Environ Model Assess (2013) 18:1–12
DOI 10.1007/s10666-012-9325-3

Species Distribution Models of Freshwater Stream Fishes in Maryland and Their Implications for Management

Kelly O. Maloney · Donald E. Weller · Daniel E. Michaelson · Patrick J. Ciccotto

Received: 20 December 2010 / Accepted: 21 May 2012 / Published online: 8 June 2012

Abstract Species distribution models (SDMs) are often used in conservation planning, but their utility can be improved by assessing the relationships between environmental and species response variables. We constructed SDMs for 30 stream fishes of Maryland, USA, using watershed attributes as environmental variables and presence/absence as species responses. SDMs showed substantial agreement between observed and predicted values for 17 species. Most important variables were natural attributes (e.g., ecoregion, watershed area, latitude/longitude); land cover (e.g., %impervious, %row crop) was important for three species. Focused analyses on four representative species (central stoneroller, creek chub, largemouth bass, and white sucker) showed the probability of presence of each species increased non-linearly with watershed area. For these species, SDMs built to predict absent, low, and high densities were similar to presence/absence predictions but provided probable locations of high densities (e.g., probability of high-density creek chub decreased rapidly with watershed area). We applied SDMs to predict suitability of watersheds within the study area for each species. Maps of suitability and the environmental and species response relationships can help develop better management plans.

Keywords Conditional random forests · Landscape · Prediction · Classification · Habitat suitability

Electronic supplementary material The online version of this article (doi:10.1007/s10666-012-9325-3) contains supplementary material, which is available to authorized users.

K. O. Maloney · D. E. Weller · D. E. Michaelson
Smithsonian Environmental Research Center,
647 Contees Wharf Rd, P.O. Box 28, Edgewater, MD 21037, USA

D. E. Weller
e-mail: wellerd@si.edu

D. E. Michaelson
e-mail: dem2k@virginia.edu

P. J. Ciccotto
Monitoring and Non-tidal Assessment,
Maryland Department of Natural Resources,
580 Taylor Avenue, C-2,
Annapolis, MD 21401, USA
e-mail: PCiccotto@dnr.state.md.us

K. O. Maloney (✉)
Northern Appalachian Research Laboratory, U.S. Geological Survey-Leetown Science Center,
176 Straight Run Road,
Wellsboro, PA 16901, USA
e-mail: kmaloney@usgs.gov

Present Address:
D. E. Michaelson
Departments of Environmental Science and Engineering Science,
University of Virginia,
Charlottesville, VA 22902, USA

1 Introduction

Conserving and protecting populations and ecosystems has been a main focus of ecological research for the past 30 years. This interest has increased dramatically partly in response to the reported loss of global biodiversity [1, 2] and the homogenization of major taxonomic groups across the landscape [3, 4]. These problems are largely driven by altered or degraded ecosystem conditions caused by land use practices [5–8]. Conservation efforts would be greatly aided by large-scale predictions of species distributions (i.e., species distribution models [9–11]) that account for such anthropogenic stressors.

Conserving fish assemblages of small streams is particularly important because many streams are impaired [12, 13] and fish assemblage integrity decreases with increasing anthropogenic stress [14–16]. However, a generalized conservation plan is likely to be ineffective because the species

in a fish assemblage have differing life history strategies, habitat requirements, and sensitivities to stressors [17–19]. Having a distribution model for each species of interest should yield more effective conservation plans.

There are many techniques to construct species distribution models (SDMs, [9, 20]), including: generalized linear models, generalized additive models, Generic Algorithm for Rule-Set Predictions, Classification and Regression Trees, and Maximum Entropy. Many of these methods have been used to predict species distributions of fishes (e.g., [21–23]), usually by predicting suitable habitats. Recent reviews suggest that SDMs can be improved by integrating species–environment relationships and incorporating more ecological theory into SDM construction (e.g., [9, 24]). These improvements could yield better models for stream fish conservation (for example, see [25]).

Here, we present SDMs for 30 common species of fish from streams in the Chesapeake Bay watershed area of Maryland, USA. Our overall goal was to predict suitable habitat from watershed attributes and present these data as habitat suitability maps for all 30 species. For each species, we used the conditional random forests machine-learning algorithm (cRF, [26, 27]) to build an SDM that predicts the probability of presence from environmental variables. The explanatory variables included watershed indicators of anthropogenic stress (e.g., % impervious cover or % row crop cover) and natural watershed attributes (e.g., % sand in soil, ecoregion, latitude, or longitude). We evaluated relationships between species presence and environmental variables using the variable importance and partial dependence plots from each cRF model (see [26, 28]). For four example species, we also built cRF SDMs that predicted the probabilities of three density categories (absent, low density, high density) rather than presence/absence.

2 Materials and Methods

2.1 Study Sites

We studied the 23,408 km^2 portion of Maryland within the Chesapeake Bay basin in the Mid-Atlantic region of the US (Online resource 1). The study area intersects six level III ecoregions including: Central Appalachians, Ridge and Valley, Blue Ridge, Northern Piedmont, Southeastern Plains, and Middle Atlantic Coastal Plains [29]. The climate ranges from cold with hot summers in the mountainous western area to temperate with hot summers towards the southeast [30]. The vegetation ranges from Northern hardwood forests in the highlands to oak, hickory, pine, and southern mixed forests of the Coastal Plains [29]. The Appalachian, Ridge and Valley, and Blue Ridge ecoregions are underlain mainly by folded and faulted sedimentary rocks; the Piedmont ecoregion is underlain by crystalline igneous and metamorphic rocks; and the Plains ecoregions are underlain by unconsolidated sediments [31]. There are cold, higher-gradient streams in the Central Appalachians and Ridge and Valley and lower-gradient, naturally acidic streams in the Coastal Plains ([32]).

2.2 Data Sets

Data on fish presence/absence and density came from the Maryland Biological Stream Survey (MBSS, [33]. The MBSS is an ongoing survey of first- to fourth-order streams (as shown on 1:100,000 US Geological Survey [USGS] maps). MBSS uses a probabilistic sampling design stratified by major watershed [34]. We used fish data collected from 1994–2004 at sites with watershed areas <200 km^2 that were in the Chesapeake Bay watershed and had a least one fish species collected. We used only the first record for sites that were sampled more than once. Of the 2,181 samples, 1,460 satisfied these conditions. There were 81 total fish species in the data set, and we selected 30 of the more widely distributed (collected in >250 sites) species for our analyses (Online resource 2). The 30 species come from 11 families and have a range of ecological characteristics and sensitivities to stressors (Online resource 2). For each species, we converted raw abundance scores to presence/absence.

We used watershed variables that influence stream conditions [5, 26, 35] to build models predicting the presence/absence of each species at each sampling point. We summarized each watershed attribute by using the ArcGIS geographic information system (ESRI, Redlands, CA, USA) to intersect watershed boundaries [36] with land cover, human population, and elevation data. We calculated the percentages of each watershed covered by impervious surface, tree cover, row-crop agriculture, pasture, and extractive cover (e.g., mines) using the 2001 National Land Cover Data [37]. Land cover did not change much between 1992 and 2001 in most of the watersheds (see Online resource 3), so we did not attempt to model the effects of land cover change in this analysis.

We used the Zonal Statistics (++) function in Hawth's Analysis Tools for ArcGIS [38] to calculate average watershed slope and elevation from a digital elevation model (DEM, 1:250,000 scale, 30 m, http://edc2.usgs.gov/geodata/index.php) and average annual precipitation for each watershed from a publicly available data set [39]. We calculated the drainage density (in kilometers per square kilometer) for each watershed by dividing the total stream length by watershed area. The average percentage of sand in soils within a watershed was calculated using STASGO soil data [40]. The percentage of calcareous bedrock came from a geological map [41] and descriptions classifying rock types as calcareous or non-calcareous [42]. Slope and elevation were highly correlated with longitude (both r>0.80), so they were removed from model development.

2.3 Presence/Absence Models

We used a random subsample (n=1162) of the MBSS sites as a training data set for development of the models, and we used the remaining sites (n=298) as an independent data set for model evaluation. We used the same training and evaluation sites for every species model. All statistical analyses were conducted using the R statistical software [43]. We developed models to predict fish presence or absence for each of the 30 species; but for clarity and conciseness, we focus on results for a set of four example species (central stoneroller, creek chub, largemouth bass, and the white sucker), which represented many of the ecological characteristics and tolerances of all 30 species (Online resource 2).

We used an ensemble tree approach to model the relationship between the presence/absence of each species and the watershed predictors. Ensemble trees are based on recursive decision trees algorithms. In a decision tree analysis, a covariate is selected from all exploratory variables and a split point value of that covariate is estimated that separates the response values into two groups. Each group is further separated into subgroups by new explanatory variables and split points. The recursive splitting procedure is stopped when a predefined stopping criterion is reached. The accuracy and the predictive ability of single tree models can be improved by using ensembles of many trees (forests; [28]. In a typical ensemble tree approach, bootstrap samples are drawn with replacement from the original data set, and observations not included in a bootstrap sample are named *out-of-bag* observations. For each bootstrap sample, a very large tree is generated and used to classify the out-of-bag observations. The final predicted class of each observation is the class most often assigned to it across all the bootstrap samples. Inferences about the strengths of relationships between predictor and response variables can be drawn from variable importance plots, which are derived for each predictor variable by randomly permuting the values of the variable for the out-of-bag observations. For more detailed descriptions of decision trees and ensemble tree methods, see References [28, 44–46].

We used the cRF [27] implementation of the ensemble approach to analyze relationships between fish presence/absence and watershed attribute predictors. cRF models are based on conditional inference trees, a method that uses classical statistical tests [47] to select split points. Splits are based on the minimum p value among all tests of independence between the response variable and each explanatory variable. We evaluated the strengths of species–environment relationships using variable importance and partial dependence plots from the cRF models [28].

Model performance for each species was evaluated with five accuracy measures calculated from the independent evaluation data set (R package PresenceAbsence [48]). The five measures were the Proportion Correctly Classified (PCC), the Kappa statistic, sensitivity, specificity, and the area under the receiver operating curve (AUC). Both PCC and Kappa are calculated from model confusion matrices, which are tables contrasting predicted vs. observed classifications. Kappa adjusts PCC for agreement due to chance alone, and Kappa ranges from −1 to 1 with increasing positive values indicating stronger agreement between the predicted and observed values [49]. We used the Kappa ranges reported by Landis and Koch [49] to infer strength of agreement—Kappa<0.00 poor, 0.00–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.81–1.00 almost perfect agreement. Sensitivity is the proportion of observed presences correctly predicted, while specificity is the proportion of observed absences correctly predicted. AUC evaluates the sensitivity and specificity of the model; values range from 0 to 1 with values >0.5 indicating model performance better than chance alone [50].

We applied the calibrated and tested model for each fish species, to predict the probability of presence/absence in the 10,806 small (upslope drainage area <200 km^2) stream reaches in the Maryland portion of the Chesapeake Bay watershed. The stream reaches and associated watersheds were taken from the 1:100 K National Hydrology Dataset plus (NHDplus; [51]) after eliminating tidal reaches adjacent to the Chesapeake Bay. We calculated watershed attributes using the methods above for the entire watershed draining to the downstream end of a reach. The predictions were summarized with watershed maps shaded by the predicted probability of presence.

2.4 Density Category Models

We also applied the cRF models to predict fish density for the set of four example species. We categorized densities into 3 bins: no individuals collected (0, Absent), Low Density, and High Density. The low density category was<0.05 individuals per square meter for the central stoneroller (n=135) and white sucker (n=337), <0.10 for creek chub (n=390), and<0.01 for largemouth bass (n=172). Sites above these thresholds were high-density (n=103, 266, 247, and 172 for central stoneroller, white sucker, creek chub, and largemouth bass, respectively). Model accuracy was evaluated using the same accuracy statistics for the presence/absence approach; however, we used a weighted kappa statistic [52], vcd function [53], and calculated AUC using the ordROC function in the nonbinROC R package [54] to address the ordinal structure of the binned densities. Sensitivity and specificity for binned analyses were calculated comparing each density category to remaining categories [55].

3 Results

3.1 Presence/Absence Models

All models correctly classified over 75%s of the sites for each species (mean PCC=0.88, SE=0.01) and performed better than chance alone (all AUCs>0.50, Table 1). The model for bluegill performed worst (76% correctly classified) and the model for the eastern mudminnow performed best (97% correctly classified). However, the kappa statistic averaged across the models for all 30 species suggested a moderate strength of agreement between observed and predicted values (mean kappa=0.59, SE=0.03, Table 1). Values of kappa ranged from 0.13 (slight agreement, Blue Ridge sculpin) to 0.93 (almost perfect agreement, eastern mudminnow). The model sensitivity (true presences) and specificity (true absences) values indicated that a lower kappa statistic in Table 1 was often due to weak ability to predict species presence. For example, the model for Blue Ridge sculpin had the lowest kappa statistic (0.13) and the lowest sensitivity (0.08) but the highest specificity (1.00). This model poorly predicted presences but predicted absences perfectly. The model for eastern blacknose dace had the highest sensitivity (0.97) but the lowest specificity (0.73); it predicted presence better than absence. Kappa for this model was 0.74, indicating a substantial strength of agreement with the observations.

Comparing the accuracy measures with the confusion matrix provides additional information on model performance. We present confusion matrices for four example species: central stoneroller, creek chub, largemouth bass, and white sucker (Online resource 4). The model for the central stoneroller predicted presence moderately well (error rate=0.32) and absence extremely well (error rate=0.04), resulting in a

Table 1 Accuracy measures for models predicting presence/absence of fish species

Common name	PCC	Kappa	Sensitivity	Specificity	AUC
American eel	0.84	0.67	0.88	0.81	0.94
Pirate perch	0.96	0.82	0.88	0.97	0.99
White sucker	0.83	0.66	0.85	0.82	0.93
Creek chubsucker	0.87	0.59	0.62	0.94	0.93
Northern hog sucker	0.90	0.59	0.57	0.96	0.95
Redbreast sunfish	0.82	0.47	0.56	0.90	0.86
Green sunfish	0.86	0.35	0.26	0.99	0.82
Pumpkinseed	0.83	0.55	0.62	0.91	0.84
Bluegill	0.76	0.51	0.68	0.82	0.83
Largemouth bass	0.80	0.37	0.34	0.96	0.84
Blue ridge sculpin	0.89	0.13	0.08	1.00	0.84
Potomac sculpin	0.93	0.67	0.57	0.99	0.96
Central stoneroller	0.92	0.68	0.68	0.96	0.96
Rosyside dace	0.85	0.66	0.73	0.91	0.92
Cutlip minnow	0.89	0.63	0.55	0.99	0.96
Common shiner	0.90	0.64	0.61	0.97	0.96
Golden shiner	0.89	0.41	0.32	0.99	0.90
Swallowtail shiner	0.90	0.51	0.45	0.97	0.93
Bluntnose minnow	0.94	0.76	0.77	0.97	0.96
Eastern blacknose dace	0.90	0.74	0.97	0.73	0.95
Longnose dace	0.89	0.77	0.86	0.91	0.96
Creek chub	0.85	0.69	0.95	0.75	0.93
Fallfish	0.88	0.47	0.38	0.98	0.90
Redfin pickerel	0.91	0.63	0.61	0.96	0.96
Yellow bullhead	0.89	0.28	0.21	0.98	0.87
Margined madtom	0.91	0.65	0.62	0.97	0.93
Fantail darter	0.92	0.69	0.64	0.98	0.98
Tessellated darter	0.86	0.71	0.91	0.82	0.95
Least brook lamprey	0.87	0.40	0.42	0.94	0.91
Eastern mudminnow	0.97	0.93	0.96	0.97	0.99
Mean (standard error)	0.88 (0.01)	0.59 (0.03)	0.62 (0.04)	0.93 (0.01)	0.92 (0.01)

These measures were calculated for the independent data set not used in model calibration

PCC proportion correctly classified, *AUC* area under the receiver operating characteristic curve

sensitivity of 0.68 and a specificity of 0.96. The kappa statistic was 0.68 (substantial agreement). The model for the creek chub predicted presence (error rate=0.05) better than absence (error rate=0.25) resulting in a higher value of sensitivity (0.95) than specificity (0.75) and a kappa statistic of 0.69 (substantial agreement). The confusion matrix for largemouth bass showed that the model predicted presence poorly (error rate=0.66) but absence very well (error rate=0.04) resulting in a very low sensitivity (0.34) and high specificity (0.96) and a kappa of 0.37 (fair agreement). The model for the white sucker predicted presence and absence equally well (error rates=0.15, 0.18, respectively) resulting in similar sensitivity (0.85) and specificity (0.82) and a kappa of 0.66 (substantial agreement, Online resource 4).

Models for all of the species were influenced most by natural landscape features (Table 2). Ecoregion was among the four most important variables for models of 27 of the 30 species; drainage density and watershed area for 19 species; and latitude, longitude, and percent sand in soils in for 13, 15, and 14 of the species, respectively. Anthropogenic land cover variables were among the four most important variables for only three species: percent impervious for the redbreast sunfish and swallowtail shiner and percent pasture for the rosyside dace. For the central stoneroller, the top seven variables were natural attributes of watersheds while five of the land cover measures were the least important variables (Table 2, Online resource 5a). For the creek chub, the five most important variables were natural attributes (Ecoreg, PerSand, Lat, Long, Precip), followed by three land use variables (PerCrop, PerPast, PerImp; Table 2, Online resource 5b). Similarly, the top four variables for the largemouth bass model were natural attributes (WSArea, DrnDns, Ecoreg, Lat), followed by PerCrop, PerTree, and PerImp (Table 2, Online resource 5c). For the white sucker model, the six most important variables were natural attributes, which were followed by five of the six land cover variables (Table 2, Online resource 5d). The relationships between explanatory and response variables from cRF models can be visualized using partial dependence plots. As an example, we show these plots for watershed area for each of the four example species—the probability of presence increased sharply with watershed area, but decreased again above ~60 km^2 for the central stoneroller and creek chub (Fig. 1).

We predicted the probability of presence for each of the 30 species in every small, nontidal stream reach in the Chesapeake Bay portion of Maryland, and we present maps of the results for the example species (Fig. 2). Maps for all 30 species are included in Online resource 6. The importance of Ecoregion is evident in the maps for central stoneroller, creek chub, and white sucker, which all have low probabilities of presence in the Southeastern Plains and Middle Atlantic Coastal Plains ecoregions (Fig. 2). The central stoneroller and largemouth bass had fewer areas of a high probability of presence than the creek chub (Fig. 2b) and white sucker (Fig. 2d).

3.2 Density Category Models

The density category models correctly classified between 63% (creek chub) and 85% (central stoneroller) of the observations (Table 3). There was substantial agreement between model predictions and observed data for the central stoneroller, creek chub, and white sucker (all Kappas between 0.61 and 0.80) and slight agreement for the largemouth bass model (Kappa=0.18). The model for the central stoneroller showed strong ability to classify true absences (sensitivity for absent class=1.00) but low specificity for the absence category (0.51). Low sensitivity statistics for both the low (0.24) and high (0.47) density bins indicate a weak ability to predict these density categories (Table 3). The creek chub model accurately predicted absence (sensitivity=0.81) and discriminated absent from the two density categories (specificity=0.89). This model predicted the low-density category better than the high-density category (Table 3). The cRF model for the largemouth bass predicted absence well (sensitivity=1.00) but both density categories weakly (absent specificity=0.16, indicating a high misclassification of both density categories as absent). The white sucker model accurately predicted absence from the two density categories (absent sensitivity=0.89, specificity=0.81); and predicted the low- and high-density categories similarly (low-density sensitivity=0.54, high-density sensitivity=0.56, Table 3).

The density predictions for the four fish species were again most influenced by natural watershed attributes (Fig. 3). Watershed area, ecoregion, and drainage density were most important for the central stoneroller (Fig 3a); ecoregion, percent sand in soils, and latitude were most important for the creek chub (Fig. 3b); and ecoregion, watershed area, and drainage density were most important for the white sucker (Fig. 3d). Important variables for the largemouth bass model included percent row crop cover as the third most important variable after watershed area and drainage density (Fig. 3c). For creek chub, the probabilities of absence and low density increased rapidly with watershed area, and the probability of high creek chub density decreased with watershed area up to ~40 km^2 (Fig. 4).

We applied the cRF model to predict the probabilities of the absent, low-density, and high-density categories for every small, nontidal stream reach in the Chesapeake Bay portion of Maryland. We present results for high-density populations (Fig. 5). Only a relatively few watersheds in the N. Piedmont were predicted with moderate probability (e.g., 0.20–0.60) to have high densities of central stoneroller (Fig. 5a). Creek chub were predicted to be in high densities in N. Piedmont and western mountainous regions (Fig. 5b); white sucker were

Table 2 Variable importance measures of all independent variables for each fish species model

Common name	WSarea	DrnDns	Ecoreg	Lat	Long	PerCalc	PerSand	Precip	PerImp	PerTree	PerCrop	PerPast	PerWet	PerBar
American eel	0.015	0.008	**0.102**	**0.057**	**0.117**	0.001	0.018	**0.028**	0.002	0.007	0.011	0.005	0.006	0.004
Pirate perch	0.011	0.005	**0.051**	**0.029**	**0.024**	0.000	**0.022**	0.003	0.003	0.002	0.009	0.001	0.004	0.001
White sucker	**0.069**	**0.045**	**0.080**	0.033	0.017	0.001	**0.037**	0.005	0.015	0.015	0.010	0.015	0.008	0.003
Creek chubsucker	0.010	**0.017**	**0.040**	**0.028**	0.011	0.001	**0.022**	0.003	0.004	0.002	0.013	0.003	0.003	0.002
Northern hog sucker	**0.063**	**0.028**	**0.039**	0.012	0.010	0.002	**0.012**	0.007	0.002	0.002	0.004	0.010	0.002	0.001
Redbreast sunfish	**0.080**	**0.033**	0.009	0.004	0.012	0.002	0.005	0.003	**0.013**	**0.021**	0.006	0.006	0.009	0.006
Green sunfish	**0.019**	0.012	**0.025**	**0.019**	0.011	0.004	0.011	**0.013**	0.006	0.007	0.007	0.004	0.003	0.011
Pumpkinseed	**0.039**	**0.024**	**0.052**	0.014	0.018	0.001	**0.026**	0.004	0.001	0.002	0.006	0.003	0.018	0.003
Bluegill	**0.023**	**0.041**	**0.018**	0.005	0.015	0.001	**0.019**	0.004	0.001	0.006	0.012	0.004	0.009	0.003
Largemouth bass	**0.029**	**0.018**	**0.013**	**0.008**	0.005	0.003	0.003	0.001	0.005	0.006	0.008	0.004	0.004	0.001
Blue ridge sculpin	0.002	0.001	**0.015**	0.004	**0.008**	0.002	0.005	**0.008**	0.003	**0.016**	0.003	0.006	0.002	0.001
Potomac sculpin	**0.022**	**0.027**	**0.031**	0.009	**0.057**	0.004	0.008	0.017	0.003	0.005	0.004	0.003	0.003	0.005
Central stoneroller	**0.032**	**0.024**	**0.037**	**0.022**	0.019	0.007	0.016	0.021	0.004	0.004	0.007	0.013	0.003	0.003
Rosyside dace	0.010	0.005	**0.113**	0.016	**0.043**	0.002	**0.036**	0.027	0.013	0.006	0.016	**0.033**	0.007	0.005
Cutlip minnow	**0.054**	**0.024**	**0.063**	0.009	**0.027**	0.001	0.014	0.023	0.001	0.003	0.003	0.014	0.001	0.002
Common shiner	**0.072**	**0.021**	**0.042**	**0.023**	0.013	0.001	0.011	0.014	0.003	0.004	0.003	0.019	0.002	0.002
Golden shiner	0.004	**0.007**	**0.019**	**0.007**	**0.006**	0.000	0.006	0.003	0.000	0.002	0.002	0.001	0.003	0.001
Swallowtail shiner	**0.044**	**0.013**	0.008	0.005	**0.021**	0.000	0.006	0.012	**0.022**	0.008	0.006	0.005	0.004	0.002
Bluntnose minnow	**0.036**	**0.026**	**0.045**	0.021	0.018	0.002	0.015	**0.035**	0.003	0.010	0.004	0.013	0.002	0.004
Eastern blacknose dace	0.005	0.002	**0.117**	**0.037**	**0.034**	0.000	**0.046**	0.004	0.008	0.006	0.009	0.007	0.010	0.003
Longnose dace	**0.063**	**0.038**	**0.141**	0.032	0.017	0.001	**0.049**	0.008	0.003	0.005	0.005	0.015	0.010	0.004
Creek chub	0.001	0.000	**0.132**	**0.052**	**0.040**	0.002	**0.073**	0.017	0.009	0.007	0.015	0.012	0.007	0.004
Fallfish	**0.041**	**0.034**	**0.029**	**0.029**	0.010	0.005	0.008	0.007	0.008	0.004	0.009	0.013	0.007	0.005
Redfin pickerel	0.007	0.005	**0.054**	0.015	**0.022**	0.000	**0.029**	0.003	0.003	0.003	0.015	0.003	**0.020**	0.001
Yellow bullhead	**0.016**	0.008	**0.009**	0.006	0.007	0.001	**0.009**	**0.015**	0.001	0.002	0.002	0.004	0.001	0.005
Margined madtom	**0.079**	**0.027**	0.011	0.007	**0.030**	0.001	0.004	**0.016**	0.005	0.004	0.007	0.010	0.002	0.002
Fantail darter	0.017	**0.020**	**0.047**	0.016	**0.092**	0.002	0.008	**0.024**	0.004	0.006	0.006	0.006	0.003	0.006
Tessellated darter	**0.079**	**0.035**	**0.037**	0.015	**0.076**	0.002	0.005	0.028	0.006	0.021	0.014	0.013	0.005	0.005
Least brook lamprey	**0.013**	0.010	**0.044**	**0.019**	0.007	0.000	**0.016**	0.002	0.009	0.002	0.007	0.004	0.005	0.001
Eastern mudminnow	0.000	0.000	**0.180**	**0.066**	**0.025**	0.001	**0.060**	0.004	0.009	0.007	0.009	0.004	0.004	0.002

Within a row, the values in bold type indicate the four most important variables in the model for that species

The variable names are *WSArea*=watershed area, *DrnDens*=drainage density, *Ecoreg*=ecoregion, *lat*=latitude, *long*=longitude, *PerCalc*=% calcareous bedrock, *PerSand*=% sand content in soils, *Precip*=precipitation, *PerImp*=% impervious cover, *PerTree*=% tree cover, *PerCrop*=% row crop cover, *PerPast*=% pasture cover, PerWet=% wetland cover, and PerBar=% barren cover in a watershed

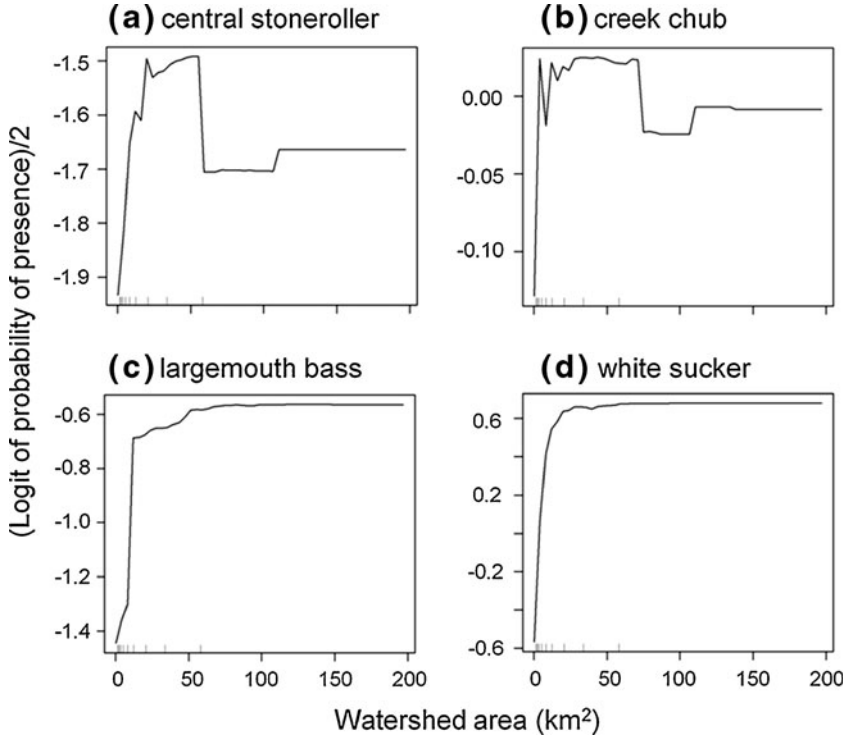

Fig. 1 Example partial dependence plots for watershed area from the presence/absence models for four example species

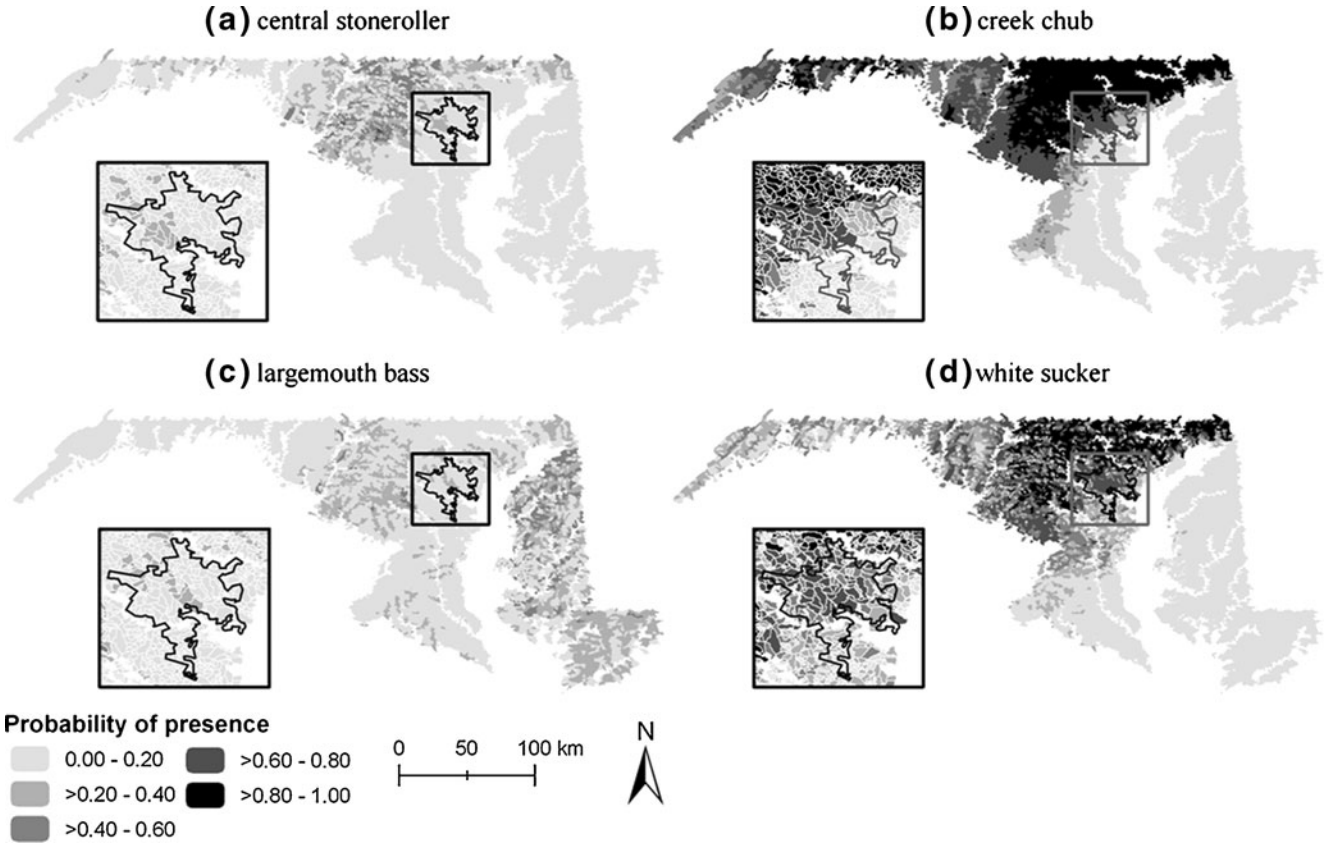

Fig. 2 Habitat suitability based on presence/absence predictions for four example species in all small, nontidal reaches in the study area (Online resource 1). The *insets* show enlarged views of the results near Baltimore, Maryland

Table 3 Accuracy measures for density category models of four example species

Common name	PCC	Kappa	Abundance class	Sensitivity	Specificity	AUC
Central stoneroller	0.85	0.65	0	1.00	0.51	0.93
			0-0.05	0.24	0.98	
			>0.05	0.47	0.99	
Creek chub						
	0.63	0.65	0	0.81	0.89	0.79
			0-0.1	0.76	0.78	
			>0.1	0.48	0.93	
Largemouth bass						
	0.77	0.18	0	1.00	0.16	0.83
			0-0.01	0.14	0.99	
			>0.01	0.05	1.00	
White sucker	0.60	0.67	0	0.89	0.81	0.79
			0-0.05	0.54	0.86	
			>0.05	0.56	0.90	

These measures were calculated for the independent data set not used in model calibration

PCC=proportion correctly classified, *AUC*=area under the receiver operating characteristic curve

predicted to be in high density in watersheds in the N. Piedmont (Fig. 5d). The largemouth bass was weakly predicted (0.20–0.40) to be in high densities in the upper Eastern Shores region of Maryland (Fig. 5c). Effects of impervious surface were evident in the creek chub predictions because few high densities were predicted within Baltimore city (inset Fig. 5b).

4 Discussion

We constructed species distribution models (SDMs) for 30 species of freshwater fishes that inhabit streams of Maryland, USA. The models showed substantial agreement between observed and predicted values for 17 of the 30 species. The most important variables for all species were natural watershed attributes, such as ecoregion, watershed area, latitude, and longitude. Land cover variables (e.g., percent impervious, percent row crop, or percent pasture) were highly important variables for only three species (redbreast sunfish, swallowtail shiner, and rosyside dace). This does not mean that land cover is unimportant for the other species. Instead, the random forest analyses identified a hierarchy of factors governing stream fishes. Large-scale natural biogeographic patterns (e.g., ecoregion) and system size were the most important factors

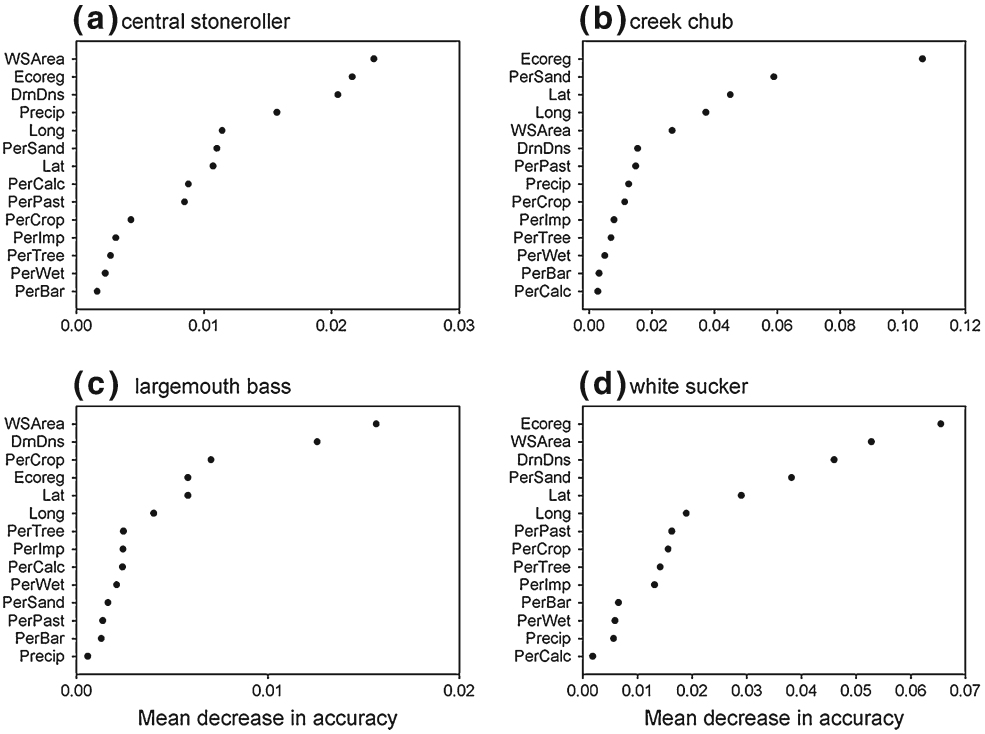

Fig. 3 Variable importance plots from the density category models for four example species; **a** central stoneroller, **b** creek chub, **c** largemouth bass, **d** white sucker. Variables with higher values of mean decrease in accuracy are more important. The variable abbreviations are in Table 2

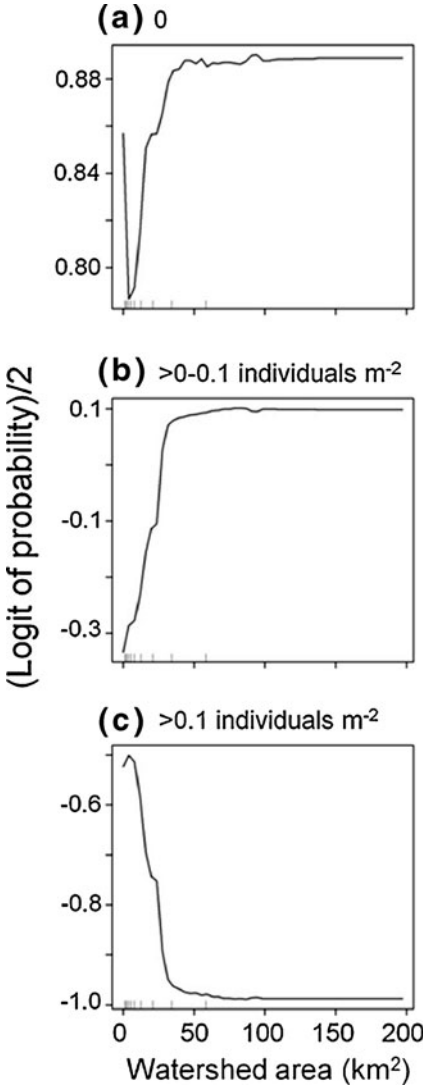

Fig. 4 Example partial dependence plots for watershed area from the density category model for creek chub

structuring fish distributions across regions. Land use factors played a role in structuring fish distributions within regions only after partitioning out the effects of position and system size, indicating a need to incorporate spatial and size components into regional fish management plans.

4.1 Importance of Watershed Attributes

The minor role of anthropogenic land cover variables in our models reflects the strong effects of natural watershed attributes on the fish assemblage across our study area. Stream size (represented here by watershed area) strongly influences stream fish assemblages [56–60], and watershed area was important in all our models. Ecoregion, latitude, and longitude all are related and indicate a strong location effect on the fish species. Location has been reported to strongly affect fish assemblages [60–62], probably because environmental similarity and dispersal decline with distance between sites [63]. Sand content in soils was important in many of the SDMs and is related to ecoregion. Soils of the Southeastern Plains and Mid Atlantic Coastal Plains ecoregions have higher sand contents than soils of the Piedmont or Highland ecoregions. Higher soil sand content often leads to higher sand in streams, which can stress many non-Coastal Plain fish species [64–66].

The importance of watershed size for many species may be due to species-specific preference for smaller systems (i.e., headwater species). For example, creek chub and central stoneroller prefer smaller systems [67, 68], which may explain the reduction in probability of presence for these two species at an area of ~60 km^2. The lack of a reduction in probability of presence for the other two species may be due to omission of streams larger than fourth order from the stream survey—extending the data set to include larger systems (up to rivers) may reveal similar patterns in the other fishes.

The initial, rapid increase in the probability of presence with watershed area up to ~10 km^2 for all four focal species may be due to our focus on systems with <200 km^2 in drainage area. Streams with <10 km^2 in drainage area were small (mean width=2.4 m) and headwaters positioned far up the stream network (e.g., mean distance to mainstem tributary >500 km^2=20.9 km). Such streams are likely controlled by stochastic processes and disconnected from source populations [69, 70], both of which may limit fish colonization. Moreover, available habitat for many species is likely less in such small systems.

The secondary role of anthropogenic land cover in our models does not imply that land use is unimportant. The cRF models first partitioned each species data set by values of controlling natural variables (see above), but then further separated the sites using land cover variables, like percent impervious cover. Natural attributes seem to drive regional differences in stream fish distribution, while land use affects patterns within regions. Clearly SDMs must consider natural attributes as well as land use to achieve high accuracy for regional analyses.

Lower importance of impervious surface in the models warrants further investigation, because impervious surfaces have drastic effects on stream ecosystems [71] and overall fish assemblage structure and integrity [15]. Three of the focal species (central stoneroller, creek chub, and white sucker) utilize mineral substrates during spawning [72], which is likely affected by altered flow regimes associated with impervious surfaces. Moreover, the herbivorous central stoneroller may lose algal resources in stream beds altered by impervious surface. Largemouth bass is a non-native fish, and the invasion and persistence of non-native species are often linked to habitat disturbances associated with landscape alterations such as urbanization [73]. Thus, the

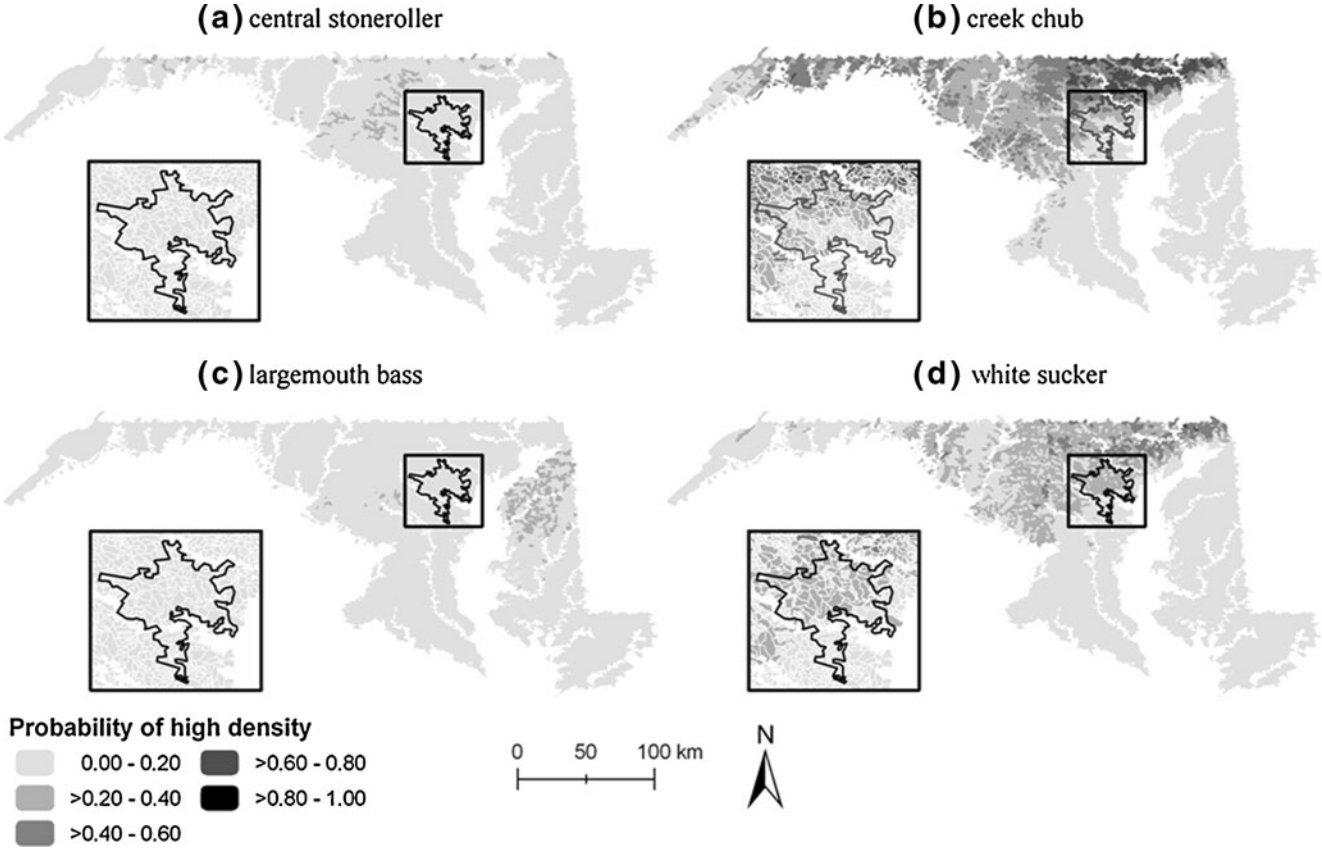

Fig. 5 Habitat suitability for high-density populations from density category predictions for four example species in all small nontidal reaches in the study area (Online resource 1). The *insets* show enlarged views of the results near Baltimore, Maryland

potential for effects of impervious surface on fishes is strong. However, our models show a lower importance of impervious surface, indicating that natural factors (e.g., stochasticity, dispersal constraints, lack of available habitat; see above) are more influential on fish presence/absence in small streams than anthropogenic factors, especially at regional scales such as Maryland. Often effects of impervious cover on fishes are identified through indirect effects on stream habitat—incorporation of these measures (e.g., local habitat) may elucidate an intermediate effect of impervious surface on fish populations.

4.2 Presence/Absence vs. Density Category Models

SDMs can be created from presence/absence, density categories, or raw abundance data. We showed that SDMs constructed using presence/absence or density categories gave qualitatively similar results for two of the four example species (creek chub and white sucker), but the density categories provide more information than just presence/absence and may help identify potential conservation sites with high density populations. Of course, field verification of the sites would be needed, but the models can help target the field efforts.

An important caveat with the binned density approach is that creation of bins reduces model accuracy. Adding categories creates additional misclassification possibilities; therefore, creating too many bins would reduce model performance. The number of density categories must balance information gained from adding categories against the reduction in model performance.

5 Conclusions

Species distribution models are important conservation and management tools. The models can reliably predict areas suitable for species occupation, and the species responses to particular environmental variables can suggest management options. Presence/absence and categorical density approaches can both provide information to help target conservation programs, mainly by suggesting sites for in-depth field surveys, especially for rare, threatened, or endangered species. SDMs could also be used to suggest stream reaches that are likely to support high densities of non-native species. These streams can be targeted by conservation biologists to search for the more altered habitats suitable for non-natives and potentially mitigate anthropogenic impacts leading to habitat degradation.

Acknowledgments We thank the Maryland Department of Natural Resources and the MBSS field crew members for providing the MBSS data and Matt Baker for watershed boundaries. We thank Lori Davias and an anonymous reviewer for constructive feedback on an earlier version of this manuscript. This research was partly funded by an REU fellowship to DEM. Additional support was provided by a Smithsonian Post-Doctoral Research Fellowship awarded to KOM.

References

1. WWF. (2006). In C. Hails, J. Loh, & S. Goldfinger (Eds.), *Living planet report 2006* (p. 41). Gland: World Wildlife Fund for Nature.
2. Allan, J. D., & Flecker, A. S. (1993). Biodiversity conservation in running waters. *BioScience, 43*(1), 32–43.
3. McKinney, M. L., & Lockwood, J. L. (1999). Biotic homogenization: A few winners replacing many losers in the next mass extinction. *Trends in Ecology & Evolution, 14*(11), 450–453.
4. Rahel, F. J. (2002). Homogenization of freshwater faunas. *Annual Review of Ecology and Systematics, 33*, 291–315. doi:10.1146/annurev.ecolysis.33.010802.150429.
5. Allan, J. D. (2004). Landscapes and riverscapes: The influence of land use on stream ecosystems. *Annual Review of Ecology, Evolution, and Systematics, 35*, 257–284.
6. Kareiva, P., Watts, S., McDonald, R., & Boucher, T. (2007). Domesticated nature: Shaping landscapes and ecosystems for human welfare. *Science, 316*(5833), 1866–1869.
7. Vitousek, P. M., Mooney, H. A., Lubchenco, J., & Melillo, J. M. (1997). Human domination of earth's ecosystems. *Science, 277* (5325), 494–499.
8. Mitsch, W. J., & Gosselink, J. G. (2000). *Wetlands* (3rd ed.). New York: John Wiley & Sons.
9. Austin, M. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling, 200*(1–2), 1–19.
10. Austin, M. P. (2002). Spatial prediction of species distribution: An interface between ecological theory and statistical modelling. *Ecological Modelling, 157*(2–3), 101–118.
11. Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling, 135*(2–3), 147–186.
12. Benke, A. C. (1990). A perspective on America's vanishing streams. *Journal of the North American Benthological Society, 9* (1), 77–98.
13. USEPA. (2006). *Wadeable streams assessment: A collaborative survey of the Nation's streams* (p. 82). Washington, DC: Office of Water, United States Environmental Protection Agency. plus appendices.
14. Allan, J. D., Erickson, D. L., & Fay, J. (1997). The influence of catchment land use on stream integrity across multiple spatial scales. *Freshwater Biology, 37*, 149–161.
15. Wang, L., Lyons, J., Kanehl, P., & Bannerman, R. (2001). Impacts of urbanization on stream habitat and fish across multiple spatial scales. *Environmental Management, 28*(2), 255–266.
16. Wang, L., Lyons, J., Kanehl, P., & Gatti, R. (1997). Influences of watershed land use on habitat quality and biotic integrity in Wisconsin streams. *Fisheries, 22*(6), 6–12.
17. Maloney, K. O., Mitchell, R. M., & Feminella, J. W. (2006). Influence of catchment disturbance on *Pteronotropis euryzonus* (Broadstripe shiner) and *Semotilus thoreauianus* (Dixie chub). *Southeastern Naturalist, 5*(3), 393–412.
18. Schleiger, S. L. (2000). Use of an index of biotic integrity to detect effects of land uses on stream fish communities in west-central Georgia. *Transactions of the American Fisheries Society, 129*, 1118–1133.
19. Schlosser, I. J. (1982). Fish community structure and function along two habitat gradients in a headwater stream. *Ecological Monographs, 52*(4), 395–414.
20. Elith, J., Graham, C. H., Anderson, R. P., Dudik, M., Ferrier, S., Guisan, A., et al. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography, 29*(2), 129–151 [Review].
21. Buisson, L., Blanc, L., & Grenouillet, G. (2008). Modelling stream fish species distribution in a river network: The relative effects of temperature versus physical factors. *Ecology of Freshwater Fish, 17*(2), 244–257. doi:10.1111/j.1600-0633.2007.00276.x [Article].
22. McNyset, K. M. (2005). Use of ecological niche modelling to predict distributions of freshwater fish species in Kansas. *Ecology of Freshwater Fish, 14*(3), 243–255. doi:10.1111/j.1600-0633.2005.00101.x.
23. Oakes, R. M., Gido, K. B., Falke, J. A., Olden, J. D., & Brock, B. L. (2005). Modelling of stream fishes in the Great Plains, USA. *Ecology of Freshwater Fish, 14*(4), 361–374.
24. Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters, 8*(9), 993–1009.
25. Hopkins, R. L., II, & Burr, B. M. (2009). Modeling freshwater fish distributions using multiscale landscape data: A case study of six narrow range endemics. *Ecological Modelling, 220*(17), 2024–2034.
26. Maloney, K. O., Weller, D. E., Russell, M. J., & Hothorn, T. (2009). Classifying the biological condition of small streams: An example using benthic macroinvertebrates. *Journal of the North American Benthological Society, 28*(4), 869–884. doi:10.1899/08-142.1.
27. Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics, 8*. doi:10.1186/1471-2105-8-25.
28. Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., & Hess, K. T. (2007). Random forests for classification in ecology. *Ecology, 88*(11), 2783–2792.
29. Omernik, J. M. (1987). Ecoregions of the conterminous United States. *Annals of the Association of American Geographers, 77*(1), 118–125.
30. Peel, M. C., Finlayson, B. L., & McMahon, T. A. (2007). Updated world map of the Koppen-Geiger climate classification. *Hydrology and Earth System Sciences, 11*, 1633–1644.
31. A brief description of the geology of Maryland (1981). Maryland Geological Survey, Baltimore, Maryland, USA. http://www.mgs.md.gov/esic/brochures/mdgeology.html.
32. MD DNR (Maryland Department of Natural Resources). (2005). *Maryland wildlife diversity conservation plan* (p. 363). Annapolis: Maryland Department of Natural Resources.
33. USEPA. (1999). In D. Boward, P. Kazyak, S. Stranko, M. Hurd, & A. Prochaska (Eds.), *From the mountains to the sea: The state of Maryland's freshwater streams* (p. 54). Washington, DC: Office of Research and Development, United States Environmental Protection Agency.
34. Southerland, M. T., Rogers, G. M., Kline, M. J., Morgan, R. P., Boward, D. M., Kazyak, P. F., et al. (2005). *Maryland Biological Stream Survey 2000-2004, Volume XVI: New biological indicators to better assess the condition of Maryland streams*. Annapolis: Maryland Department of Natural Resources, Chesapeake Bay and Watershed Programs.
35. Pyne, M. I., Rader, R. B., & Christensen, W. F. (2007). Predicting local biological characteristics in streams: A comparison of landscape classifications. *Freshwater Biology, 52*(7), 1302–1321.
36. King, R. S., Baker, M. E., Whigham, D. F., Weller, D. E., Jordan, T. E., Kazyak, P. F., et al. (2005). Spatial considerations for linking watershed land cover to ecological indicators in streams. *Ecological Applications, 15*(1), 137–153.

37. Homer, C., Huang, C. Q., Yang, L. M., Wylie, B., & Coan, M. (2004). Development of a 2001 National Land-Cover Database for the United States. *Photogrammetric Engineering and Remote Sensing, 70*(7), 829–840.
38. Beyer, H. L. (2007). *Hawth's analysis tools for ArcGIS*. Available at: http://www.spatialecology.com/htools.
39. PRISM Climate Group (2006). Parameter-elevation regressions on independent slopes model (PRISM). http://www.prism.oregonstate.edu/.
40. Soil Survey Staff (2009). U.S. general soil map (STATSGO2) for Maryland. Natural Resources Conservation Service, United States Department of Agriculture.
41. Dicken, C. L., Nicholson, S. W., Horton, J. D., Kinney, S. A., Gunther, G., Foose, M. P., et al. (2005). Preliminary integrated geologic map databases for the United States: Delaware, Maryland, New York, Pennsylvania, and Virginia. Version 1.1. http://pubs.usgs.gov/of/2005/1325/. Accessed 01 June 2009.
42. Neuendorf, K. K. E., Mehl, J. P., Jr., & Jackson, J. A. (2005). *Glossary of geology* (5th ed.). Alexandria: American Geological Institute.
43. R Development Core Team. (2008). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
44. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont: Wadsworth International Group.
45. De'ath, G., & Fabricius, K. E. (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology, 81*(11), 3178–3192.
46. Loh, W. Y. (2008). Classification and regression tree methods. In F. Ruggeri, R. S. Kenett, & F. W. Faltin (Eds.), *Encyclopedia of statisitcs in quality and reliability* (pp. 315–323). Chichester: Wiley.
47. Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics, 15*(3), 651–674.
48. Freeman, E. (2009). Presence-absence model evaluation, PresenceAbsence package v. 1.1.3. http://cran.r-project.org/web/packages/PresenceAbsence/index.html.
49. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.
50. Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*(4857), 1285–1293.
51. USGS (US Geological Survey) (2006). National hydrography dataset (NHD) plus. http://www.horizon-systems.com/nhdplus/.
52. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. doi:10.1177/001316446002000104.
53. Meyer, D., Zeileis, A., Hornik, K. (2008). vcd: Visualizing categorical data. R package version 1.2-0, CRAN.R-project.org.
54. Nguyen, P. (2008). The nonbinROC package.
55. Kuhn, M. (2010). caret: Classification and regression training. R package version 4.34. http://cran.r-project.org/web/packages/caret/index.html.
56. Matthews, W. J., & Robison, H. W. (1998). Influence of drainage connectivity, drainage area and regional species richness on fishes of the Interior Highlands in Arkansas. *The American Midland Naturalist, 139*(1), 1–19.
57. Zorn, T. G., Seelbach, P. W., & Wiley, M. J. (2002). Distributions of stream fishes and their relationship to stream size and hydrology in Michigan's Lower Peninsula. *Transactions of the American Fisheries Society, 131*(1), 70–85.
58. Argent, D. G., Bishop, J. A., Jay, R., Stauffer, J., Carline, R. F., & Myers, W. L. (2003). Predicting freshwater fish distributions using landscape-level variables. *Fisheries Research, 60*(1), 17–32.
59. Walters, D. M., Leigh, D. S., Freeman, M. C., Freeman, B. J., & Pringle, C. M. (2003). Geomorphology and fish assemblages in a Piedmont river basin, U.S.A. *Freshwater Biology, 48*(11), 1950–1970.
60. Sheldon, A. L. (1968). Species diversity and longitudinal succession in stream fishes. *Ecology, 49*(2), 194–198.
61. Huet, M. (1959). Profiles and biology of western European streams as related to fish management. *Transactions of the American Fisheries Society, 88*(3), 155–163.
62. Marsh-Matthews, E., & Matthews, W. J. (2000). Geographic, terrestrial and aquatic factors: Which most influence the structure of stream fish assemblages in the midwestern United States? *Ecology of Freshwater Fish, 9*(1–2), 9–21.
63. Nekola, J. C., & White, P. S. (1999). The distance decay of similarity in biogeography and ecology. *Journal of Biogeography, 26*(4), 867–878.
64. Cordone, A. J., & Kelley, D. W. (1961). The influences of inorganic sediment on the aquatic life of streams. *California Fish and Game, 47*, 189–228.
65. Newcombe, C. P., & Jensen, J. O. T. (1996). Channel suspended sediment and fisheries: A synthesis for quantitative assessment of risk and impact. *North American Journal of Fisheries Management, 16*(4), 693–727.
66. Waters, T. F. (1995). *Sediment in streams: Sources, biological effects, and control*. Bethesda: American Fisheries Society.
67. Boschung, H. T., Jr., & Mayden, R. L. (2004). *Fishes of Alabama*. Washington D.C.: Smithsonian Books.
68. Cyterski, M., & Barber, C. (2006). Identification and prediction of fish assemblages in streams of the Mid-Atlantic Highlands, USA. *Transactions of the American Fisheries Society, 135*(1), 40–48.
69. Grant, E. H. C., Lowe, W. H., & Fagan, W. F. (2007). Living in the branches: Population dynamics and ecological processes in dendritic networks. *Ecology Letters, 10*(2), 165–175. doi:10.1111/j.1461-0248.2006.01007.x.
70. Thorp, J. H., Thoms, M. C., & Delong, M. D. (2006). The riverine ecosystem synthesis: Biocomplexity in river networks across space and time. *River Research and Applications, 22*(2), 123–147. doi:10.1002/rra.901.
71. Walsh, C. J., Roy, A. H., Feminella, J. W., Cottingham, P. D., Groffman, P. M., & Morgan, R. P. (2005). The urban stream syndrome: Current knowledge and the search for a cure. *Journal of the North American Benthological Society, 24*(3), 706–723.
72. Jenkins, R. E., & Burkhead, N. M. (1994). *Freshwater fishes of Virginia*. Bethesda: American Fisheries Society.
73. Meador, M. R., Coles, J. F., & Zappia, H. (2005). Fish assemblage responses to urban intensity gradients in contrasting metropolitan area: Birmingham, Alabama and Boston, Massachusetts. *American Fisheries Society Symposium, 47*, 409–423.