MFU ~	SLP ~	SINU ~
Rhinichthys atratulus	NR	BFI ↗	PPT ↗	TMI ↗	TMA ↘	MVU ↘	ELE ↗	TM ↘	MFU ~	C_AG ↗	SINU ~
	IR	NT ~	HCI ~	ELE ↗	PT ~	PPT ↗	C_AG ↗	POP ~	BFI ~	MFU ~	RDX ↗
	RR	PPT ↘	SINU ↗	POP ↗	ELE ↗	TMI ↘	MFU ~	NT ~	TMA ↘	MVU ~	TM ↘
Nocomis leptocephalus	NR	MVU ↗	TMA ↗	BFI ↘	ELE ↘	SLP ↗	TM ~	C_UB ~	TMI ~	POP ~	MFU ↗
	RR	BFI ~	MVU ↗	SINU ↗	POP ~	NT ~	C_UB ↘	PPT ↘	C_AG ↘	TMI ~	MFU ↗
Notropis rubellus	NR	BFI ↗	SLP ↘	ELE ~	C_AG ↘	PPT ↗	TM ↘	MFU ~	HCI ↗	SO ↘	NT ~
	IR	TM ~	MVU ↘	MFU ~	POP ~	SLP ↗	TMA ↘	ELE ~	HCI ↗	C_AG ↘	PPT ↘
Notropis blennius	NR	ELE ↘	PPT ~	POP ~	TMI ~	RDX ↗	BFI ↘	MFU ~	MVU ~	NT ~	C_AG ↘
	RR	NT ~	TMI ↘	PPT ↘	PT ↗	SINU ↗	SLP ~	MVU ↘	TM ↗	BFI ↗	ELE ↗
Notropis buccatus	NR	BFI ↗	HCI ↗	ELE ↗	PPT ~	C_AG ~	TMA ↘	TMI ~	TM ↗	MFU ~	SINU ~
	IR	SLP ↗	TMA ↘	TMI ↘	HCI ↘	PPT ↗	SINU ↗	RDX ↗	NT ~	PT ~	TM ↗
Campostoma anomalum	NR	PPT ↗	BFI ↗	TMA ↘	HCI ↗	NT ~	TMI ↗	POP ~	SLP ↘	C_AG ↗	ELE ↗
	RR	ELE ↗	TM ~	SINU ~	SLP ↗	MFU ~	NT ~	TMA ~	PPT ↘	PT ~	BFI ↗

Pimephales notatus	NR	PPT ⮥	BFI ⮥	POP ~	HCI ⮥	SLP ↴	ELE ⮥	MFU ~	NT ~	C_AG ~	MVU ~
	RR	PPT ↴	POP ~	TMI ~	MVU ~	MFU ~	ELE ~	SINU ⮥	SLP ~	BFI ⮥	HCI ↴
Cyprinella spiloptera	NR	ELE ↴	TMA ↴	RDX ↴	SINU ⮥	C_AG ↴	MFU ~	PPT ↴	PT ~	POP ~	C_UB ⮥
	IR	POP ~	MFU ~	HCI ↴	TMI ↴	C_AG ~	TMA ⮥	ELE ⮥	TM ↴	C_UB ~	MVU ⮥
Luxilus chrysocephalus	NR	TM ⮥	TMA ↴	POP ~	PT ~	HCI ⮥	ELE ⮥	PPT ~	MVU ⮥	SINU ⮥	BFI ⮥
	IR	C_AG ~	PPT ⮥	HCI ⮥	TMI ~	MFU ~	TMA ↴	POP ~	ELE ~	TM ~	NT ~
	RR	TM ⮥	TMA ~	SINU ⮥	TMI ⮥	BFI ~	ELE ⮥	MFU ~	SLP ~	MVU ~	C_AG ⮥
Ameiurus natalis	BR	SLP ⮥	MFU ~	HCI ↴	BFI ~	POP ~	PPT ~	C_AG ~	ELE ⮥	SINU ~	TMI ~
	IR	SINU ⮥	POP ~	ELE ⮥	NT ~	HCI ~	TMI ⮥	TMA ~	PPT ~	C_AG ↴	SLP ~
	CR	TMA ~	MVU ↴	PPT ⮥	SLP ⮥	POP ⮥	MFU ~	ELE ~	PT ~	NT ~	TMI ⮥
Aphredoderus sayanus	BR	C_AG ⮥	HCI ~	BFI ↴	TMA ⮥	SLP ~	NT ~	POP ~	PT ↴	TMI ↴	MVU ↴
	CR	TMA ↴	ELE ~	TMI ⮥	SLP ~	PPT ⮥	C_AG ↴	SINU ⮥	POP ~	HCI ~	BFI ↴
Menidia beryllina	BR	BFI ↴	SLP ⮥	HCI ~	TMI ↴	C_AG ~	TMA ⮥	PPT ⮥	MFU ⮥	MVU ⮥	NT ~
	CR	PPT ↴	MVU ⮥	TMI ⮥	MFU ~	ELE ↴	SINU ~	BFI ↴	SLP ~	C_UB ⮥	POP ⮥
Lepomis humilis	BR	ELE ↴	TM ↴	SLP ⮥	TMI ↴	HCI ~	SINU ~	PPT ~	MVU ↴	C_AG ↴	MFU ~
	IR	ELE ~	RDX ⮥	C_AG ~	NT ~	TMI ⮥	POP ⮥	SINU ⮥	MFU ~	SLP ⮥	PT ~
Etheostoma gracile	BR	PPT ⮥	ELE ~	MFU ~	C_AG ⮥	TMA ⮥	PT ~	POP ~	RDX ~	TM ↴	SLP ~
	CR	ELE ↴	PPT ⮥	TMI ↴	C_AG ⮥	SINU ⮥	PT ~	TMA ~	MFU ~	SLP ~	HCI ⮥
Etheostoma caeruleum	NR	MFU ~	HCI ⮥	TMA ↴	PT ~	SINU ~	ELE ↴	POP ~	MVU ~	SLP ↴	NT ~
	IR	HCI ↴	MVU ~	ELE ⮥	POP ~	MFU ~	C_AG ~	SLP ~	PT ~	C_UB ~	TMA ↴
Etheostoma flabellare	NR	PPT ⮥	BFI ⮥	TMI ⮥	TMA ↴	HCI ⮥	ELE ⮥	POP ~	TM ⮥	SLP ~	MFU ~
	RR	SINU ~	C_AG ↴	BFI ⮥	PPT ~	NT ~	TMI ↴	MVU ⮥	POP ~	MFU ~	PT ~
Macrhybopsis hyostoma	BR	BFI ↴	SLP ↴	MVU ↴	TMA ⮥	POP ~	NT ~	MFU ~	TMI ⮥	C_AG ↴	BFI ↴
	CR	MFU ⮥	TMA ↴	TMI ~	ELE ⮥	MVU ~	BFI ⮥	MVU ~	PT ~	NT ⮥	C_UB ~
Percina roanoka	NR	HCI ↴	MVU ~	SINU ~	SLP ↴	TM ⮥	PT ~	C_AG ↴	NT ~	TMI ⮥	PPT ⮥
	RR	TM ~	PPT ~	SLP ~	SINU ⮥	TMI ↴	SLP ~	POP ~	SINU ⮥	C_UB ⮥	NT ~
Chrosomus oreas	NR	C_AG ~	HCI ⮥	BFI ↴	TMA ⮥	ELE ↴	TMI ↴	POP ~	PPT ↴	TM ⮥	SO ⮥
	RR	ELE ⮥	TMA ↴	TMI ↴	BFI ⮥	MVU ⮥	TM ⮥	MFU ~	C_AG ~	SLP ⮥	PT ↴

Table B.2. Comparing the performance of non–spatial models and spatial models in terms of among–basins transferability AUC (the area under the receiver–operating–characteristic curve) for 5 fish species in the New River (NR), Illinois River (IR) and RR–Roanoke River (RR). These five species were re-examined in the spatial models because they had poor transferability (AUC < 0.5) in the three non–spatial models (GLM–logistic model, BRT–boosted regression trees, and MaxEnt).

	Basin		Non–spatial			Spatial		
Species name	Model	Prediction	GLM	BRT	MaxEnt	GLM	BRT	MaxEnt
Cyprinella spiloptera	NR	IR	0.510	0.474	0.473	0.510	0.519	0.527
Nocomis leptocephalus	NR	RR	0.477	0.481	0.489	0.513	0.506	0.509
Campostoma anomalum	NR	RR	0.505	0.471	0.500	0.524	0.554	0.508
Percina roanoka	NR	RR	0.420	0.453	0.441	0.526	0.518	0.533
Chrosomus oreas	NR	RR	0.453	0.483	0.518	0.499	0.510	0.554

Figure B.1. Partial dependence plot in the boosted regression trees showing the unimodal relationship between the occurrence of *Chrosomus oreas* (Mountain redbelly dace) and elevation. The optimal elevation for *C. oreas*, between 550 and 1000 m, is shown by combining the data from the New River basin and Roanoke River basin.



Chapter 4: Temporal transferability of stream fish distribution models: can uncalibrated SDMs predict distribution shifts under climate change?

Jian Huang[1], Emmanuel A. Frimpong[1*]

[1]Department of Fish and Wildlife Conservation, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

**Correspondence*: Emmanuel A. Frimpong, PhD, tel. +540 231 6881, fax +540 231 7580, e-mail: frimp@vt.edu

Abstract. Species distribution models (SDMs) have been widely used to assess the impact of climate and landscape changes. However, accuracy of predications of SDMs with independent data is seldom assessed. The SDMs to be used for projection over time should be well calibrated with high discrimination power. Here, we used logistic regression, boosted regression trees, and MaxEnt to model the habitat suitability of 16 fish species in the New River basin based on historical species occurrence obtained during 1950-1990. We evaluated the transferability of these SDMs in terms of discrimination power and calibration based on two independent testing datasets. One testing dataset was obtained by single-visit sampling during 1995-2010 and the other by multi-year occupancy-based sampling during 2012-2014. Annual average temperature in the New River basin has increased by about 0.5 °C from 1960s to 2010s. Discrimination power of the SDMs was moderate to good in the evaluations with the independent datasets, with AUC > 0.6 for 14 (87.5%) of species in the logistic regression and boosted regression trees. With observed species prevalence (frequency of occurrence) as the discrimination cutoff, logistic regression had the highest overall accuracy for 13 of the 16 species and highest specificity for 10

species, whereas MaxEnt had the highest sensitivity for 14 species. Biases and misspecified spread of predicted probability of species occurrence were common in the temporal model transfers, no matter what modeling approach and sources of testing data were used. We suggest reclassifying predicted probability of species occurrence (prevalence) to ordinal ranks to deal with (under- and over- estimation) bias. Inappropriate spread of predictions could be remedied by using training datasets with large sample size and good coverage of environmental gradients, including proximal habitat factors, and fine-tuning variable selection with regularization or cross validation.

Key words: Species distribution model, occupancy model, discrimination power, calibration, temporal transferability, climate change, New River, fish

INTRODUCTION

Climate change has been recognized as one of the most influential disturbances to alter species distributions (Thuiller, 2007; Brown *et al.*, 2011). Further global warming would profoundly impact ectothermic taxa whose growth, recruitment, and dispersal are influenced by temperature (Buisson *et al.*, 2008). Climate change has notably caused distribution shifts of many cold-water fish species with physiological optima less than 20°C (Britton *et al.*, 2010), and would potentially threaten other species with low ability of expansion and adaptation. In a study examining the effects of simulated global warming (i.e., water temperature increase by 0.8-4 °C by the year 2060) on the distribution of 50 stream fishes in the US Midwest, Lyons *et al.* (2010) predicted that all 19 cold-or cool-water fishes examined were predicted to decline but 23 of 31 warm-water fishes to increase in distribution. Species distribution models (SDMs) are commonly

used quantitative approaches to predict spatial distribution shifts under climate change scenarios (Steen *et al.*, 2010; Brown *et al.*, 2011). Successful predictive SDMs could provide critical information for ecologists and conservation agencies to delineate suitable habitats for species at risk (Elith *et al.*, 2006; Sindt *et al.*, 2012) and enable stakeholders and managers to prevent economic hardship by altering stocking policies (Flebbe et al. 2006).

However, one step researchers usually neglect before predicting future spatial distribution under climate and landform change is to assess the temporal transferability of SDMs developed using historical or current observations. Assessing transferability (or generality) is the process to evaluate how accurately and reliably SDMs predict species distribution in a different region or time frame (Randin *et al.*, 2006; Elith & Leathwick, 2009; Wenger & Olden, 2012). In the assessment of transferability, SDMs are firstly developed based on training datasets, and their performance in the predictions on independent testing datasets are measured in terms of accuracy and precision. Spatial transferability is evaluated if the testing datasets and training datasets are sampled from different regions in a similar time frame, whereas temporally transferability is measured when the testing datasets and training datasets are sampled in different time frames.

We assessed temporal transferability in this study since we transferred species distribution model among temporally independent datasets. Transferring species distribution models over time is widely recognized as risky because these applications inevitably involve extrapolation into new environmental space (Elith & Leathwick, 2009; Schibalski *et al.*, 2014). The predictions of species occurrence could be unreliable if the environmental gradients and species–habitat relationships vary unpredictably among the examined time frames (Strauss & Biedermann, 2007; Elith & Leathwick, 2009). In our previous study (Huang & Frimpong, in review), we found that spatial transfer of models can be constrained by the mismatches in the

range and location of the environmental gradients in the training region and prediction region. These same constraints might also hinder temporal transference of SDMs. Additionally, transferability depends on model type (Meynard & Quinn, 2007; Wenger & Olden, 2012), the quality of training data (Strauss & Biedermann, 2007; Wang & Jackson, 2014), and species traits (Bulluck *et al.*, 2006; Randin *et al.*, 2006).

Most previous studies (e.g., Meynard & Quinn, 2007; Strauss & Biedermann, 2007; Wenger & Olden, 2012) assessed the temporal or spatial transferability of species distribution models by means of discrimination power, i.e., a model's ability to discriminate occupied and unoccupied sites (Pearce & Ferrier, 2000). Sensitivity (true positive or presence rate), specificity (true negative or absence rate), accuracy (correct classification rate), and AUC (area under the Receiver-Operating-Characteristic (ROC) curve) are commonly used measures of discrimination power. Calibration that describes a more stringent agreement between predictions and observations (Pearce & Ferrier, 2000) should also be included in the assessment of temporal model transferability (Strauss & Biedermann, 2007). Pearce and Ferrier (2000) viewed calibration as a process to evaluate a model's reliability in predicting the probability of a site being occupied. In a well calibrated model, if the predicted probabilities of species presence was between 0.5~0.7 for 100 sites, then approximately 60 sites would actually be occupied by the species. Two measurable components, bias and spread are evaluated in model calibration. Calibration is necessary because even a model with large discrimination power could have the problem of bias and misspecification in the spread of predictions of the probability of species occurrence. In a regression of the observed presence/absence against the logit (predicted probability of presence) in a logistic model, the intercept and slope term correspond to the bias

and specification respectively (Pearce & Ferrier, 2000). Likelihood ratio tests can then be used to examine whether the intercept and slope deviate from 0 and 1, respectively (Miller *et al.*, 1991).

Depending on the source of testing data, evaluations of model discrimination power and calibration have several forms with varied efficacies. In cross validation, the testing datasets are iteratively sampled from the original data and predicted by the models developed based on training datasets. Model performance measures such as overall accuracy and AUC described above then can be obtained by relating the observed and predicted values of the testing datasets. In external evaluations, the testing datasets are obtained independent of the training data, for example, from different crews, time frames or regions. Most studies used independent testing data from single-visit samplings (Randin *et al.*, 2006; Tuanmu *et al.*, 2011; Wang & Jackson, 2014). Alternatively, and perhaps more robustly, independent testing data from occupancy-based sampling can be used in the assessment of model transferability. In occupancy-based sampling, temporally replicated observations of species presence/absence at each site are obtained, which allows separation of the probability that a species is present (occupancy) from the conditional probability of detecting the species (MacKenzie *et al.*, 2002). However, occupancy-based design requires intensive time and labor. Thus, it is worth comparing model evaluation based on occupancy-based and single-visit, independent datasets because if these two approaches yield equivalent results then we can have more confidence in the single-visit evaluations which are less costly in time and labor.

In this study, we evaluate the temporal transferability of species distribution models for 16 fish species in the New River basin (Virginia, West Virginia, and North Carolina, USA) in terms of discrimination power (sensitivity, specificity, overall accuracy, and AUC) and calibration (bias and spread). We take a retrospective approach by predicting current spatial

species distribution with SDMs developed based on historical data. Different from simulation based studies, climate and landscape data and biological responses in this study are real, including biological data acquired from independent sources or our field surveys. The temporal transferability of three widely used statistical models (logistic regression, boosted regression trees, and MaxEnt) are compared in terms of discrimination power and reliability of predicted probability of species presence in two independent testing datasets. One testing dataset contained single-visit samples collected during 1995-2010 (Esselman *et al.*, 2013) and the other contained multi-year occupancy-based samples collected during 2012-2014.

METHODS

Study system and species

The New River (Figure 4.1) originates in North Carolina, heads north and drains through southwestern Virginia and West Virginia for a total drainage area of 21,700 km^2 (Jenkins & Burkhead, 1994). The New River is heterogeneous geologically as it spans four types of major eastern U.S. physiographic provinces: Blue Ridge, Piedmont, Valley-and-Ridge Appalachia and Appalachian Plateaus (Fenneman, 1946). At the Kanawha Falls, the New River and Gauley River merge to form the Kanawha River, which then flows into the Ohio River and Mississippi River systems. Owing to isolation and habitat heterogeneity, the New River has maintained a particularly high proportion of endemic fish species. We modeled the distribution and tested our hypotheses on 16 fish species (Table 4.1) found in the wadeable streams in the New River Basin. These 16 species are diverse in traits (e.g., body size, habitat preference) and rarity (Pritt & Frimpong, 2009); six of them (*Notropis scabriceps*, *Percina gymnocephala*, *Etheostoma osburni*, *Nocomis platyrhynchus*, *Etheostoma kanawhae*, and *Phenacobius teretulus*) are New River endemic species. *Micropterus salmoides*, *Micropterus dolomieu*, and *Ambloplites rupestris* are

warm-water species, *Oncorhynchus mykiss, Salmo trutta* and *Salvelinus fontinalis* are cold-water species, while the other 10 species are cool-water species *(*Cherry *et al.* 1975, Cherry *et al.* 1977, Shingleton *et al.* 1981, Beitinger 2000, Lyons *et al.* 2009). We determined the thermal classification of three endemic species, *Phenacobius teretulus, Etheostoma kanawhae* and *Percina gymnocephala*, based on their habitat preferences (Menhinick 1991, Jenkins & Burkhead 1994, Stauffer et al. 1995) and our field observations since their temperature tolerance have not been well documented in the literature. The 16 species examined in our study, except for warm-water species (*A. rupestris, M. dolomieu*, and *M. salmoides*), were hypothesized to move northward or into higher elevation streams given only water warming, although this may be compounded by disruptions in landscape connectivity (e.g., waterfalls and dams) and future landscape transformations (e.g., urbanization, deforestation).

Developing species distribution models

Historical species occurrence data (2,269 presence records, and 14,291 absences) used in the models were extracted from the *IchthyMaps* database, a public database that contains fish occurrence records sampled primarily during 1950 to 1990 in the United States (Frimpong, Huang, Liang & Ostroff, in review). These species occurrences were spatially joined to the 1,035 National Hydrography Dataset (NHD) inter-confluence segments (study unit) and linked to habitat features. Each stream segment has been reported to contain at least two non-game fish species. Game species are treated separately in this dataset because their occurrences are usually surveyed in targeted efforts by anglers and agencies rather than in community-based sampling designs. The presences of game species are therefore uninformative in determining the absence of non-game species that are usually sampled as a community, whereas the converse is not true.

Fourteen habitat factors assumed to be critical for fish distribution are listed in the Table 4.2. The climate data (e.g., temperature, precipitation) during 1950 to 2014 were retrieved from the PRISM climate group (2004). We took the 20-year mean for the annual minimum temperature, maximum temperature, mean temperature, and annual total precipitation. The values of habitat factors in Year 1961-1980 were used together with historical fish occurrence data to develop species distribution models, while the habitat data in year 1995-2014 were used with current (1995-2014) fish occurrence data in model evaluations. The land cover data in 1980's (Price *et al.*, 2006) and 2006 (Fry *et al.*, 2011), respectively corresponding to historical and current fish occurrence data, were obtained from the United States Geological Survey (USGS) Land Cover Institute. The habitat condition index (HCI) was retrieved from the National Fish Habitat Action Plan (NFHAP) databases (National Fish Habitat Board, 2012), and other variables were retrieved from NHDplusV1 and NHDplusV2 (EPA, USGS and Horizon Systems Corporations, 2010, 2012). Enduring variables such as elevation, slope and stream order remain unchanged during the two periods (i.e., 1961-1980, 1995-2014). The spatial autocorrelation in the habitat factors was filtered by the principal coordinates of neighbor matrices (PCNM; Borcard & Legendre, 2002). 'Spatialized' habitat factors (predicted by PCNM spatial eigenvectors in a multivariate regression model) were used as predictors in the logistic regression, BRT, and MaxEnt.

Based on the historical fish occurrence data and habitat factors, we developed species distribution models for each of the 16 species using three approaches, logistic regression under the Lasso regularization (Tibshirani, 1996), boosted regression trees (BRT) model (Friedman, 2001), and maximum entropy presence–only model (MaxEnt; Phillips *et al.*, 2006). BRT and MaxEnt are widely used in the studies of species distribution models (e.g., Elith *et al.*, 2006).

The Lasso (least absolute shrinkage and selection operator) is used to regularize logistic regressions through adding a constraint (or penalty) of absolute sum of coefficients to the likelihood optimization (Friedman *et al.*, 2010). The penalty parameter can be tuned in the cross validation to balance model accuracy and simplicity (Tibshirani, 1996). We implemented the Lasso-regularized logistic regression with the package 'glmnet' (Friedman *et al.*, 2010), and BRT and MaxEnt models with the package 'dismo' in the R program (R core team, 2014). AUC values of these 3 types of models were compared in the Friedman Rank Sum Test which is a nonparametric one-way analysis of variance for repeated measures developed by Friedman (1937). Statistics in this test are calculated in terms of rank, so the assumption of normality in the response is not required. Null hypothesis of this test is that these three types of models (logistic regression, BRT and MaxEnt) have similar discrimination power in terms of AUC.

Discrimination power

The performance of these three types of species distribution models were measured by sensitivity, specificity, overall accuracy (AUC). Sensitivity, specificity and overall accuracy are threshold-dependent measures in which a threshold of predicted probability of species presence ($\hat{p}$) is required to determine the species' occurrence at each site. The observed prevalence for each species (proportion of presences in all observations) was used as cutoff to calculate sensitivity, specificity and overall accuracy (Franklin & Miller, 2009). For example, if the observed prevalence for a species is 0.23, a stream segment with $\hat{p} \geq 0.23$ would be classified as occupied site. The Receiver-Operating-Characteristic (ROC) curve is created by plotting the sensitivity against (1- specificity) as discrimination threshold changes from 0 to 1. The AUC (which is a threshold-independent measure) is calculated by integrating the area under the ROC curve. A chance model has a mean AUC of 0.5. The classification proposed by Swets (1988) has

been routinely used to evaluate AUC in model training: AUC between 0.7-0.9 indicates moderate discrimination power and between 0.9-1 is considered excellent. However, a lower threshold, for example AUC of 0.6 - 0.7 (Randin *et al.*, 2006; Strauss & Biedermann, 2007), is often used to distinguish transferable from non-transferable SDMs. AUC was also calculated in the model training, 5-fold cross validation, and external evaluations with the two independent datasets. The procedure of 5–fold cross validation was: 1) the original dataset was randomly partitioned into 5 subsets of equal size, 2) each subset was retained as testing data to validate the model developed based on the remaining 4 subsets, 3) and the mean of 5 AUC values was calculated for each species.

Model calibration

We measured bias and spread using Cox (1958)'s approach for model calibration (Figure 4.2). The observed occurrences (π_i *or* $p(y_i = 1)$) in the single-visit samples were regressed against the logit of the predicted probability of presence ($\hat{p}_i$) in a logistic regression for model calibration (Fiugre C.1):

$$Logit(\pi_i) = a + b * Logit\,(\hat{p}_i) \qquad (1)$$

In the testing based on occupancy sampling design, we used the linear model:

$$\psi_j = a + b * \hat{p}_j \qquad (2)$$

In the model (1) and (2), *a* is intercept, *b* is slope, π_i is the observed probability of species presence at site *i* in the (single-visit) testing data, ψ_i is the probability of species occurrence estimated at site *j* in the occupancy-based testing data and is assumed to be true, $\hat{p}_i$ and $\hat{p}_j$ are predicted probability of species presence at each site. We then tested whether intercept (*a*) and slope (*b*) term of the calibration line significantly deviated from zero and one respectively using three likelihood ratio tests (Miller *et al.*, 1991; Pearce & Ferrier, 2000) at significance level of

0.01: Test 1: S_1 = Deviance (a=0, b=1) - Deviance ($a=\hat{a}$, $b=\hat{b}$); Test 2: S_2 = Deviance (a=0, b=1) - Deviance ($a=\hat{a}$, b=1); Test 3: S_3 = Deviance ($a=\hat{a}$, b=1) - Deviance ($a=\hat{a}$, $b=\hat{b}$). We used significance level of 0.01 instead of 0.05 to control the type I (false positive) error in the multiple hypothesis tests. The statistic S_1, S_2, and S_3 follow X^2 distribution with 2, 1, and 1 degree of freedom respectively. The null hypothesis and alternative hypothesis of these tests are shown in the Figure 4.2. In the deviance, $\hat{a}$ and $\hat{b}$ are the fitted values of parameters in the model (1) or (2). If bias (i.e., consistent over- or under- estimation of the probability of species presence) occurs, the intercept would significantly deviate from 0. In the case of misspecified spread, i.e., the predicted probabilities spread extremely (close to 0 or 1) or clumpy, the slope term would significantly deviate from 1.

Evaluations with single-visit samples

The species distribution models were evaluated in the independent single-visit testing data (sampled during 1995 and 2010) compiled by NFHAP (Esselman *et al.*, 2013). The current spatial distribution of 16 fish species were predicted by the species distribution models built based on historical data. Discrimination power (e.g., sensitivity, specificity, overall accuracy, and AUC) and calibration (bias and spread) were evaluated based on predicted and observed fish occurrence.

Evaluations with occupancy-based samples

We conducted occupancy-based field survey to evaluate the performance of species distribution models developed by logistic regression, BRT and MaxEnt. A total of 80 stream segments (sites) were sampled in a 3-year occupancy-based survey (Figure 4.1). These 80 segments are overall evenly distributed across the New River basin, and approximately 1/3 of them are located in each of the three major physiographic provinces. These segments are all

wadeable in the summer, but represent a wide range of stream size, geomorphology, landscape characteristics and disturbance. Each segment was sampled with a single-pass electrofishing during May to August. The sampling crew included 4-5 members equipped with 2 backpack electrofishers (Smith-Root LR-24) and four nets (14×17×8 inch trapezoid, 0.25 inch Mesh).

We used a basic occupancy model (MacKenzie *et al.*, 2002) to estimate the probability of species presence in the R package "unmarked" (Fiske & Chandler, 2011). This occupancy model assumes that the population is closed so the estimated probability of occupancy would not change during the study years. We used the habitat factors with 1995-2014 values (Table 4.2) as covariates to estimate site-specific probability of presence, and used water temperature (°C), discharge (cfs) and sampling effort (seconds) as sampling covariates to estimate site- and time-specific probability of detection. Water temperature and sampling effort were recorded after sampling at each site. To estimate discharge at a site, we first linked each site with its downstream USGS gage station (http://waterdata.usgs.gov/nwis/rt). The discharge at a site was estimated as (real-time discharge at USGS gage) × (watershed area of site) / (watershed area of USGS gage). For each species, we compared and ranked models using AIC (Akaike Information Criterion). The site-specific probabilities of presence for each species were predicted by the occupancy models weighted by AIC.

RESULTS

Evaluations with the training datasets

The spatial distribution of 16 fish species were well described in the model development using historical training data with moderate to high AUC values, particularly in the BRT and MaxEnt models. AUC was higher than 0.75 for all species in these two models. In general, performance in terms of AUC decreased in the 5-fold cross validation for all three types of

models. Temperature measures (annual minimum, maximum and mean during 1961-1980) were key predictors and were found negatively related to the occurrence of 10 species (*Notropis scabriceps, Phenacobius teretulus, Etheostoma osburni, Percina gymnocephala, Cottus kanawhae, Rhinichthys atratulus, Chrosomus oreas, Noturus insignis, Micropterus salmoides,* and *Oncorhynchus mykiss*) in the partial dependence plots of BRT models (Table C.1). Other important habitat factors included: BFI (base flow index), stream order, and elevation. Four species (*N. scabriceps, P. teretulus, Etheostoma kanawhae,* and *P. gymnocephala*) showed positive relations with BFI while *Etheostoma osburni, Nocomis platyrhynchus, C. kanawhae, R. atratulus,* and *Ambloplites rupestris* showed negative relationship with BFI. Seven species (*N. scabriceps, E. kanawhae, P. gymnocephala, N. platyrhynchus, A. rupestris, Micropterus dolomieu,* and *Micropterus salmoides*) favored moderate to large streams. Two trout species, *Salvelinus fontinalis* and *Salmo trutta,* showed negative relationships with stream order where *S. fontinalis* occupied stream with higher elevation and denser forest cover.

Evaluations with single-visit testing samples

The temporal transferability of species distribution models varied among species (Figure 4.3, Figure C.2), according to the evaluation based on the single-visit testing data. Generally, cool water species showed the best temporal transferability. For *E. osburni, C. kanawhae,* and *N. insignis*, the performance of the three types of models were consistently high (AUC > 0.9). Good temporal transferability was also observed for *N. scabriceps, N. platyrhynchus,* and *A. rupestris.* The model transferability in terms of AUC for all three trout species (*S. fontinalis, S. trutta,* and *O. mykiss*) was poor (Figure 4.3). Friedman Rank Sum Test showed that the discrimination power of the performance of logistic regression, BRT and MaxEnt models were not different in the testing based single-visit samples (p-value = 0.13). Overall accuracy was largely consistent

with AUC in the model transfers: SDMs were most transferable for *E. osburni*, *C. kanawhae*, and *N. insignis*, and least transferable for three trout species. The logistic regression showed highest accuracy for 13 out of 16 species (Figure C.2). For most species, the MaxEnt models had highest sensitivity, but lowest specificity, suggesting that MaxEnt models tend to make positive prediction "boldly" at the cost of high false absence rate. Conversely, the logistic regressions had lowest sensitivity but highest specificity for most species. The sum of sensitivity and specificity of BRT was the highest of the three model types for 10 of the 16 species.

Overall the SDMs were not well calibrated in terms of bias and spread in the independent testing datasets (Table 4.3). The number of poorly calibrated models in terms of [bias, spread] were [10, 6], [10, 7], [11, 6] respectively for logistic regressions, BRT, and MaxEnt. Only *C. kanawhae* was well calibrated in all three models. Six species (*P. teretulus, E. osburni, P. gymnocephala, N. insignis, C. oreas* and *M. salmoides*) were well calibrated in two models. In the MaxEnt models, overestimation bias and clumped spread of predictions were common, corresponding to their high sensitivity but low specificity.

Evaluations with occupancy-based testing samples

Based on data collected by occupancy-based design during 2012-2014, we estimated the probability of species presence and probability of detection. To control the number of the candidate occupancy models for each species, we limited the number of habitat covariates to be five at most, and number of sampling covariate to be 0 or 1. There are 14 habitat covariates (Table 4.2), and three sampling covariates (sampling effort, discharge, and water temperature at sampling), so the number of candidate occupancy models for each species is 61,208, i.e., $4 * \left[\sum_{k=0}^{5}\binom{14}{k}\right]$ where $\binom{14}{k}$ means the combination of selecting *k* out of 14 habitat covariates. The 'best' model for each species was identified by comparing AIC of 61,208 models with varied

combination of habitat covariates and sampling covariates. The predicted probability of species occurrence was predicted by averaging the occupancy models based on AIC weights. The habitat covariates and sampling covariates in the 'best' model with the smallest AIC value were used to evaluate species-habitat relationships. Temperature (maximum and mean), stream order, BFI, % urban and % agriculture in the watershed were the important habitat covariates for the 16 fish species (Table 4.1). Discharge was the key factor that affected the detection of nine species (e.g., *N. scabriceps, E. kanawhae,* and *E. osburni*). Except for *C. kanawhae,* the New River fish examined were negatively associated with % urban in the watershed and habitat condition index. Cool- and cold- water species such as *S. fontinalis, N. scabriceps, P. teretulus, E. osburni* and *P. gymnocephala* favored elevated mountain streams with low maximum July temperature, which were consistent with the patterns revealed in the partial dependence plot based on historical data (Table C.1). In contrast, *M. dolomieu* and *N.insignis* favored warmer streams with moderate agricultural disturbance. The three trout species (*S. fontinalis, S. trutta,* and *O. mykiss*) favored larger pristine mountain streams.

The temporal transferability measured in the occupancy-based samples also varied among species (Figure 4.3, Figure C.2). Good transferability in terms of AUC was found for *E. osburni* and *E. kanawhae* in the BRT and logistic regression. According to the partial dependence plots (Figure C.3) and histogram (Figure C.4), we found that cool water species such as *E. osburni* showed negative response to the increase of temperature and such pattern was consistent between 1961-1980 data and 1995-2014 data. For other species, the AUC in the occupancy-based evaluation ranged from 0.5 to 0.8. Friedman's Rank Sum Test showed that the discrimination power (AUC) of BRT was significantly better than MaxEnt in the occupancy-based evaluation (p-value = 0.022). Consistent with the independent evaluation, the logistic

regression had highest accuracy and specificity while the MaxEnt model had the highest sensitivity for most species in the occupancy-based evaluation.

In the occupancy-based evaluation, the number of poorly calibrated models in terms of [bias, spread] were [12, 6], [9, 3], [10, 11] respectively for logistic regression, BRT, and MaxEnt. The frequency of underestimating and overestimating probability of species occurrence were largely even. Clumpy spread in the prediction of probability of species presence was more frequent; namely, underestimation occurred when the 'true' probability of species occurrence was > 0.5 while overestimation occurred when the 'true' probability of species occurrence was < 0.5 (Pearce & Ferrier, 2000). Generally, the BRT models were well calibrated for the six New River endemic species (Table 4.3).

DISCUSSION

Adequate predictive species distribution model depends on our understanding of patterns (e.g., magnitude, rate, frequency) and impacts of climate changes. Annual average (air) temperature in the New River basin has increased by 0.48 °C during the past six decades (Table 4.2). Climate change impacts ecosystems slower but longer than other disturbances (Thuiller, 2007), but these disturbances could be compounded at multiple spatial and temporal scales (Steen *et al.*, 2008; Lyons *et al.*, 2010; Bond *et al.*, 2011). Temperature measures (annual minimum, maximum and mean) were found negatively related to presence for more than half of the New River fish species modeled in this study. We confirmed that further global warming would be devastating to both cool water species (e.g., *N. scabriceps, P. teretulus, E. osburni, R. atratulus,* and *C. oreas*) and cold-water species, In addition to a general pole-ward shift, these species are predicted to expand upstream with concurrent loss of headwater biodiversity under global warming (Thuiller, 2007; Buisson & Grenouillet, 2009). These fish species may respond

differently to climate and landscape change, owing to variations in their life history, thermal tolerance, biogeographic affinities, mobility, and adaption through modifying phenological and physiological traits (Thomas *et al.*, 2004; Perry *et al.*, 2005). Flebbe *et al.* (2006) also predicted that more than 50% of total trout habitat area in the southern Appalachians would be lost by the year 2100 given two global climate circulation models, and only small refuges in elevated headwater streams would remain, which may pose a great challenge for fisheries managers to find mitigation strategies.

Generally, the discrimination power of species distribution models for New River fish species were moderate to good in the temporal model transfers. Yet, temporal transferability was species-specific: the AUC and overall accuracy were high for few New River endemic species, such as *N. scabriceps* and *E. osburni*. The fish-habitat relationships for the endemic species were more accurately described likely because the whole range of the environmental gradients for the species were evaluated, free of the problem of mismatches in the range and location of variables in the model transfers. Additionally, the spatial distribution and composition of stream fish assemblages are strongly affected by physiographic region and drainage basin (Hocutt & Wiley 1986; Jenkins & Burkhead 1994). Stream fish faunas are distinctive according to physiographic region (Angermeier & Winston 1998; Angermeier & Winston 1999). The endemic species in the New River basin showed different physiographic preference: *Notropis scabriceps* (New River shiner), *Phenacobius teretulus* (Kanawha minnow) and *Percina gymnocephala* (Appalachia darter) are often common in the upper section of the New River Basin within the Blue Ridge province but usually rare or uncommon in the Valley and Ridge province (Jenkins & Burkhead 1994); *Etheostoma kanawhae* (Kanawha darter) is restricted to the Blue Ridge Province while *Etheostoma osburni* (Candy darter) is restricted to Valley and Ridge and Appalachian Plateau

province (Jenkins & Burkhead 1994). Physiographic provinces can be used as regional units in hierarchical model because of the strong impacts of land forms on species distribution, habitat quality and disturbance patterns.

Discrimination power (AUC and overall accuracy) was low for three trout species (*S. fontinalis, S. trutta,* and *O. mykiss*), *M. salmoides,* and *M. dolomieu.* For more widely-spread species (e.g., *M. salmoides,* and *M. delomieu*), New River might only cover a small proportion of the whole distribution range, thus the fish-habitat relationships described in the SDMs might be unreliable (Jackson *et al.*, 2001). It has been widely recognized that distribution of generalist species, such as *M. salmoides* and *O. mykiss*, are particularly difficult to predict (Brotons *et al.*, 2004; Guisan *et al.*, 2007) and transfer over time and space (Randin *et al.*, 2006; Strauss & Biedermann, 2007; Schibalski et al., 2014). Additionally, the spatial distribution and abundance of the game species in this study, particularly the three trout species in the New River drainage, might be reflecting management interventions (e.g., stocking strategies, capture-return policy) of the fish and wildlife agencies in the three states of the basin (Virginia, West Virginia, and North Carolina).

Bias and misspecified spread in the predicted probability of occurrence of the New River fish species were common in the external evaluation based on either the single-visit testing data or occupancy-based testing data (Table 4.3). Pearce and Ferrier (2000) attributed the bias in the estimations to the inconsistent species prevalence between the training and testing data. Considerable bias and misspecified spread in predictions of probability of species occurrence were also observed in the temporal model transfers for phytophagous insects (Strauss & Biedermann, 2007). We found that, if the prevalence of training dataset was higher than occupancy-based testing samples, then the model transferred would likely suffer from

overestimation bias; otherwise underestimation tended to occur in model transfers (Table 4.3). There might be some other reasons causing poor calibration when a model is applied to new data. The observed prevalence might not be a good estimator of the 'true' species prevalence. Another major explanation for limited transferability is the inherent difficulty in extrapolating SDMs to novel environment (Elith & Leathwick, 2009; Schibalski *et al.*, 2014). After all, conventional species distribution models are static and they may sensitively capture the changes in the predictors, but they could not explicitly account for population dynamics or dispersal (Guisan & Zimmermann, 2000; Bulluck *et al.*, 2006), interactions of habitat factors, and dynamic species-habitat relationships.

If overestimation occurs, predictions of the probability of species presence need to be adjusted to lower values or reclassified to ordinal scale (e.g., probability of species presence: low, slight, moderate, high, and extremely high); otherwise, for example, an actual probability of occurrence of 0.4 might be represented by an estimated probability of 0.8. Using ordinal scale for habitat suitability was implemented in Ottaviani *et al.* (2004) and Vaughan and Ormerod (2005), and also recommended by Strauss and Biedermann (2007). In the case of underestimation, predictions of the probability of species presence need to be adjusted to larger values or to ordinal values. The major cause of misspecified spread of the predictions is over-fitting and under-fitting (Pearce & Ferrier, 2000). Over-fitting creates the extreme spread of predictions of presence probability, either close to 0 or 1. Models with too many predictors or complex high-order terms tend to have problems of over-fitting. Over-fitting was not very common in the logistic regression and BRT, likely because these two approaches have built-in cross validation and the complexity was further controlled by Lasso-regularization and simple classifiers respectively in this study. However, we found that many models had the problem of lack-of-fit,

particularly the MaxEnt models. Under-fitting resulted in the clumped spread of predictions of presence probabilities, namely, underestimating when the actual probability is > 0.5, and overestimating when the actual probability is < 0.5. When under-fitting occurs, one can remedy by loosening the penalization (i.e., choose smaller penalty parameter in the tuning process) in the Lasso, increasing the complexity of trees or decreasing learning rate in the BRT, and using higher-order and non-linear terms in the MaxEnt. It is worth noting that SDMs, constrained by the availability of testing data, unlikely can be calibrated to suit each climate and landform scenarios for each species. A more efficient approach might be generalizing the associations of model transferability and biological/ecological trait (Strauss & Biedermann, 2007), rarity and management strategies of species.

The assessment of model generality in terms of discrimination power and calibration would be affected by the source of training data and testing data. Using training dataset with large sample size is one the most efficient ways to develop accurate and reliable SDMs to be transferred over time or space. Our SDMs based on training data of 1,035 samples had moderate to good discrimination power. The assessment of model performance has been predominantly based on internal evaluations such as model fitting and cross validation. However, we found that conclusions on the 'best' model in the cross validation and external evaluations might be inconsistent (Figure 4.3). This emphasizes the need for assessing temporal transferability of SDMs under climate and landscape change because good performance in cross validation does not guarantee transferable models. Additionally, the temporal transferability of SDMs evaluated based on single-visit samples and occupancy-based samples were not consistent for some species, for example, *N. platyrhynchus* and *E. kanawhae.* Some inconsistency in the model evaluations are expected because sample size, species prevalence, sampling approach, and

environmental gradients covered might be different among the testing datasets. The occupancy-based testing data would be more desirable because of its higher accuracy determining in presence and absence after accounting for detection rate. Yet, a tradeoff of such decomposition of probabilities is that intensive time and labor are required in the occupancy-based design. Occupancy-based designs need temporally replicated sampling over seasons or years, rendering such designs easily constrained by funding and logistical challenges such as maintaining the same crew. Under most logistical situations, the single-visit models appear more practical than the occupancy models. A comparative study (Welsh *et al.*, 2013) indicated that adjusting non-detection with occupancy models would not necessarily improve accuracy and precision in estimation, compared to ignoring non-detection completely, particularly when the data are sparse.

Discrimination power and calibration could be improved by fine-tuning the set of variables in the model. Both Sindt *et al.* (2012) and Onikura *et al.* (2012) agreed that the main cause of lower accuracy from SDMs development to model validation is the unavailability or exclusion of local habitat predictors. Austin and Van Niel (2011) and Franklin *et al.* (2013) suggested that local habitat factors (e.g., light, geomorphology) need to be included in predictive SDM under climate change scenarios. Performance of SDMs also relies on how well we could account for the effects of landscape configuration (e.g., connectivity, heterogeneity) and spatial autocorrelation (Elith *et al.*, 2006; Zimmermann *et al.*, 2010). Several studies have indicated that landscape composition and configuration are important predictors of species spatial distribution in the climate change context (Steen *et al.*, 2008; Franklin *et al.*, 2013). Comparatively, the studies that addressed single type of disturbance would likely estimate current distribution range with bias and misspecified spread due to the exclusions of other constraints in the fundamental

niche of the focal species, and would be even more misleading when the study objective was predicting future ecological patterns (Onikura, 2012; Sindt *et al.*, 2012). The commonly used approach that evaluates species' response purely along environmental gradients (i.e., without considering spatial autocorrelation) in the SDMs could only capture a portion of factors and mechanisms that drive current species distribution and future shift (Borcard & Legendre, 2002). SDM frameworks that suit hierarchical structure and stream connectivity and incorporate cutting-edge geospatial methodologies are needed to fine-tune the predictions of distribution changes (e.g., fragmentation of population) and shifts (e.g., upstream-ward, north-ward) of different species and of different causes.

ACKNOWLEDGEMENTS

This work was funded by the US Geological Survey Aquatic gap analysis program. A special thanks to Brandon Peoples, Stephen Floyd, Joe Buckwalter, Steve Watkins, Caitlin Worsham and numerous summer interns for help with fish sampling in the New River during 2012-2014. We also thank Dr. Donald Orth for valuable suggestions in the manuscript.

REFERENCES

Angermeier PL, Winston MR (1998) Local vs. regional influences on local diversity in stream fish communities of Virginia. *Ecology*, 79, 911-927.

Austin MP, Van Niel KP (2011) Improving species distribution models for climate change studies: variable selection and scale. *Journal of Biogeography*, 38, 1-8.

Beitinger T, Bennett W, Mccauley R (2000) Temperature tolerances of north american freshwater fishes exposed to dynamic changes in temperature. *Environmental Biology of Fishes*, 58, 237-275.

Bond N, Thomson J, Reich P, Stein J (2011) Using species distribution models to infer potential climate change-induced range shifts of freshwater fish in south-eastern Australia. *Marine and Freshwater Research*, 62, 1043-1061.

Borcard D, Legendre P (2002) All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological modelling*, 153, 51-68.

Britton JR, Cucherousset J, Davies GD, Godard MJ, Copp GH (2010) Non-native fishes and climate change: predicting species responses to warming temperatures in a temperate region. *Freshwater Biology*, 55, 1130-1141.

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

Brown CJ, Schoeman DS, Sydeman WJ *et al.* (2011) Quantitative approaches in climate change ecology. *Global Change Biology*, 17, 3697-3713.

Buisson L, Thuiller W, Lek S, Lim P, Grenouillet G (2008) Climate change hastens the turnover of stream fish assemblages. *Global Change Biology*, 14, 2232-2248.

Buisson L, Grenouillet G (2009) Contrasted impacts of climate change on stream fish assemblages along an environmental gradient. *Diversity, Distributions*, 15, 613-626.

Cherry DS, Dickson KL, Cairns Jr J (1975) Temperatures selected and avoided by fish at various acclimation temperatures. *Journal of the Fisheries Research Board of Canada*, 32, 485-491.

Cherry DS, Dickson KL, Cairns Jr J, Stauffer JR (1977) Preferred, avoided, and lethal temperatures of fish during rising temperature conditions. *Journal of the Fisheries Research Board of Canada*, 34, 239-246.

Cox DR (1958) Two further applications of a model for binary regression. *Biometrika*, 45, 562-565.

Bulluck L, Fleishman E, Betrus C, Blair R (2006) Spatial and temporal variations in species occurrence rate affect the accuracy of occurrence models. *Global Ecology and Biogeography*, 15, 27-38.

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

Elith J, Graham CH, Anderson RP *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29, 129-151.

EPA (Environmental Protection Agency), USGS (United States Geological Survey), Horizon Systems Corporations. (2010) NHDPlusV1 Data. Available at: http://www.horizon–systems.com/nhdplus/data.php

EPA (Environmental Protection Agency), USGS (United States Geological Survey), Horizon Systems Corporations. (2012) NHDplusV2 Data. Available at: http://www.horizon–systems.com/nhdplus/NHDplusV2_data.php.

Esselman PC, Infante DM, Wang L, Wu D, Cooper AR, Taylor WW (2011) An index of cumulative disturbance to river fish habitats of the conterminous United States from landscape anthropogenic activities. *Ecological Restoration*, 29, 133-151.

Esselman PC, Infante DM, Wieferich D *et al.* (2013) National Fish Habitat Action Plan (NFHAP) 2010 community fish data. National Fish Habitat Partnership Data System. Available at: http://dx.doi.org/doi:10.5066/F7QN64RG.

Fenneman NM, Johnson DW (1946) Physiographic divisions of the conterminous U.S. U.S. Geological Survey, Reston, Virginia.

Fiske I, Chandler R (2011) unmarked: An R Package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software*, 43, 1-23.

Flebbe PA, Roghair LD, Bruggink JL (2006) Spatial modeling to project southern Appalachian trout distribution in a warmer climate. *Transactions of the American Fisheries Society,* 135, 1371-1382.

Franklin J, Davis FW, Ikegami M, Syphard AD, Flint LE, Flint AL, Hannah L (2013) Modeling plant species distributions under future climates: how fine scale do climate projections need to be? *Global Change Biology*, 19, 473-483.

Friedman M (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association,* 32, 675–701.

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29, 1189-1232.

Friedman J, Jerome F, Trevor H, Rob T (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33, 1-22.

Fry J, Xian G, Jin S *et al.* (2011) Completion of the 2006 national land cover database for the conterminous United States. *PEandRS*, 77, 858-864.

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

Guisan A, Zimmermann NE, Elith J, Graham CH, Phillips S, Peterson AT (2007) What matters for predicting the occurrences of trees: techniques, data, or species' characteristics? *Ecological Monographs*, 77, 615-630.

Hijmans, RJ, Phillips S, Leathwick J, Elith J (2013) dismo: Species distribution modeling. R package version 0.9-3. Available at: http://CRANR–project.org/package=dismo.

Hocutt CH, Wiley EO (1986) *The zoogeography of North American freshwater fishes,* New York, Wiley.

Jackson DA, Peres-Neto PR, Olden JD (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 RE, Burkhead NM (1994) *Freshwater fishes of Virginia*. American Fisheries Society, Bethesda, Maryland.

Legendre P, Borcard D, Blanchet FG, Dray S (2012) PCNM: MEM spatial eigenfunction and principal coordinate analyses. R package version 2.1-2/r106. Available at: http://R-Forge.R-project.org/projects/sedar/.

Lyons J, Zorn T, Stewart J, Seelbach P, Wehrly K, Wang L (2009) Defining and characterizing coolwater streams and their fish assemblages in Michigan and Wisconsin, USA. North *American Journal of Fisheries Management,* 29, 1130-1151.

Lyons J, Stewart JS, Mitro M (2010) Predicted effects of climate warming on the distribution of 50 stream fishes in Wisconsin, USA *Journal of Fish Biology*, 77, 1867-1898.

MacKenzie DI, Nichols JD, Lachman GB, Droege S, Royle AJ, Langtimm, CA (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83, 2248-2255.

Menhinick EF (1991) *The freshwater fishes of North Carolina*, Raleigh, N.C.; Charlotte, N.C., North Carolina Wildlife Resources Commission ; Distributed by Larkin Distributors.

Meynard CN, Quinn JF (2007) Predicting species distributions: a critical comparison of the most common statistical models using artificial species. *Journal of Biogeography*, 34, 1455-1469.

Miller ME, Hui SL, Tierney WM (1991) Validation techniques for logistic regression models. *Statistics in Medicine*, 10, 1213-1226.

Onikura N, Nakajima J, Miyake T, Kawamura K, Fukuda S (2012) Predicting distributions of seven bitterling fishes in northern Kyushu, Japan. *Ichthyological Research*, 59, 124-133.

Ottaviani DG, Boitani J, Boitani L (2004) Two statistical methods to validate habitat suitability models using presence-only data. *Ecological Modeling*, 179, 417-443.

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

Perry AL, Low PJ, Ellis JR, Reynolds JD (2005) Climate change and distribution shifts in marine fishes. *Science*, 308, 1912-1915.

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190, 231-259.

Price CV, Nakagaki N, Hitt KJ, Clawges RC (2006) Enhanced historical land-use and land-cover data sets of the US Geological Survey, USGS Digital Data Series 240. Available at: http://pubs.usgs.gov/ds/2006/240.

Pritt JJ, Frimpong EA (2010) Quantitative determination of rarity of freshwater fishes and implications for imperiled-species designations. *Conservation Biology*, 24, 1249-1258.

PRISM Climate Group (2004) PRISM Climate Data. Available at: http://prism.oregonstate.edu. Oregon State University, Corvallis, Oregon.

R Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R–project.org/.

Randin CF, Dirnböck T, Dullinger S, Zimmermann NE, Zappa M, Guisan A (2006) Are niche-based species distribution models transferable in space? *Journal of Biogeography*, 33, 1689-1703.

Schibalski A, Lehtonen A, Schröder B (2014) Climate change shifts environmental space and limits transferability of treeline models. *Ecography*, 37, 321-335.

Shingleton MV, Hocutt CH, Stauffer JR (1981) Temperature preference of the New River shiner. *Transactions of the American Fisheries Society*, 110, 660-661.

Sindt AR, Pierce CL, Quist MC (2012) Fish species of greatest conservation need in wadeable iowa streams: current status and effectiveness of aquatic gap program distribution models. *North American Journal of Fisheries Management*, 32, 135-146.

Stauffer JR, Jr., Boltz JM, White LR (1995) The fishes of West Virginia. *Proceedings of the Academy of Natural Sciences of Philadelphia*, 146, 1-389.

Steen PJ, Zorn TG, Seelbach PW, Schaeffer JS (2008) Classification tree models for predicting distributions of michigan stream fish from landscape variables. *Transactions of the American Fisheries Society*, 137, 976-996.

Steen PJ, Wiley MJ, Schaeffer JS (2010) Predicting future changes in muskegon river watershed game fish distributions under future land cover alteration and climate change scenarios. *Transactions of the American Fisheries Society*, 139, 396-412.

Strauss B, Biedermann R (2007) Evaluating temporal and spatial generality: How valid are species–habitat relationship models? *Ecological Modelling*, 204, 104-114.

Swets J (1988) Measuring the accuracy of diagnostic systems. *Science*, 240, 1285-1293.

Thomas CD, Cameron A, Green RE *et al.* (2004) Extinction risk from climate change. *Nature*, 427, 145-148.

Thuiller W (2007) Biodiversity: Climate change and the ecologist. *Nature*, 448, 550-2.

Tibshirani R (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267-288.

Tuanmu M-N, Viña A, Roloff GJ, Liu W, Ouyang Z, Zhang H, Liu J (2011) Temporal transferability of wildlife habitat models: implications for habitat monitoring. *Journal of Biogeography*, 38, 1510-1523.

Vaughan IP, Ormerod SJ (2005) The continuing challenges of testing species distribution models. *Journal of Applied Ecology*, 42, 720-730.

Wang L, Jackson D (2014) Shaping up model transferability and generality of species distribution modeling for predicting invasions: implications from a study on Bythotrephes longimanus. *Biological Invasions*, 16, 2079-2103.

Welsh AH, Lindenmayer DB, Donnelly CF (2013) Fitting and interpreting occupancy models. *PLoS ONE*, 8, e52015.

Wenger SJ, Olden JD (2012) Assessing transferability of ecological models: an underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, 3, 260-267.

Zimmermann NE, Edwards TC, Graham CH, Pearman PB, Svenning J-C (2010) New trends in species distribution modelling. *Ecography*, 33, 985-989.

Table 4.1. Results of occupancy models used to estimate the probability of occurrence of 16 fish species at 80 NHD inter-confluence segments. Habitat covariates and sampling covariates listed are those in the 'best' model with lowest AIC. The '1' in column 'Sample covariate' means that only intercept is fitted in the 'best' model. The descriptions of the habitat covariates are shown in the Table 4.2.

Species	Common name	Habitat covariates	Sampling covariates
Notropis scabriceps	New River shiner	ELE (+) SO(+) HCI (-) PPT (-) L_UB(-)	Discharge (+)
Phenacobius teretulus	Kanawha minnow	ELE (+) SLP (-) SO (+) TMAX(-) TMEAN(+)	Sampling effort (+)
Etheostoma kanawhae	Kanawha darter	TMIN (+) TMEAN(-)	Discharge (+)
Etheostoma osburni	Candy darter	BFI(-) ELE(+) SLP(-) L_UB(-)	Discharge (+)
Percina gymnocephala	Appalachia darter	BFI (+) ELE (+) SO (+) TMAX (-) L_AG (+)	1
Nocomis platyrhynchus	Bigmouth chub	SO (-) HCI (+) TMAX (-) TMEAN (+) L_AG (+)	1
Cottus kanawhae	Kanawha sculpin	HCI (+) TMIN (-) TMAX (+) L_AG (+) L_UB (+)	1
Rhinichthys atratulus	Blacknose dace	HCI (-) L_AG (-)	Discharge (+)
Chrosomus oreas	Mountain redbelly dace	HCI(-) TMIN(+) TMEAN(-) L_AG(-) L_UB(-)	Discharge (+)
Noturus insignis	Margined madtom	TMAX(+) L_AG(+)	Discharge (+)
Ambloplites rupestris	Rock bass	SLP(-) SO(+) L_AG (+)	Discharge (-)
Micropterus dolomieu	Smallmouth bass	TMAX(+) TMEAN(-) L_AG(+)	Water temperature (+)
Micropterus salmoides	Largemouth bass	BFI (+) SLP (+) SINU (-) L_AG (+) L_UB(-)	Discharge (-)
Salvelinus fontinalis	Brook trout	ELE (+) SINU (-)	1
Salmo trutta	Brown trout	BFI(+) TMAX(-) L_AG(-) L_UB(-)	1
Oncorhynchus mykiss	Rainbow trout	BFI(+) SO (+) L_AG(-) L_UB(-)	Discharge (+)

Table 4.2. The list of habitat factors used to develop species distribution models for 16 stream fish species in the New River.

Variable	Source	Descriptions	Mean ± S.D. (Year 1961-1980)	Mean ± S.D. (Year 1995-2014)
BFI	NHDplusV1	The ratio of base flow (i.e., the component of stream flow attributed to ground–water discharge) to total flow (%).	44.49 ± 9.03	44.49 ± 9.03
ELE	NHDplusV2	Mean elevation (meter)	756.39 ± 168.74	756.39 ± 168.74
SLP	NHDplusV2	Mean slope (degrees)	12.39 ± 4.78	12.39 ± 4.78
SO	NHDplusV2	Stream order (Strahler 1952)	1.90 ± 1.25	1.90 ± 1.25
SINU	NHDplusV2	Sinuosity–reach length divided by length of straight line between two nodes.	1.16 ± 0.27	1.16 ± 0.27
HCI	NFHAP	An index of cumulative disturbance in the watershed based on 15 disturbance variables (Esselman et al., 2011). A HCI of 0 indicates pristine condition.	3.20 ± 0.62	3.20 ± 0.62
RL	NFHAP	Total road length in the watershed (km).	3576.18 ± 7150.27	3576.18 ± 7150.27
L_UB	USGS	Percentage of agriculture in the watershed	2.28 ± 9.40	2.94 ± 9.58
L_AG	USGS	Percentage of forest in the watershed	25.67 ± 30.90	26.43 ± 29.49
L_FR	USGS	Percentage of urban in the watershed	69.07 ± 32.60	70.08 ± 25.14
PPT	PRISM	20–year mean annual precipitation (mm).	1101.02 ± 152.94	1130.27 ± 156.92
TMIN	PRISM	20–year mean annual temperature (°C).	-6.77 ± 1.15	-5.24 ± 1.13
TMAX	PRISM	20–year mean annual maximum temperature (°C).	27.13 ± 1.07	26.19 ± 1.18
TMEAN	PRISM	20–year mean annual minimum temperature (°C).	10.45 ± 0.849	10.93 ± 0.99

Table 4.3. Calibrating species distribution models (Lasso-logistic regression regularized by Lasso, BRT-boosted regression model, MaxEnt-Maximum Entropy) for New River fish species in the temporally independent data. P.t, Pi, and P.o are the observed species prevalence in the training data, independent data, and occupancy-based data respectively. C.b and C.s are respectively the calibrations on bias (o-overestimation, u-underestimation) and spread (s- spread too dispersed, c- too clumpy or concentrated) of predictions of the probability of species occurrence. If a model is well calibrated in terms of both bias and spread, then the cells of C.b and C.s are null, otherwise at least one cell for the species would not be null.

				Single-visit test						Occupancy-based test					
	Prevalence			Lasso		BRT		MaxEnt		Lasso		BRT		MaxEnt	
Species	P.t	P.i	P.o	C.b	C.s	C.b	C.s	C.b	C.s	C.b	C.s	C.b	C.s	C.b	C.s
N. scabriceps	0.21	0.03	0.14	o		u	s	o			c				
P. teretulus	0.15	0.06	0.20			o		o						o	c
E. kanawhae	0.19	0.09	0.25	o		o		o	c	o	c		s		c
E. osburni	0.12	0.01	0.13			o		o							c
P. gymnocephala	0.14	0.08	0.22					o		o	c			o	c
N. platyrhynchus	0.26	0.09	0.33	u	s	u	s	o	s					o	c
C. kanawhae	0.11	0.01	0.27							o	c		c	o	c
R. atratulus	0.14	0.47	0.81	o	c	o	c	u	c	u		u		u	c
C. oreas	0.20	0.11	0.62	o	c					u		u		u	
N. insignis	0.11	0.02	0.32				s			u		o	c	u	
A. rupestris	0.19	0.22	0.66	u	s	u	s			u		u		u	
M. dolomieu	0.17	0.13	0.47	u			s			o	c	u			c
M. salmoides	0.03	0.02	0.11					o	c	o		o			c
S. fontinalis	0.08	0.39	0.22	u	s	u		o		u	s	u		o	c
S. trutta	0.04	0.40	0.34	o		o	c	o	c	u		u			
O. mykiss	0.06	0.17	0.33	o	c	o		o	c	u		u		o	c

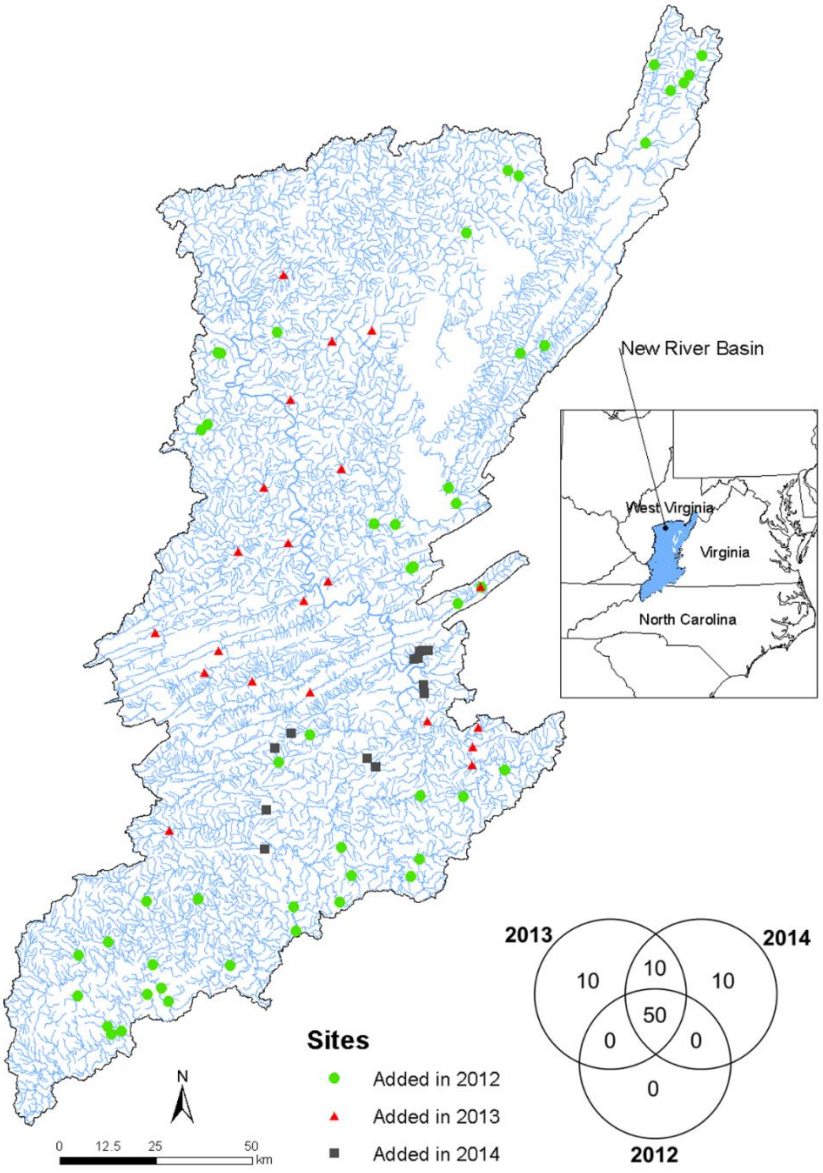


Figure 4.1. The occupancy-based sampling scheme to collect fish species in the New River basin (located in the eastern U.S.) during 2012-2014. The dots in the map represent sites (i.e., inter-confluence stream segments) where we sampled fish. Totally 50 sites were sampled in year 2012, 20 and 10 new sites were added in the following years.

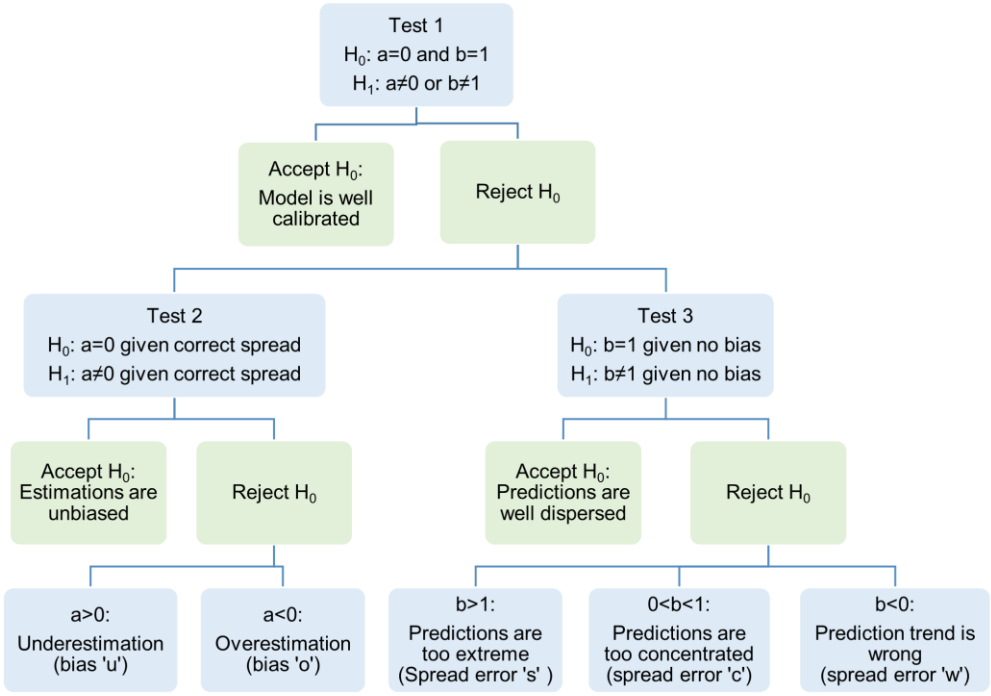


Figure 4.2. The species distribution models for 16 New River fish species were calibrated in terms of bias and spread (Miller *et al.*, 1991; Pearce & Ferrier, 2000). Test 1 likelihood ratio tests to examine whether the intercept (*a*) and slope (*b*) of the calibration line significantly deviate from 0 and 1 respectively. Test 2 and Test 3 test the intercept and slope separately if the null hypothesis of good calibration is rejected in the Test 1. If the null hypothesis is rejected in the Test 2 and Test 3, fitted value of *a* and *b* in the alternative hypothesis would be evaluated to make a decision (‘u’, ‘o’, ‘s‘, ‘c’, and ‘w’).

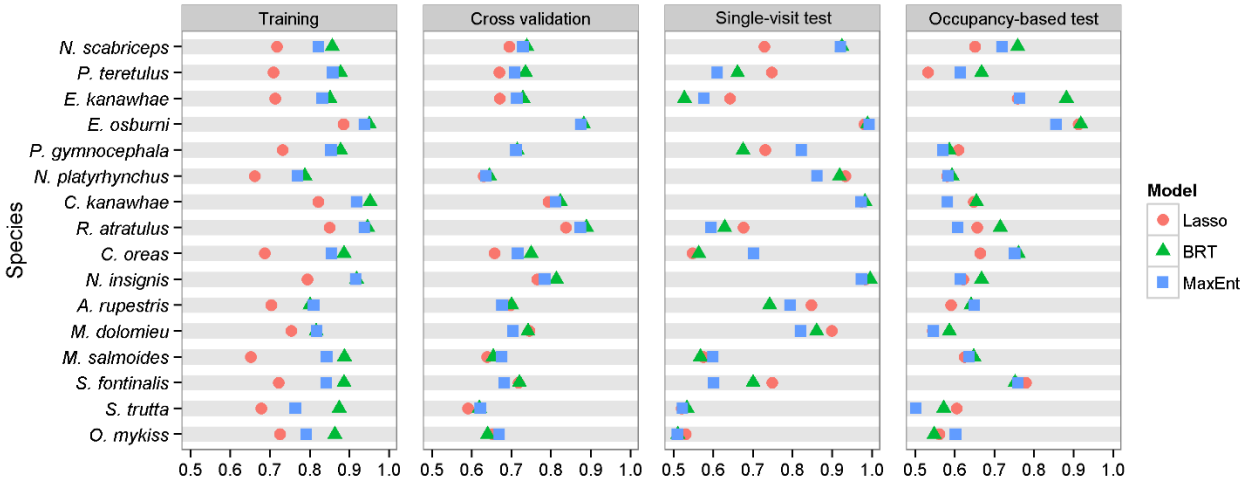


Figure 4.3. AUC (area under the Receiver-Operating-Characteristic curve) in the model training, 5-fold cross validation, (single-visit) independent test, and occupancy-based test. The performance of three types of species distirbution models, Lasso-regularized logistic regression (Lasso), boosted regression trees (BRT) and MaxEnt models, were evaluated.

Figure C.1. Examples of calibration curves drawn by regressing logit of the predicted probability ($\hat{p}_i$) against the observed occurrence (π_i) at evaluation sites in a logistic (calibration) model. Term *a* and *b* are the intercept and slope of the logistic (calibration) model respectively. A species distribution model is well calibrated if $\pi_i = \hat{p}_i$ and the calibration curve has intercept of 0 and slope of 1. Bias (e.g., underestimation indicated by the blue line) occurs if intercept term *a* deviates from 0 (blue line), and spread error (red line) occurs if slope term *b* deviates from 1.

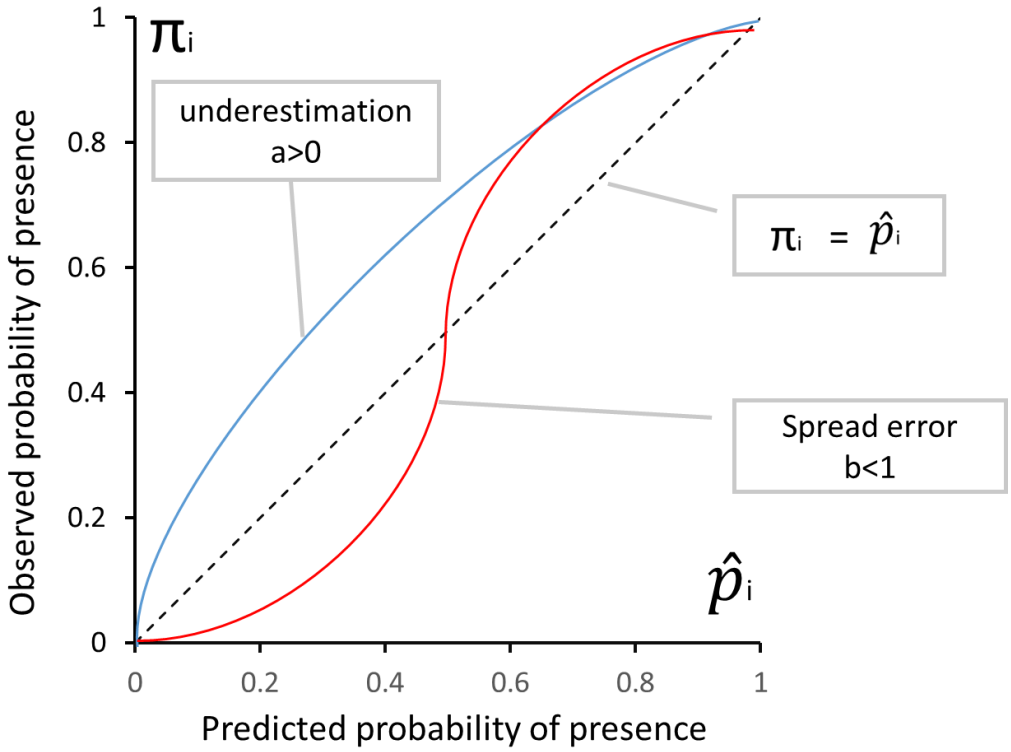


Table C.1. Key predictors and trends of the partial dependence plots in the boosted regression trees. The relationship of fish distribution and habitat factors could be positive (+), negative (-), unimodal (^), and other complicated nonlinear (*). The descriptions of predictors are given in the Table 4.2.

Species	Key predictors				
N. scabriceps	BFI(+)	TMIN(-)	SO(+)	PPT(+)	L_UB(*)
P. teretulus	BFI(+)	TMEAN(-)	SINU(+)	L_UB(-)	TMIN(-)
E. kanawhae	BFI(+)	PPT(+)	ELE(-)	L_FR(-)	SO(+)
E. osburni	BFI(-)	TMEAN(-)	PPT(-)	SINU(-)	TMIN(-)
P. gymnocephala	BFI(+)	HCI(^)	SO(+)	TMEAN(-)	SINU(+)
N. platyrhynchus	BFI(-)	SO(+)	RL(+)	TMEAN(*)	PPT(+)
C. kanawhae	BFI(-)	TMIN(-)	PPT(^)	L_AG(+)	SO(^)
R. atratulus	BFI(-)	TMEAN(-)	SINU(+)	ELE(+)	TMAX(-)
C. oreas	BFI(*)	PPT(-)	SO(-)	HCI(-)	TMEAN(-)
N. insignis	PPT(-)	TMEAN(-)	TMIN(-)	BFI(-)	HCI(-)
A. rupestris	BFI(-)	SO(+)	ELE(-)	L_FR(-)	RF(+)
M. dolomieu	SO(+)	BFI(-)	ELE(-)	L_UB(+)	RL(+)
M. salmoides	SO(+)	TMIN(-)	TMAX(-)	BFI(*)	L_FR(-)
S. fontinalis	SO(-)	SLP(-)	ELE(+)	L_FR(+)	PPT(+)
S. trutta	SO(-)	SLP(-)	TMEAN(^)	ELE(-)	HCI(-)
O. mykiss	TMEAN(-)	SO(^)	PPT(-)	L_FR(+)	HCI(+)

Figure C.2. Discrimination power (overall accuracy, sensitivity, and specificity) of species distribution models in the independent testing (single-visit testing in the Panel A, and occupancy-based testing in the Panel B). Three types of species distirbution models, Lasso-regularized logistic model (Lasso), boosted regression trees (Lasso) and MaxEnt models, were compared.

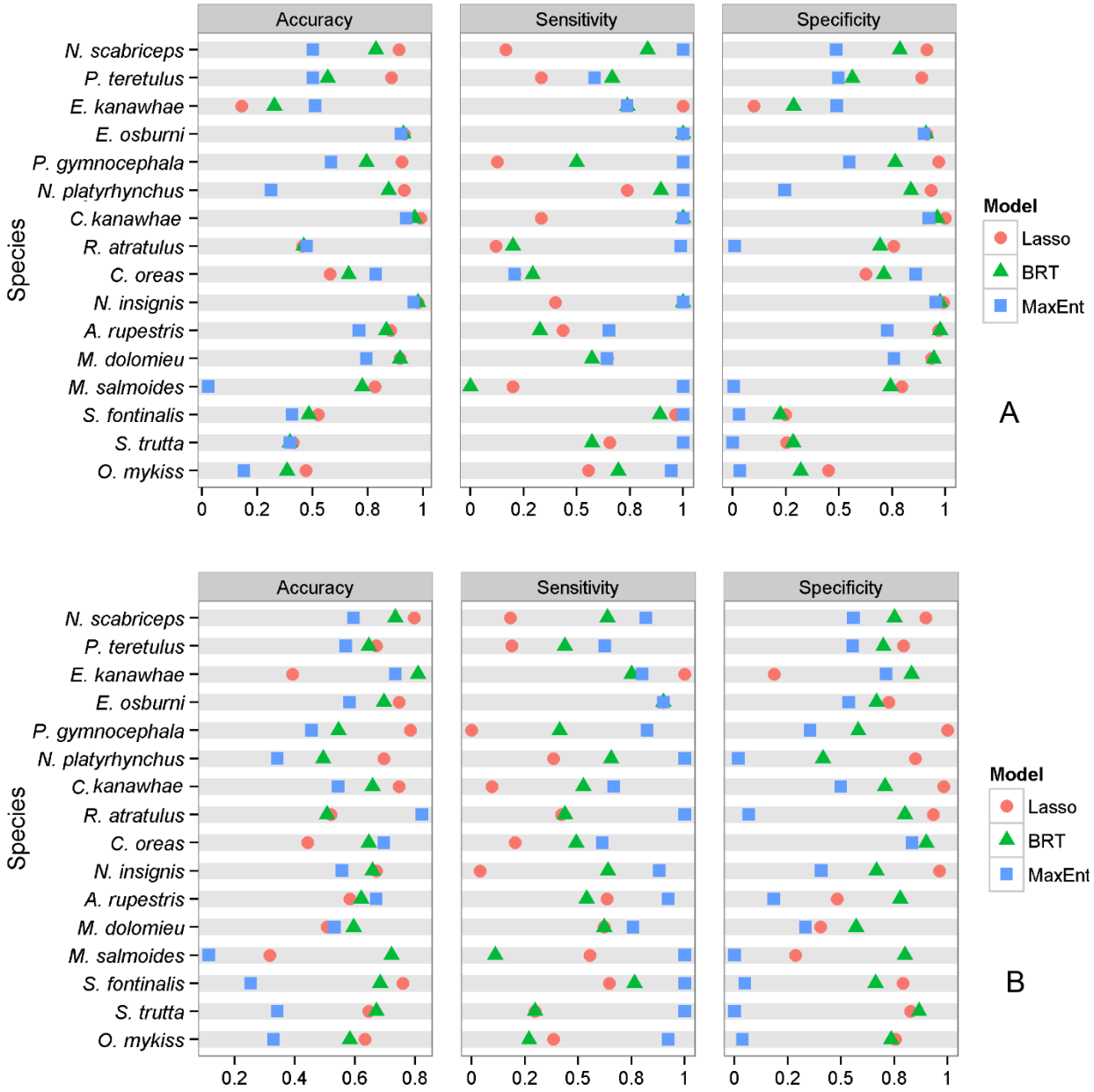


Figure C.3. Partial dependent plots showing that *Etheostoma osburni* (Candy darter) and minimum January temperature are negatively related. We used partial dependent plots to show the marginal effect of minimum temperature on the probability of occurrence of *Etheostoma osburni* (Candy darter) in different time frames. A partial dependence plot shows the dependence of the probability of species occurrence on a predictor variable, marginalizing over the values of all other predictor variables in the model. The plot in the panel A is created based on historical data in 1961-1980, and the plot in the Panel B is based on current data in 1995-2010.

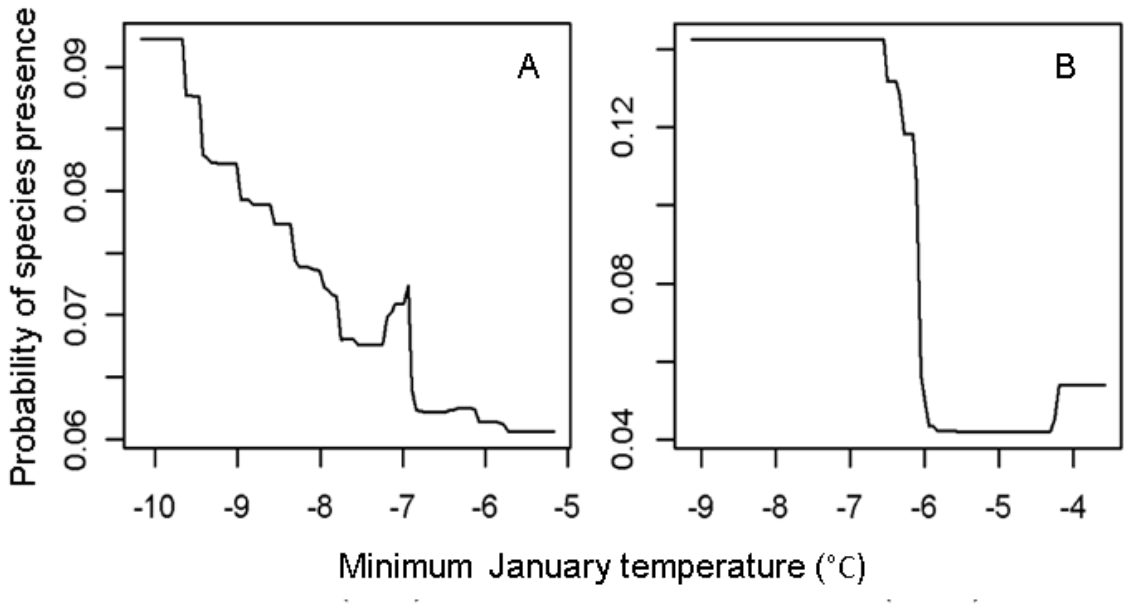
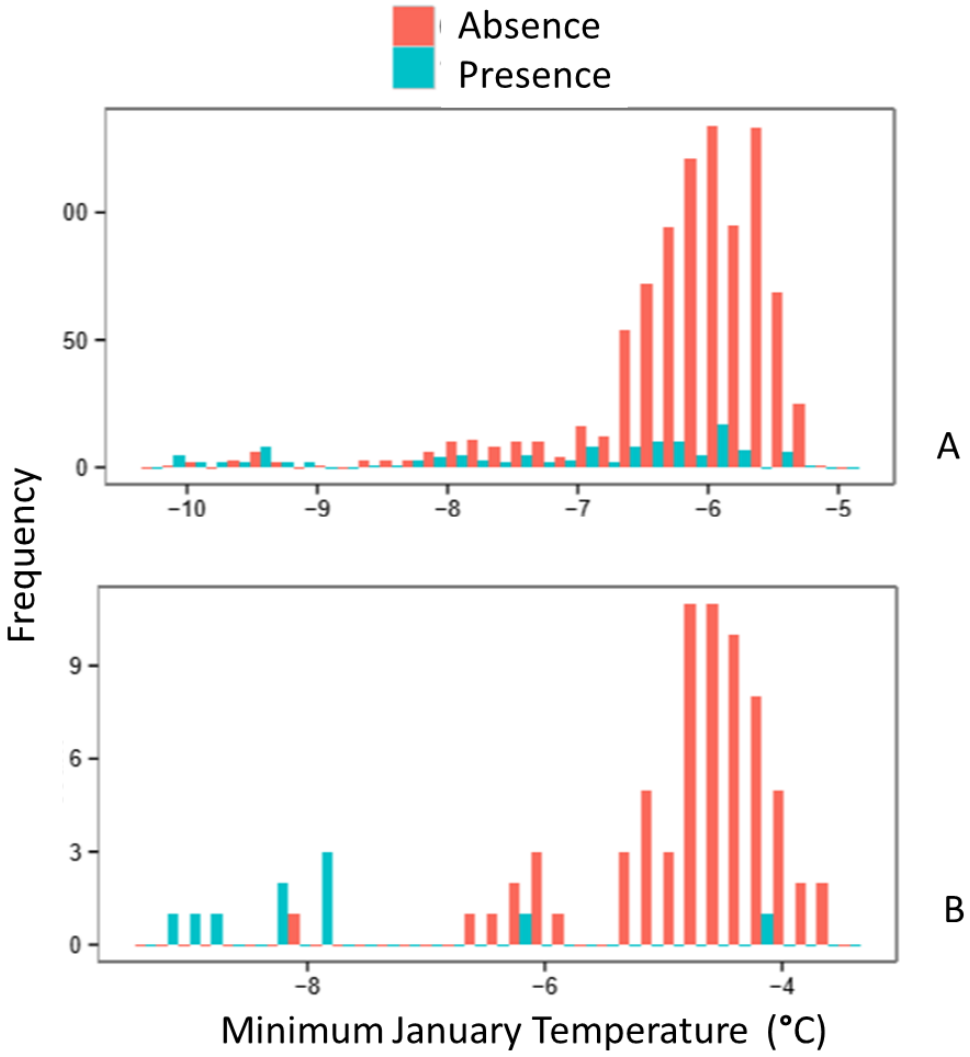


Figure C.4. Histograms showing that *Etheostoma osburni* (Candy darter) and minimum January temperature were negatively correlated. Panel A is based on 1961-1980 data, and panel B is based on 1995-2010 data.



General Conclusion

This study demonstrated the success of using high-resolution metacommunity database developed by integrating historical freshwater fish occurrences for modeling species distributions. The metacommunity database constitutes a valuable resource to provide species presence and absence data to address issues in ecology, biogeography, conservation and resource management at large extent but high resolution. Presence-absence models (boosted regression trees and logistic regression) outperformed presence-only MaxEnt models at the NHD-segment level in model validation. Boosted regression trees are recommended to select variables and describe fish-habitat relationships by partial dependence curves. This study revealed that the performance of a species distribution model also depends on species' traits and incorporation of spatial autocorrelation. Discrimination power in the cross validation was higher for species with large geographic ranges and small local populations. I also found that the effect of prevalence on model fitting could be confounded by a statistical artifact. Incorporating spatial autocorrelation significantly improved discrimination power for a few fish species in the Brazos River Basin. The habitat suitability described could be corrected after detaching the spatial components from the environmental variables, although the discrimination power in the spatial model would not increase conspicuously. The framework of synthesizing accumulated species presence records from different sources into a metacommunity database and then inferring absence of species will serve as a comprehensive tool for understanding species-habitat relationships at multiple spatial scales and help improve conservation and management of taxa.

Spatial transferability of species distribution models was limited for over 70% of stream fish species examined, no matter what modeling approach was used. Predicting spatial distribution of species in new environment is more risky and difficult than interpolations in the

sampled area, owing to spatial heterogeneity in the habitats and ecological relationships. This study reinforced the need for assessment of transferability in studies involving predictions of probability of species occurrence to new environment.

I explored the solutions to improve spatial transferability of species distribution models for stream fishes. Solution 1 is to incorporate dispersal and landscape permeability in the SDMs. Constrained by watershed boundaries or anthropological barriers (e.g., dams, roads), as fish could not occupy all suitable habitat. Accounting for dispersal and landscape connectivity could identify the suitable but unreachable habitat, which in turn could reduce the rate of false positive misclassification rate. Solution 2 is to use model techniques whose complexity can be explicitly controlled. This study showed that conventional logistic model outperformed advanced machine–learning models in the spatial transfers. Over-fitting and multicollinearity in the logistic model can be further mitigated by Lasso regularizations. Solution 3 is to detach the spatial information in the habitat factors, which enables prediction of species occurrence based on reliable ecological relationships solely. Solution 4 is to extend the environmental gradients or sample size of the training data (with spatial balance). Predictions of species occurrence are more stable if the training data ranges encompass prediction data ranges.

Annual average temperature in the New River basin increased by about 0.5 °C from 1970s to 2010s. Over 60% of fish species examined were found negatively correlated with temperature. The overall trend is that these fish species would shift their distribution range northward or upstream but they may respond to climate change differently due to variations in adaptability, thermal tolerance, life history and biogeographic affinities.

The discrimination power was moderate to good in the temporal transfers for species distribution models of 16 New River fish species. The discrimination power was relatively high

for endemic species. Bias and misspecified spread occurred commonly in the temporal transfers of species distribution models, according to calibration based on two independent testing datasets. Predicted probabilities of species presence need to be scaled to ordinal ranks (e.g., low, slight, moderate, high, and extremely high) to avoid overestimation and underestimation. Fine-tuning predictor variables with regularization and cross validation can effectively mitigate the inappropriate spread in the predictions caused by over-fitting or lack-of-fit. Additionally, the solution 1 (incorporating dispersal and landscape permeability), solution 3 (filtering spatial components in the habitat factors) and solution 4 (extending the sample size and environment gradients of the training dataset) proposed in chapter 3 could improve the accuracy and reliability in the temporal transfers of species distribution models as well.

General References

Araújo, M. B. and A. Guisan. 2006. Five (or so) challenges for species distribution modelling. Journal of Biogeography 33:1677-1688.

Barbosa, A. M., R. Real, and J. Mario Vargas. 2009. Transferability of environmental favourability models in geographic space: The case of the Iberian desman (*Galemys pyrenaicus*) in Portugal and Spain. Ecological Modelling 220:747-754.

Bond, N., J. Thomson, P. Reich, and J. Stein. 2011. Using species distribution models to infer potential climate change-induced range shifts of freshwater fish in south-eastern Australia. Marine and Freshwater Research 62:1043-1061.

Breiman, L. 2001. Random Forests. Machine Learning 45:5-32.

Brotons, L., W. Thuiller, M. B. Araújo, and A. H. Hirzel. 2004. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. Ecography 27:437-448.

Bulluck, L., E. Fleishman, C. Betrus, and R. Blair. 2006. Spatial and temporal variations in species occurrence rate affect the accuracy of occurrence models. Global Ecology and Biogeography 15:27-38.

Chen, P., E. O. Wiley, and K. McNyset. 2007. Ecological niche modeling as a predictive tool: silver and bighead carps in North America. Biological Invasions 9:43-51.

Chu, C., N. E. Mandrak, and C. K. Minns. 2005. Potential impacts of climate change on the distributions of several common and rare freshwater fishes in Canada. Diversity and Distributions 11:299-310.

Comte, L., and G. Grenouillet. 2013. Do stream fish track climate change? Assessing distribution shifts in recent decades. Ecography 36:1236-1246.

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

Diez, J. M., and H. R. Pulliam. 2007. Hierarchical analysis of species distributions and abundance across environmental gradients. Ecology 88:3144-3152.

Dormann, C. F., J. M. McPherson, M. B. Araújo, R. Bivand, J. Bolliger, G. Carl, R. G. Davies, A. Hirzel, W. Jetz, W. Daniel Kissling, I. Kühn, R. Ohlemüller, P. R. Peres-Neto, B. Reineking, B. Schröder, F. M. Schurr, and R. Wilson. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30:609-628.

Elith, J., C. H. Graham, R. P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, L. Jin, L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. McC. Overton, A. T. Peterson, and S. J. Phillips. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129-151.

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

Elith, J., S. J. Phillips, T. Hastie, M. Dudík, Y. E. Chee, and C. J. Yates. 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17:43-57.

Fielding, A. H., and P. F. Haworth. 1995. Testing the generality of bird-habitat models. Conservation Biology 9:1466-1481.

Franklin, J., and J. A. Miller. 2009. Mapping species distributions : spatial inference and prediction. Cambridge University Press, Cambridge; New York.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. The Annals of Statistics 29:1189-1232.

Frissell, C., W. Liss, C. Warren, and M. Hurley. 1986. A hierarchical framework for stream habitat classification: Viewing streams in a watershed context. Environmental Management 10:199-214.

Grenouillet, G., L. Buisson, N. Casajus, and S. Lek. 2011. Ensemble modelling of species distribution: the effects of geographical and environmental ranges. Ecography 34:9-17.

Gilpin, M. E., I. Hanski, and L. S. o. London. 1991. Metapopulation dynamics: empirical and theoretical investigations. Academic Press.

Griffith, D. A. and P. R. Peres-Neto. 2006. Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. Ecology 87:2603-2613.

Gumbel, E. J. 1961. Bivariate logistic distributions. Journal of the American Statistical Association 56:335-349.

Hirzel, A. H., P. Bertrand, P.-A. Oggier, C. Yvon, C. Glenz, and R. Arlettaz. 2004. Ecological requirements of reintroduced species and the implications for release policy: the case of the bearded vulture. Journal of Applied Ecology 41:1103-1116.

Hanley, J. A., and B. J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristics curve. Radiology 143: 29-36.

Jenkins, R. E., and N. M. Burkhead. 1994. Freshwater fishes of Virginia. American Fisheries Society.

Lassalle, G., M. Béguer, L. Beaulaton, and E. Rochard. 2008. Diadromous fish conservation plans need to consider global warming issues: An approach using biogeographical models. Biological Conservation 141:1105-1118.

Leathwick, J. R., D. Rowe, J. Richardson, J. Elith, and T. Hastie. 2005. Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. Freshwater Biology 50:2034-2052.

Lee, D. S., S. P. Platania, G. H. Burgess. 1980. Atlas of North American freshwater fishes. North Carolina State Museum of Natural History.

Leibold, M. A., M. Holyoak, N. Mouquet, P. Amarasekare, J. M. Chase, M. F. Hoopes, R. D. Holt, J. B. Shurin, R. Law, D. Tilman, M. Loreau, and A. Gonzalez. 2004. The metacommunity concept: a framework for multi-scale community ecology. Ecology Letters 7:601-613.

Lyons, J., J. S. Stewart, and M. Mitro. 2010. Predicted effects of climate warming on the distribution of 50 stream fishes in Wisconsin, U.S.A. Journal of Fish Biology 77:1867-1898.

MagalhÃEs, M. F., D. C. Batalha, and M. J. Collares-Pereira. 2002. Gradients in stream fish assemblages across a Mediterranean landscape: contributions of environmental factors and spatial structure. Freshwater Biology 47:1015-1031.

Meynard, C. N., and J. F. Quinn. 2007. Predicting species distributions: a critical comparison of the most common statistical models using artificial species. Journal of Biogeography 34:1455-1469.

Miller, M.E., S. L. Hui, and W. M. Tierney. 1991. Validation techniques for logistic regression models. Statistics in Medicine 10: 1213-1226.

Mugodo, J., M. Kennard, P. Liston, S. Nichols, S. Linke, R. Norris, and M. Lintermans. 2006. Local stream habitat variables predicted from catchment scale characteristics are useful for predicting fish distribution. Hydrobiologia 572:59-70.

Niu, S. Q., M. P. Franczyk, and J. H. Knouft. 2012. Regional species richness, hydrological characteristics and the local species richness of assemblages of North American stream fishes. Freshwater Biology 57:2367-2377.

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

Peterson, A. T., M. Pape, and M. Eaton. 2007. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. Ecography 30:550-560.

Randin, C. F., T. Dirnböck, S. Dullinger, N. E. Zimmermann, M. Zappa, and A. Guisan. 2006. Are niche-based species distribution models transferable in space? Journal of Biogeography 33:1689-1703.

Sindt, A. R., C. L. Pierce, and M. C. Quist. 2012. Fish species of greatest conservation need in wadeable Iowa streams: current status and effectiveness of Aquatic Gap Program distribution models. North American Journal of Fisheries Management 32:135-146.

Steen P. J., T. G. Zorn, P. W. Seelbach, and J. S. Schaeffer. 2008. Classification tree models for predicting distributions of michigan stream fish from landscape variables. Transactions of the American Fisheries Society 137: 976-996.

Steen, P. J., M. J. Wiley, and J. S. Schaeffer. 2010. Predicting future changes in Muskegon River watershed game fish distributions under future land cover alteration and climate change scenarios. Transactions of the American Fisheries Society 139:396-412.

Stewart-Koster, B., M. J. Kennard, B. D. Harch, F. Sheldon, A. H. Arthington, and B. J. Pusey. 2007. Partitioning the variation in stream fish assemblages within a spatio-temporal hierarchy. Marine and Freshwater Research 58:675-686.

Stockwell, D., and A. Peterson. 2002. Effects of sample size on accuracy of species distribution models. Ecological Modelling 148:1 - 13.

Strauss, B., and R. Biedermann. 2007. Evaluating temporal and spatial generality: How valid are species–habitat relationship models? Ecological Modelling 204:104-114.

Wang, L., and D. Jackson. 2014. Shaping up model transferability and generality of species distribution modeling for predicting invasions: implications from a study on Bythotrephes longimanus. Biological Invasions 16:2079-2103.

Wenger, S. J., and J. D. Olden. 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. Methods in Ecology and Evolution 3:260-267.

Wiener, N. 1949. Extrapolation, interpolation, and smoothing of stationary time series, with engineering applications. Technology Press of the Massachusetts Institute of Technology, Cambridge.

Wilson, D. S. 1992. Complex interactions in metacommunities, with implications for biodiversity and higher levels of selection. Ecology 73:1984-2000.

Yu, D., M. Chen, Z. Zhou, R. Eric, Q. Tang, and H. Liu. 2013. Global climate change will severely decrease potential distribution of the East Asian coldwater fish *Rhynchocypris oxycephalus* (Actinopterygii, Cyprinidae). Hydrobiologia 700:23-32.

Zarkami, R., R. Sadeghi, and P. Goethals. 2012. Use of fish distribution modelling for river management. Ecological Modelling 230:44-49.