Transactions of the American Fisheries Society 137:976–996, 2008

DOI: 10.1577/T07-119.1

[Article]

Classification Tree Models for Predicting Distributions of Michigan Stream Fish from Landscape Variables

PAUL J. STEEN*
School of Natural Resources and Environment, University of Michigan, 170 Dana Building, Ann Arbor, Michigan 48109, USA

TROY G. ZORN
Michigan Department of Natural Resources, Marquette Fisheries Research Station, 484 Cherry Creek Road, Marquette, Michigan 49855, USA

PAUL W. SEELBACH
Michigan Department of Natural Resources, Institute for Fisheries Research, 212 Museums Annex Building, Ann Arbor, Michigan 48109, USA

JEFFREY S. SCHAEFFER
U.S. Geological Survey, Great Lakes Science Center, 1451 Green Road, Ann Arbor, Michigan 48105, USA

Abstract.—Traditionally, fish habitat requirements have been described from local-scale environmental variables. However, recent studies have shown that studying landscape-scale processes improves our understanding of what drives species assemblages and distribution patterns across the landscape. Our goal was to learn more about constraints on the distribution of Michigan stream fish by examining landscape-scale habitat variables. We used classification trees and landscape-scale habitat variables to create and validate presence–absence models and relative abundance models for Michigan stream fishes. We developed 93 presence–absence models that on average were 72% correct in making predictions for an independent data set, and we developed 46 relative abundance models that were 76% correct in making predictions for independent data. The models were used to create statewide predictive distribution and abundance maps that have the potential to be used for a variety of conservation and scientific purposes.

Environmental complexity and species interactions make it difficult to learn the exact abiotic habitat constraints on a population. Researchers commonly use statistical models for this by searching for patterns between species occurrences or abundances and the environmental characteristics of sampled locations. These models are used for two important purposes: (1) to formulate and test hypotheses about the factors and processes that exert important effects on organisms and (2) to make predictions of species distributions and abundances for use in management and conservation decisions.

Traditionally, fish habitat requirements have been described from site- or local-scale environmental variables (Fausch et al. 1988). Habitat variables measured at this scale are useful to managers because small-scale habitat can be manipulated (Fausch et al. 1988; Vaughan and Ormerod 2003). Local-scale variables, such as cover and substrate, are measured on short river reaches and affect food, refuge habitat, spawning habitat, and ultimately fish abundance. Three well-known modeling approaches—the U.S. habitat suitability index, river invertebrate prediction and classification system, and Australian rivers assessment scheme—are based on local-scale environmental variables (Seelbach et al. 2002).

Modeling at a site-scale level is generally expensive, and in some cases it is impossible to measure site attributes everywhere within a study region (Seelbach et al. 2002). Beyond this practical concern, an important tenet of ecology states that "different processes are likely to be important on different scales"(Levin 1992); researchers may be completely unaware of important large-scale processes that impact fish if only site-scale habitat data are studied (Wiley et al. 1997; Fausch et al. 2002; Allan 2004).

In the past 15 years, the advent of powerful geographical information systems (GIS) tools has made it possible to study spatial variation in fish distributions and abundance from a larger landscape perspective and to incorporate habitat attributes measured at larger spatial scales. Modeling based on GIS uses a variety of large-scale, map-based variables

* Corresponding author: psteen@usgs.gov

Received May 29, 2007; accepted October 19, 2007
Published online June 5, 2008

(e.g., geology and climate) that influence an aquatic system's hydrological and thermal characteristics (Wiley et al. 1997). Modeling at this scale often incorporates land use patterns as well, because they influence amounts and rates at which sediment, pollutants, and water are delivered to the system (Schlosser 1991).

Fish species are clearly influenced by processes that operate on larger spatial scales and slower temporal scales than those measured locally (Richards et al. 1996; Leftwich et al. 1997; Rathert et al. 1999; Allan 2004). Although fish are responding mechanistically to what is happening in their immediate surroundings, those local-scale factors are directly caused by the larger landscape. For example, although stream temperature is measured at a specific location, it is controlled by a combination of local- and landscape-scale processes (Wehrly et al. 2003, 2006). Also, a stream's hydrologic flow regime, which is crucial for fish communities, is driven by factors measured at a catchment scale (Poff et al. 1997).

Models based on landscape-scale processes are becoming more common. Wiley et al. (1997) produced trout population density models using only landscape-scale variables, whereas Zorn et al. (1998) used catchment area and low-flow yield as key variables in predicting fish assemblages in Michigan. Zorn et al. (2004) also used landscape-scale variables with multiple linear regression to predict fish assemblages. Close associations have also been recognized between fish assemblages and hydrologic variability, watershed size, gradient, and percent forest cover (Poff and Allan 1995; Maret et al. 1997).

In addition to providing understanding into processes that drive the fish distributions, there are many other reasons to develop models that study the relationship of landscape-scale environmental variables and fish populations. Such models provide insight to how aquatic ecological systems function, predict potential population sites, and identify areas for population restoration (Fausch et al. 1988; Maret et al. 1997; Wiley et al. 1997; Olden 2001; Olden and Jackson 2002). This is especially important for Michigan stream fish communities. Michigan possesses a diverse array of streams ranging from nationally renowned trout fisheries to diverse warmwater and coolwater communities that support recreational angling for a variety of game species. In addition, maintaining the diversity of nongame stream fishes is an important conservation goal. Both fisheries managers and nongame biologists need further understanding of the processes that regulate stream fish communities within the state. However, accumulation of broad knowledge about Michigan stream communities has been hindered because although historical fish data are plentiful, a relatively small percentage of stream reaches has been sampled.

In this study, our goal was to learn more about large-scale factors that influence the distribution of Michigan stream fish. To do this, we used landscape-scale habitat variables and three sources of data on Michigan fish distributions to create and validate models that predicted presence–absence (PA) and relative abundance (RA) of Michigan fishes. Specific objectives were to (1) build classification tree models for Michigan stream fish, (2) assess each model for validity using an independent data set, (3) describe the general structure and behavior of the models, (4) understand patterns in model error and model limitations, and (5) use the models to describe relationships between fish communities and landscape-scale habitat variables.

Methods

Habitat variables.—Data for predictor variables used in this study were obtained through the combined efforts of the Great Lakes Aquatic Gap Analysis Program (GLGAP; GLSC 2006) and the Classification and Impairment Assessment of Upper Midwest Rivers (CIAUMR; Brenden et al. 2006; UM 2006). These groups have established a high-resolution, GIS-linked database containing characteristics of Michigan's rivers. The database was referenced to a group of ArcGIS line coverages (Environmental Systems Research Institute [ESRI], Inc., Redlands, California) in which each river was divided into interconfluence reaches. Line coverages were based on the U.S. Geological Survey's National Hydrography Dataset (USGS 2006) at the 1:100,000 scale but were updated to provide more accurate representation of Michigan rivers (Brenden et al. 2006). The database describes 31,817 Michigan stream reaches (86,983 km of stream length) and includes information on a wide variety of landscape-scale environmental variables for each stream reach, such as soil permeability, land cover, stream position, bedrock and surficial geology, modeled water temperature, climate data, modeled exceedence flows, and modeled phosphorus (Brenden et al. 2006).

The database contained approximately 320 variables for each stream reach; we chose to combine some and remove others, resulting in a list of 23 variables that we hypothesized to have the most direct mechanistic relationships to fish distributions (Table 1). Reducing the number of predictors was essential for reducing collinearity between variables, improving model interpretability, and reducing probability of spurious correlations. Not all correlated variables were removed;

TABLE 1.—List of habitat and land use stressor variables used in the creation of presence–absence and relative abundance models for predicting distributions of Michigan stream fishes. The descriptive statistics summarize the entire Michigan stream population as described by the database constructed by the Gap Analysis Program and the Classification and Impairment Assessment of Upper Midwest Rivers.

Variable code	Variable description	Minimum	Maximum	Mean	SD
	Temperature				
WATER_TEMP	Water temperature (°C), predicted July mean	12.3	26.2	19.5	3.0
WT_MAAT	Mean annual air temperature (°C)	3.7	9.8	7.3	1.7
	Position in catchment				
CATCHAREA	Area of the watershed (km^2)	0.72	14,103.5	721.0	1,680.6
	Connectivity				
UP_POND	Distance (m) upstream to closest pond ≥5 acres	0	57,566.4	8,948.0	10,580.0
DOWN_POND	Distance (m) downstream to closest pond ≥10 acres or one of the Great Lakes	0	195,470.1	29,732.2	35,989.0
LINKDCATCH	Distance (m) from downstream reach with ≥10% catchment area than target reach	0	58,851.0	2,871.0	7,115.2
DOWN_LENGTH	Distance (m) from downstream end of reach to one of the Great Lakes	0	130,093.1	31,886.8	31,417.6
	Geology and hydrology				
WT_FINE	Fine-grain surficial geology (% of watershed)	0	1	0.11	0.22
WT_COARSE	Coarse-grain surficial geology (% of watershed)	0	1	0.65	0.36
TEN_YIELD	10% exceedence flow yield ($m^3/s/km^2$)	0.0075	0.0416	0.0186	0.0037
NINETY_YIELD	90% exceedence flow yield ($m^3/s/km^2$)	0.0001	0.0264	0.0039	0.0031
GRADIENT	Channel gradient	0	0.0288	0.0026	0.0038
TEN_POWER	High-flow-based specific power ($m^3/s/km^2$)	0	0.0073	0.0005	0.0008
NINETY_POWER	Summer-flow-based specific power ($m^3/s/km^2$)	0	0.0021	0.0001	0.0002
	Land use				
WT_FOREST	Forest land cover (% of watershed)	0.02	0.95	0.41	0.24
WT_WETLAND	Wetland land cover (% of watershed)	0	0.56	0.15	0.08
WT_AGR	Agricultural land use (% of watershed)	0	0.95	0.28	0.25
WT_URBAN	Urban land use (% of watershed)	0	0.64	0.05	0.07
RT_FOREST	Forest land cover (% of riparian network)	0.02	0.90	0.28	0.16
RT_WETLAND	Wetland land cover (% of riparian network)	0.01	0.94	0.37	0.17
RT_AGR	Agricultural land use (% of riparian network)	0	0.94	0.17	0.20
RT_URBAN	Urban land use (% of riparian network)	0	0.56	0.04	0.06
	Water quality				
TOTAL_P_PPM	Total phosphorus, predicted (mg/L)	0.01	0.25	0.05	0.04

for example, it was important to retain the different types of land use and land cover because these variables are important for managers as examples of landscape-scale variables that can be manipulated. Choosing these variables was a key step in the modeling process, and our decision was based on past work on Michigan fish (Zorn et al. 2004) as well as preliminary classification trees in which we included all possible variables. The variables that we retained and their importance for fish distribution and abundance are discussed in the next several paragraphs.

Water temperature has important effects on growth and survival of fish and affects dissolved oxygen levels (Smale and Rabeni 1995; Wehrly et al. 2003, 2006; Bailey and Alanara 2006; Rand et al. 2006). Because water temperature data were not available for every stream reach, a temperature model was developed to make predictions of mean July stream temperature. In addition to water temperature, we also used mean annual air temperature, which reasonably approximates groundwater temperature and thus water temperature during base flow conditions.

Of the different types of land use data available, we used percent of forest, wetlands, agriculture, and urbanization on two scales: a 120-m (60 m on each side of the stream) riparian network stream buffer for the stream reach of interest and all streams upstream, and the total catchment area (km^2) of the stream reach. The riparian area of a stream is an important indicator of erosion control, pollution filtering capacity, shading, and woody debris potential, whereas land use of the entire catchment area has important effects on water chemistry and stream hydrology (Wang et al. 1997, 2003; Snyder et al. 2003).

Surficial geology has impacts on water chemistry and hydrology (Bent 1971). We obtained surficial geology data from 1:250,000-scale maps. We calculated the area consisting of coarse-textured geological configurations (outwash, coarse textured end moraine and till, lacustrine sand and gravel, dune sand) for the watershed

of each stream reach and divided this area by total watershed area to produce the percent of coarse surficial geology in the watershed. This was also done with fine-textured surficial configurations (fine-textured till, fine-textured end-moraine, and lacustrine clay and silt).

Several habitat variables were built from GIS-obtained information to serve as surrogates for site-scale habitat features that are important in shaping fish communities (Table 1). The 90% exceedence flow yield (exceedence flow/catchment area), which indicates the relative contribution of groundwater, served as a replacement for velocity at base flow; specific stream power at 90% exceedence flow (10 × 90% exceedence flow × gradient/catchment area) can indicate a stream's substrate (e.g., a high stream power indicates scouring of fine sediment from the channel bed). The 10% exceedence flow is a measure of a stream's peak flow that can limit recruitment and abundance of the fish population, and specific stream power at 10% exceedence flow is a measure of the stream's maximum erosive force and sediment transport capability. All flow estimates were standardized as yields by dividing values by catchment area.

Phosphorus is an essential nutrient that can limit productivity in aquatic systems (Vanni 1987; Vanni et al. 1997; Zorn et al. 2003). Because total phosphorus measurements were not available for every Michigan stream reach, we predicted it using a multiple regression equation based on 1985–1992 Michigan Rivers Inventory (MRI) phosphorus measurements and the other variables in Table 1:

$$\begin{aligned}\log_e(\text{total P}) = {} & -6.996 \\ & + (\text{percent of agriculture in watershed} \\ & \quad \times 1.497) \\ & + [\log_e(\text{stream power at} \\ & \qquad 90\% \text{ exceedence flow}) \\ & \qquad \times -0.222] \\ & + (10\% \text{ exceedence flow yield} \\ & \quad \times 59.977),\end{aligned}$$

where $N = 172$, $P < 0.001$, and adjusted $R^2 = 0.54$ (Seelbach and Wiley 1997).

There were several measured connectivity variables that take advantage of the stream connection properties inherent to NHD (Brenden et al. 2006). Variables built from these analyses include distance from the stream to one of the Great Lakes and distance from the stream to upstream and downstream lakes and ponds. Streams reaches that were disconnected from the Great Lakes by dams or waterfalls were noted. We expected these variables to be important for lake fish species that migrate into streams for certain portions of their life cycle (e.g., Chinook salmon *Oncorhynchus tshawytscha*) or fish that live in both lakes and rivers (e.g., most centrarchids). Also, the variable LINKDCATCH was created to measure the distance from the stream reach of interest to the closest downstream reach in a stream with a 10% greater catchment area than that of the stream of interest. This distance might prove useful for explaining occurrences of large-river fish in small tributaries or small-stream fish in nearby larger rivers.

Fish distribution.—We used three fish databases to create and validate the models. The MRI data set contains fish samples obtained during the 1980s and 1990s. The samples cover the geographic extent of Michigan but do have a bias against larger, non-wadeable rivers and Upper Peninsula rivers (Seelbach and Wiley 1997). We compiled fish counts (1980–2002) from the Fish Collection System (FCS) of the Michigan Department of Natural Resources' Fisheries Division. These records were collected with a wide variety of catch techniques, including electrofishing, rotenone, and seining. Given the poor catch efficiency of seining methods, we only recorded the presence of fish caught by seining and did not consider missing species as absent. We also used the Michigan Fish Atlas created by the University of Michigan's Museum of Zoology (Bailey et al. 2000). This database has Michigan fish occurrence records that date back to the mid-19th century. However, to match the time frame of the MRI and FCS data, we only used data from collections made during 1980–2000. These records were also collected with a wide variety of catch techniques and provide good spatial coverage of the state.

For all three data sets, we deleted replicate samples so that a stream reach was represented by only one sampling effort. When different samples for the same reach disagreed on species presence or abundance, we kept the observation that indicated presence or higher abundance. This method assumed that the stream reach has the potential to hold the higher amount of fish and that lower fish counts were due to disturbance unrelated to the habitat factors.

Presence–absence modeling procedure.—In a previous study, we modeled brook trout *Salvelinus fontinalis* with several different analytical techniques and determined that a classification tree method was successful for modeling with landscape-scale data (Steen et al. 2006). An explanation of classification trees has been provided by previous authors (Breiman et al. 1984; Bell 1999; De'ath and Fabricius 2000; Vayssieres et al. 2000; De'ath 2002; Holland et al. 2005; Taverna et al. 2005; Baker et al. 2006; Usio et al. 2006).

We decided to use classification trees to develop the models for all common species of Michigan stream fish. We created a species-specific PA classification

TABLE 2.—List of Michigan fish species that were modeled to determine distributions based on presence–absence (PA) and relative abundance (RA). Numbers refer to each species occurrence in the PA and RA training data (Michigan Rivers Inventory [MRI]; numbers with asterisks are from MRI and Michigan Fish Atlas). Species with insufficient data for modeling are not listed.

Species	PA	RA
Amiidae		
Bowfin *Amia calva*	77*	
Aphredoderidae		
Pirate perch *Aphredoderus sayanus*	32	24
Antherinopsidae		
Brook silverside *Labidesthes sicculus*	58*	
Catostomidae		
Quillback *Carpiodes cyprinus*	72*	
Longnose sucker *Catostomus catostomus*	41	
White sucker *Catostomus commersonii*	375	277
Creek chubsucker *Erimyzon oblongus*	39	
Lake chubsucker *Erimyzon sucetta*	57*	
Northern hog sucker *Hypentelium nigricans*	182	109
Spotted sucker *Minytrema melanops*	67*	
Silver redhorse *Moxostoma anisurum*	31	34
River redhorse *Moxostoma carinatum*	25*	
Black redhorse *Moxostoma duquesnei*	36	
Golden redhorse *Moxostoma erythrurum*	111	82
Shorthead redhorse *Moxostoma macrolepidotum*	56	24
Greater redhorse *Moxostoma valenciennesi*	35	38
Centrarchidae		
Rock bass *Ambloplites rupestris*	243	161
Green sunfish *Lepomis cyanellus*	200	128
Pumpkinseed *Lepomis gibbosus*	197	124
Warmouth *Lepomis gulosus*	97*	
Orangespotted sunfish *Lepomis humilis*	61*	
Bluegill *Lepomis macrochirus*	284	99
Longear sunfish *Lepomis peltastes*	40	
Smallmouth bass *Micropterus dolomieu*	157	89
Largemouth bass *Micropterus salmoides*	180	96
White crappie *Pomoxis annularis*	29*	
Black crappie *Pomoxis nigromaculatus*	85	110
Cobitidae		
Oriental weatherfish *Misgurnus anguillicaudatus*	29*	
Cottidae		
Mottled sculpin *Cottus bairdii*	83	172
Slimy sculpin *Cottus cognatus*	60	61
Cyprinidae		
Central stoneroller *Campostoma anomalum*	87	72
Redside dace *Clinostomus elongatus*	45*	
Lake chub *Couesius plumbeus*	43*	
Spotfin shiner *Cyprinella spiloptera*	68	39
Common carp *Cyprinus carpio*	150	76
Brassy minnow *Hybognathus hankinsoni*	77*	
Striped shiner *Luxilus chrysocephalus*	71*	
Common shiner *Luxilus cornutus*	263	203
Redfin shiner *Lythrurus umbratilis*	71*	37
Northern pearl dace *Margariscus margarita*	91	
Hornyhead chub *Nocomis biguttatus*	142	92
River chub *Nocomis micropogon*	41	
Golden shiner *Notemigonus crysoleucas*	46	32
Emerald shiner *Notropis atherinoides*	38*	
Blacknose shiner *Notropis heterolepis*	58	
Rosyface shiner *Notropis rubellus*	59	50
Sand shiner *Notropis stramineus*	39	
Mimic shiner *Notropis volucellus*	31	

TABLE 2.—Continued.

Species	PA	RA
Northern redbelly dace *Phoxinus eos*	51	
Finescale dace *Phoxinus neogaeus*	37*	
Bluntnose minnow *Pimephales notatus*	235	177
Fathead minnow *Pimephales promelas*	49	
Longnose dace *Rhinichthys cataractae*	74	69
Western blacknose dace *R. obtusus*	202	144
Creek chub *Semotilus atromaculatus*	332	243
Esocidae		
Redfin pickerel *Esox americanus*	81	28
Northern pike *Esox lucius*	168	182
Muskellunge *Esox masquinongy*	73*	
Fundulidae		
Banded killifish *Fundulus diaphanous*	57*	
Blackstripe topminnow *F. notatus*	48*	
Gadidae		
Burbot *Lota lota*	54	
Gasterosteidae		
Brook stickleback *Culea inconstans*	81	63
Ictaluridae		
Black bullhead *Ameiurus melas*	74	55
Yellow bullhead *Ameiurus natalis*	135	78
Brown bullhead *Ameiurus nebulosus*	74*	34
Channel catfish *Ictalurus punctatus*	51	26
Stonecat *Noturus flavus*	118	76
Tadpole madtom *Noturus gyrinus*	35	44
Lepisosteidae		
Longnose gar *Lepisosteus osseus*	25*	
Moronidae		
White perch *Morone americana*	32*	
White bass *Morone chrysops*	30*	
Percidae		
Eastern sand darter *Ammocrypta pellucida*	30*	
Greenside darter *Etheostoma blennioides*	35	
Rainbow darter *Etheostoma caeruleum*	151	95
Iowa darter *Etheostoma exile*	133*	
Least darter *Etheostoma microperca*	83*	
Johnny darter *Etheostoma nigrum*	289	208
Yellow perch *Perca flavescens*	101	65
Northern logperch *Percina caprodes*	92	69
Blackside darter *Percina maculata*	212	156
Walleye *Sander vitreus*	53	
Petromyzontidae		
Chestnut lamprey *Ichthyomyzon castaneus*	37	
Northern brook lamprey *Ichthyomyzon fossor*	63	
Silver lamprey *Ichthyomyzon unicuspis*	29*	
American brook lamprey *Lampetra appendix*	53	
Sea lamprey *Petromyzon marinus*	131	
Sciaenidae		
Freshwater drum *Aplodinotus grunniens*	50*	
Salmonidae		
Coho salmon *Oncorhynchus kisutch*	37*	
Rainbow trout *O. mykiss*	128	109
Chinook salmon *O. tshawytscha*	45*	
Brown trout *Salmo trutta*	196	159
Brook trout *Salvelinus fontinalis*	186	165
Umbridae		
Central mudminnow *Umbra limi*	259	179
Number of species	93	46

tree model for each of the 93 fish species that had more than 25 occurrences in our training data set (Table 2). We used the MRI data set as our training data set because it had higher sample sizes for most of the nongame fishes than did the FCS data set, which we used as our testing data. For 11 species, either the number of occurrences in the FCS data were low (<3) or the identifications of the fish were suspect. For these species, we withheld 20% of the MRI data from training to serve as a test data set (Table 3). We used the Michigan Fish Atlas data (Bailey et al. 2000) as a supplemental training database; if the MRI data did not contain at least 25 species occurrences, we added Michigan Fish Atlas data to the MRI data for model training purposes.

After the training data for a given species were pruned down through the procedures above, they were entered into Classification and Regression Trees version 5.0 (Steinberg and Colla 1997). This program produced a series of differently sized classification trees, each with a different misclassification rate for both the training data and an independent data set created from cross validation of the training data. Next, we selected the tree that minimized error in both the training and cross validation data sets. If a tree was greater than seven terminal nodes but had a lower error rate than a smaller tree, we selected the smaller tree despite its higher error rate. We believed that as the number of terminal nodes increased beyond seven, interpretation of a tree would become more difficult and such trees would contain more spurious variable splits. This decision represents our desire to have trees that are accurate yet easy to interpret. Certainly, this is not an objective decision, but it reflects our judgment and preference.

Using this tree as a starting point, we determined whether the variable splits in the tree could possess ecological meaning. Splits that lacked ecological meaning were those created at an unreasonable value (e.g., the most common spurious split was a percent land use split of <1%). Because it was unlikely that these values had any significance for fish distribution, we removed these variables from the analysis and recreated the tree to develop a better model. If there were no spurious variable splits, we accepted the tree as the final PA model. Figure 1 gives an example of a final PA model for the brown bullhead.

The FCS test data set was applied to the final model to get a benchmark of the model's accuracy by estimating the percentage of observations predicted correctly. In addition, we calculated the true skill statistic (TSS) for the FCS data. The TSS and its predecessor, Cohen's kappa, are relatively new ways to measure the accuracy of PA models and address the problem reported by Fielding and Bell (1997) of inflated accuracy ratings for rare species. The TSS is a PA assessment score that accounts for errors and success via random guessing; it ranges from −1 to 1, where 1 indicates perfect prediction and values of −1 to 0 indicate that prediction success is worse than the success attained by random guessing (Allouche et al. 2006). However, we primarily examined the percentage accuracy rating rather than TSS because percent accuracy is more intuitive than TSS and creates more interesting and more easily understood results. In addition, results indicated that TSS consistently underestimated the value of models with a large discrepancy between the number of observations indicating presence and the number indicating absence.

Presence–absence model error.—We identified sites from the FCS test data set that had misclassified fish predictions—in other words, sites where predicted PA did not match the observation. These types of errors are usually described as false positive (predicted present when observed absent) and false negative (predicted absent when observed present). For example, when a FCS sampling site had 10 false positive errors, this meant that 10 fish species were predicted to be present in the stream but were not observed there.

We examined the correlation matrix of false positive and false negative errors for a site and the habitat values for the stream reach where the sampling site was located. We did this to determine whether there were any patterns between model error and the habitat variables; such patterns can indicate whether streams with particular habitat tend to have more- or less-accurate models. To prevent the models that performed poorly from interfering with these results, we only looked at PA models with a TSS value greater than 0 and an accuracy of at least 60% (in both absence and presence) in making predictions for the test data set.

Relative abundance modeling procedure.—For the RA models, we selected MRI data obtained from two-pass electrofishing depletion samples and converted the fish counts to estimated catch per hectare. The FCS data set and Michigan Fish Atlas data set were not used in RA modeling.

We built the RA models on an individual species basis. For each species with 25 or more occurrences in the MRI data, we divided fish density estimates into three logarithmic-scale categories: low (1–10 fish/ha), medium (11–100 fish/ha), and high (>100 fish/ha). We also tried dividing density estimates into categories by equal interval and by natural breaks. However, the models performed the same or worse using these category breaks, so for simplicity we used the logarithmic scale, so that each fish species had the same abundance categories.

TABLE 3.—Sample size and percent correct agreement between predicted presence–absence (PA) values and observed values in a test data set for each PA model used to predict Michigan stream fish distributions. The list is sorted by the average of accuracy for percent present and percent absent (average accuracy); average accuracy does not consider differences in number between percent present and percent absent. The true skill statistic (TSS) for each PA model is also shown.

Species	Present		Absent		Average accuracy	TSS
	Number	Percent	Number	Percent		
Black redhorse	12	91.7	788	94.9	93.3	0.21
White perch	27	100.0	781	81.3	90.7	0.16
Channel catfish	54	81.5	760	98.0	89.8	0.73
Greenside darter[a]	8	100.0	72	79.2	89.6	0.35
Greater redhorse	13	84.6	801	93.3	89.0	0.17
Redfin shiner	21	95.2	803	82.6	88.9	0.12
Golden redhorse	47	83.0	780	94.0	88.5	0.44
Silver redhorse	11	81.8	802	94.3	88.1	0.16
White bass	19	94.7	793	79.3	87.0	0.10
Rosyface shiner[a]	15	100.0	84	71.4	85.7	0.38
Lake chub	3	100.0	803	70.0	85.0	0.01
Chinook salmon	60	88.3	786	80.2	84.3	0.24
Spotfin shiner	49	75.5	781	92.8	84.2	0.38
Mimic shiner	17	88.2	786	78.2	83.2	0.08
Blackstripe topminnow[a]	12	91.7	104	74.0	82.8	0.28
Walleye	149	71.8	698	93.0	82.4	0.63
Sea lamprey	4	100.0	801	64.7	82.3	0.01
River chub	24	70.8	800	93.0	81.9	0.22
Common carp	156	84.6	723	76.1	80.4	0.39
Emerald shiner	24	70.8	796	89.7	80.3	0.16
Tadpole madtom	22	72.7	802	87.4	80.1	0.13
Sand shiner	22	72.7	785	86.6	79.7	0.12
Black crappie	85	72.9	751	86.0	79.5	0.34
Stonecat	81	66.7	758	92.1	79.4	0.44
Yellow bullhead	97	78.4	745	78.9	78.6	0.29
Pirate perch	26	76.9	780	79.7	78.3	0.10
Slimy sculpin	28	85.7	775	70.3	78.0	0.09
Spotted sucker	12	91.7	801	63.8	77.7	0.03
Brook trout	504	75.6	586	79.7	77.7	0.55
Shorthead redhorse	30	63.3	781	90.0	76.7	0.18
Mottled sculpin[a]	15	80.0	51	72.5	76.3	0.39
White crappie	12	75.0	789	76.4	75.7	0.04
Brook silverside	7	85.7	787	65.6	75.7	0.02
Central stoneroller	105	73.3	731	77.2	75.2	0.27
Muskellunge	53	84.9	739	64.4	74.7	0.13
Rock bass	302	73.8	663	75.4	74.6	0.44
Northern pike	251	61.8	667	87.4	74.6	0.51
Coho salmon	75	72.0	763	76.0	74.0	0.19
Longnose sucker	7	85.7	802	62.2	74.0	0.02
River redhorse	3	66.7	788	81.2	74.0	0.01
Fathead minnow	37	83.8	777	63.4	73.6	0.09
Smallmouth bass	185	61.6	721	85.0	73.3	0.41
Longnose gar	11	63.6	800	83.0	73.3	0.04
Quillback	180	61.1	794	84.9	73.0	0.38
Chestnut lamprey	5	60.0	802	85.8	72.9	0.02
Redfin pickerel	101	66.3	694	78.7	72.5	0.25
Northern logperch	104	63.5	746	80.6	72.1	0.25
Longnose dace	134	67.2	717	76.7	72.0	0.28
Brassy minnow	5	80.0	801	63.5	71.8	0.01
Green sunfish	357	77.0	592	66.4	71.7	0.41
Striped shiner[a]	18	61.1	101	81.8	71.5	0.30
Yellow perch	221	61.9	650	80.2	71.0	0.38
Northern hog sucker	99	68.7	699	73.2	70.9	0.21
Finescale dace[a]	10	60.0	104	81.7	70.9	0.19
Largemouth bass	275	61.1	630	80.5	70.8	0.40
Creek chub	401	75.1	398	64.6	69.8	0.40
Bluntnose minnow	235	70.6	685	68.9	69.8	0.31
Common shiner	353	68.3	621	71.0	69.7	0.37
Brook stickleback	117	75.2	718	63.9	69.6	0.19
Oriental weatherfish[a]	8	75.0	103	64.1	69.6	0.11
Orangespotted sunfish[a]	15	66.7	106	70.8	68.7	0.18
Rainbow trout	363	67.8	783	68.3	68.0	0.32
Johnny darter	271	72.7	519	63.2	67.9	0.32
Warmouth	22	72.7	776	63.1	67.9	0.04

TABLE 3.—Continued.

Species	Present		Absent		Average accuracy	TSS
	Number	Percent	Number	Percent		
Rainbow darter	98	60.2	693	75.6	67.9	0.19
Black bullhead	78	65.4	762	70.1	67.8	0.13
Pumpkinseed	116	66.4	676	69.1	67.7	0.19
Brown trout	711	70.0	531	65.3	67.7	0.35
Hornyhead chub	137	73.7	737	61.3	67.5	0.19
Iowa darter	10	70.0	800	62.3	66.1	0.02
Brown bullhead	33	60.6	777	71.6	66.1	0.06
Redside dace	5	60.0	803	71.9	65.9	0.01
Northern redbelly dace	46	69.6	763	61.9	65.7	0.07
Burbot	98	53.0	752	77.7	65.4	0.16
Lake chubsucker	5	60.0	786	70.4	65.2	0.01
Central mudminnow	481	69.0	514	61.1	65.1	0.30
Blackside darter	259	60.2	669	69.7	65.0	0.25
Golden shiner	18	61.1	775	68.1	64.6	0.03
Bluegill	284	60.2	641	68.6	64.4	0.25
White sucker	761	66.8	379	60.7	63.7	0.25
Least darter	5	60.0	785	64.1	62.0	0.01
Bowfin	24	62.5	782	61.5	62.0	0.03
Silver lamprey[a]	10	60.0	90	63.3	61.7	0.09
Banded killifish	14	71.4	105	51.4	61.4	0.09
Longear sunfish	8	50.0	783	71.6	60.8	0.01
Northern pearl dace	16	62.5	795	52.6	57.5	0.01
Western blacknose dace	464	85.6	514	24.1	54.9	0.15
Northern brook lamprey	19	31.6	796	77.6	54.6	0.01
American brook lamprey	8	25.0	799	84.0	54.5	0.01
Creek chubsucker	14	14.3	781	84.6	49.5	0.00
Freshwater drum	33	36.4	781	62.5	49.5	0.00
Eastern sand darter	8	37.5	106	59.4	48.5	−0.01
Blacknose shiner	17	17.6	796	56.9	37.3	−0.02

[a] Of the MRI data, 20% was withheld to serve as a test data set.

To build the trees, we followed the same steps used in the PA models, except that we used three density categories instead of PA categories. Because the only density data available were from the MRI data set, we withheld 20% of the MRI sample for model validation. Several fish had greater than 25 samples but too few fish in a category to allow for withholding of a 20% validation sample (e.g., 2 observations in the low density category, 4 in medium, and 30 in high). In these cases, we used the 10-fold cross validation procedure of Steinberg and Colla (1997) to assess model performance. In this cross validation process, one-tenth of the data is held back, and the rest of the data is used to create the tree; error estimates are made for the withheld data. This is repeated until all the data have been withheld and tested, and the final testing accuracy is determined from the combination of all data subsamples.

If an RA model had an accuracy worse than random (i.e., $<33.3\%$ for any category) when applied to the test data or when used in cross validation, we created a two-category classification tree for that species. For such models, we dropped the medium density category so that the species was only predicted at a low and high RA. This also involved dropping the training data for the medium category and making the assumption that in the real world, fish density does not fall within this range. This resulted in models that were simpler and more removed from reality than the three-category models, but we think this was necessary to build RA models with good accuracy levels for these species.

Model analysis and predictions.—The most important variables predicting PA or RA for all species were determined by counting the number of times each variable occurred in the model set. We then more closely examined how the top-five variables were split in the trees to determine whether any overall patterns were caused by these variables. To prevent the poorly performing models from interfering with these results, we only looked at PA models or two-category RA models that had at least 60% accuracy in one or both categories (absence or presence; low or high density) when applied to the test data set. To include a PA model in the analysis, we also required the model to have a TSS greater than 0.

For every species, we applied the PA model to every stream reach in Michigan. For species with an RA model, we applied the model to every stream in which the species was predicted as present, and we combined the two models to produce predictions with

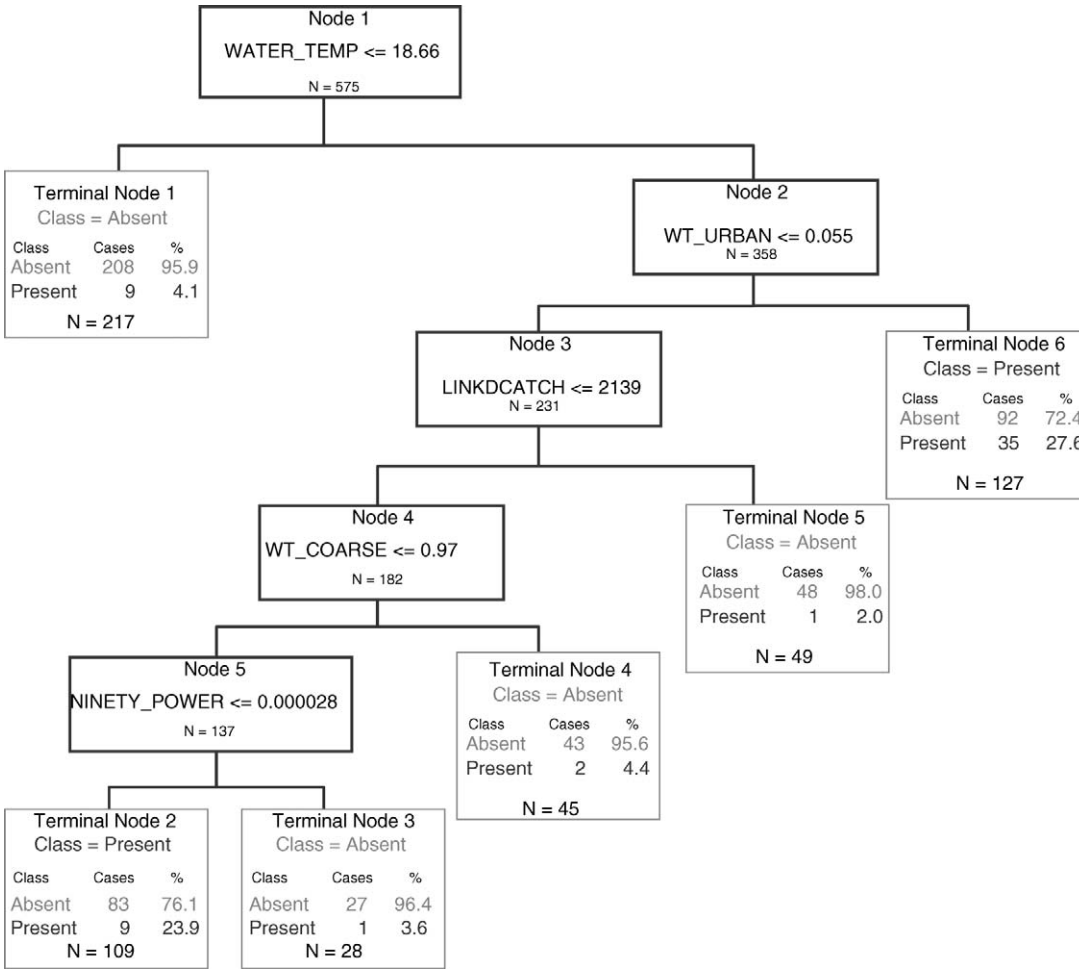


FIGURE 1.—Classification tree of a presence–absence model used for predicting the distribution of brown bullheads in Michigan streams. Variable codes are described in Table 1. An observation less than or equal to the split value (given in each box) is sent to the lower left node; otherwise, it goes to the lower right node. The terminal node indicates the final classification of the observation. Terminal nodes 2 and 6 indicate how the classification tree dealt with uneven sample sizes for presence and absence; even though they had more absence observations than presence observations, these nodes were classified as indicating presence because they contained the majority of the presence observations from the previous variable split.

three or four categories: fish absence, low RA, medium RA (when available), and high RA. The predictions were joined to the updated 1:100,000-scale NHD in the GIS to produce a statewide distribution map for each fish.

Results

Presence–Absence Models

We developed PA models for 93 Michigan stream fish (Table 2). Despite the addition of the Michigan Fish Atlas data, we did not have enough data (<25 occurrences) to create PA models for 52 of the 145 fish species found in Michigan (Bailey and Smith 2002). However, although 18 of these species are found in streams, 34 are primarily or exclusively lake species and our samples did not include lakes.

Each PA model had two measurements of percent accuracy in making predictions for the test data: percent correctly predicted presences and percent correctly predicted absences. The mean of these two scores provided an accuracy measurement (hereafter, average accuracy) that we used to compare individual species models.

For all 93 PA models combined, we predicted 72% of the test data observations correctly; 44% of the models had an average accuracy of 65–75%, including

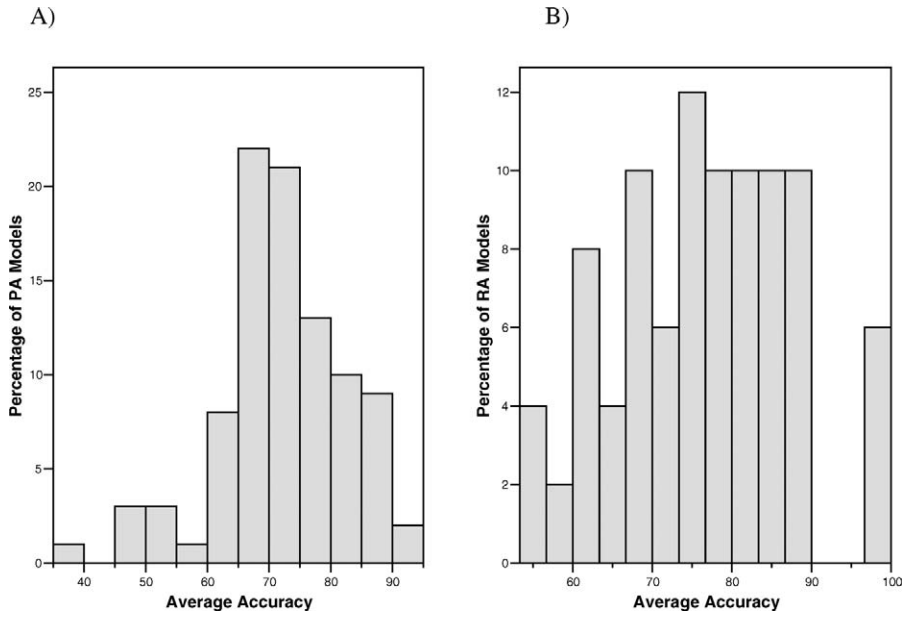


FIGURE 2.—Histograms showing the percentage of Michigan stream fish models that fell within certain ranges of average accuracy level: **(A)** 93 presence–absence (PA) models and **(B)** 46 relative abundance (RA) models.

models for the rock bass, northern pike, smallmouth bass, and yellow perch (Table 3; Figure 2). Four species models (creek chubsucker, freshwater drum, eastern sand darter, and blacknose shiner) had predictions that were worse than random (average accuracy $< 50\%$). However, 21% of PA models had an average accuracy greater than 80%, including models for the greenside darter, redfin shiner, and white perch. Fish species associated with big, slow rivers were modeled particularly well. Models for four redhorse species (black, greater, golden, and silver redhorses) had an average accuracy greater than 88%, and models for two others were nearly as accurate (river redhorse: 74.0%; shorthead redhorse: 76.7%). The channel catfish model had an average accuracy of 89.8%, and the common carp model had an average accuracy of 80.4%. Although coldwater species were not modeled as accurately as the redhorse species, brook trout, slimy sculpin, mottled sculpin, Chinook salmon, and coho salmon models all had average accuracies of about 75%.

We recorded the frequency of each habitat variable included in PA models with an average accuracy greater than 60% and a TSS greater than 0. The two variables that appeared most often were water temperature (in 45 of the 82 models) and catchment area (44 models; Table 4). Other frequently occurring variables included air temperature, predicted total phosphorus, and the 10% exceedence flow yield. All land use variables included in the models occurred with approximately the same frequency, although land use measured on the larger watershed scale occurred slightly more frequently (on average, in 14 of the 82 models) than land use measured on the riparian scale (11 models).

We examined the PA models to see if there were any patterns associated with the variable splits of the five most frequently occurring variables. Patterns in the variable splits indicate whether these important variables have consistent effects on the fish. The pattern was quite clear for water temperature; in 39 of the 45 models containing water temperature, an increase in water temperature resulted in fish presence. Not surprisingly, coldwater species were associated with five of the other six models. Brook trout, brown trout, rainbow trout, mottled sculpin, and slimy sculpin were predicted to be absent when the temperature was above 19.9°C on average. An increase in temperature resulted in absence of pirate perch also, but the split value was quite high (23°C), so this species should not be grouped with the others. Models of coolwater species (e.g., muskellunge, brook stickleback, and brassy minnow) did not have consistent water temperature patterns.

A catchment area increase resulted in a prediction of presence in 39 of the 44 models containing this

TABLE 4.—Number of times each habitat variable (defined in Table 1) was included in 82 Michigan stream fish presence–absence (PA) models with accuracy greater than 60% and the number of times each habitat variable was included in the 43 relative abundance (RA) models with accuracy greater than 60%.

PA models			RA models		
Variable code	Number	Percent	Variable code	Number	Percent
WATER_TEMP	45	54.9	CATCHAREA	18	41.9
CATCHAREA	44	53.7	TOTAL_P__PPM	14	32.6
WT_MAAT	26	31.7	WT_COARSE	12	27.9
TOTAL_P_PPM	24	29.3	NINETY_YIELD	11	25.6
TEN_YIELD	22	26.8	LINKDCATCH	10	23.3
WT_FOREST	17	20.7	GRADIENT	9	20.9
WT_COARSE	15	18.3	WT_MAAT	9	20.9
UP_POND	15	18.3	WATER_TEMP	8	18.6
TEN_POWER	15	18.3	RT_AGR	7	16.3
NINETY_YIELD	14	17.1	WT_WETLAND	7	16.3
RT_AGR	13	15.9	RT_WETLAND	7	16.3
WT_WETLAND	13	15.9	TEN_YIELD	6	14.0
WT_AGR	13	15.9	NINETY_POWER	6	14.0
WT_URBAN	12	14.6	RT_FOREST	6	14.0
RT_FOREST	12	14.6	UP_POND	6	14.0
RT_WETLAND	11	13.4	DOWN_POND	4	9.3
NINETY_POWER	10	12.2	WT_FINE	4	9.3
DOWN_POND	8	9.8	TEN_POWER	4	9.3
RT_URBAN	8	9.8	RT_URBAN	4	9.3
WT_FINE	7	8.5	WT_FOREST	3	7.0
GRADIENT	7	8.5	DOWN_LENGTH	3	7.0
LINKDCATCH	6	7.3	WT_AGR	2	4.7
DOWN_LENGTH	6	7.3	WT_URBAN	1	2.3

variable; a phosphorus increase resulted in a presence prediction for 18 of the 24 models that included this variable. The results for air temperature and 10% exceedence flow yield were ambiguous; neither presence nor absence predictions predominated when the variable value increased.

We looked at the correlation matrix between the absolute number of errors (i.e., not percentage error) made at a site in the testing data and the habitat variables for the stream. For false negative errors, the highest correlation was rather small (10% exceedence flow yield: $r = 0.17$). However, the number of false positive errors made at a site was correlated with several habitat variables. The strongest correlation was between number of false positive errors and water temperature ($r = 0.66$), indicating that as temperature increased, more species were predicted to be in streams where they were not observed. Similarly, catchment area ($r = 0.35$) and agriculture (percent agricultural land use in riparian network: $r = 0.43$; percent agricultural land cover in watershed: $r = 0.50$) were also positively correlated with the number of false positive errors at a site. On the other hand, percent of forest in the riparian zone ($r = -0.58$) and watershed ($r = -0.57$) were negatively correlated to number of false positive errors, indicating that as percent forest increased, fewer errors were made for a stream.

Relative Abundance Models

We created 46 RA models; 10 models had three abundance levels and 36 models had two abundance levels. We did not have enough data to create RA models for 47 of the species with PA models. Similar to the PA models, the RA models predicted some species very well (e.g., brook stickleback and pumpkinseed), but other species were not modeled much more accurately than random guessing (e.g., rainbow darter and rosyface shiner; Tables 5, 6; Figure 2). Overall, the accuracy of the RA models, especially the two-category models, exceeded our expectations. The average three-category model predicted low abundances correctly 71.8% of the time, medium abundances 58.5% of the time, and high abundances 79.4% of the time (Table 5). On average, the two-level model predicted low abundances 80.2% of the time and high abundances 76.9% of the time (Table 6).

We recorded the frequency at which each habitat variable occurred in the more accurate RA models (i.e., all three-level models and two-category models with >60% accuracy for both categories; Table 4). Catchment area was the most important (41.9% of models), followed by predicted total phosphorus (32.6%) and percentage of coarse surficial geology in the watershed (27.9%). Although water temperature was in about 50% of the PA models and air

TABLE 5.—Sample size (*N*) and percent correct agreement (PC) between predicted relative abundance (RA) category and observed values in a test data set for each three-category RA model. The list is sorted by the average PC for low-, medium-, and high-density categories. The average does not consider differences in *N* among categories.

Species	Low		Medium		High		Average PC
	N	PC	*N*	PC	*N*	PC	
Brook stickleback	6	66.6	5	100.0	5	80.0	82.2
Northern pike	21	85.7	20	60.0	5	100.0	81.9
Brown bullhead[a]	19	79.0	11	63.6	4	100.0	80.9
Central stoneroller	8	87.5	5	60.0	5	80.0	75.8
Longnose dace	9	77.8	3	66.7	5	60.0	68.2
Black crappie[a]	68	66.2	47	55.3	5	80.0	67.2
Greater redhorse[a]	15	53.3	20	35.0	3	100.0	62.8
Tadpole madtom[a]	9	66.7	19	52.6	26	68.8	62.7
Redfin shiner[a]	12	75.0	21	33.3	4	75.0	61.1
Silver redhorse[a]	20	60.0	12	58.3	2	50.0	56.1

[a] The species was tested with a cross validation procedure rather than a 20% validation set withheld from the original data (Steinberg and Colla 1997).

TABLE 6.—Sample size (*N*) and percent correct agreement (PC) between predicted relative abundance (RA) category and observed values in a test data set for each two-category RA model. The list is sorted by the average PC for low- and high-density categories. The average does not consider differences in *N* between categories.

Species	Low		High		Average PC
	N	PC	*N*	PC	
Channel catfish	4	100.0	3	100.0	100.0
Golden shiner	6	100.0	2	100.0	100.0
Pirate perch	2	100.0	4	100.0	100.0
Common carp	10	80.0	9	100.0	90.0
Pumpkinseed	18	94.4	13	84.6	89.5
Rock bass	14	100.0	26	76.9	88.5
Stonecat	6	100.0	13	76.9	88.5
Shorthead redhorse	4	75.0	3	100.0	87.5
Slimy sculpin	8	87.5	7	85.7	86.6
Bluntnose minnow	11	90.9	33	81.8	86.4
Yellow bullhead	10	80.0	9	88.9	84.5
Black bullhead	8	87.5	5	80.0	83.8
Grass pickerel	5	100.0	3	66.7	83.3
Golden redhorse	6	83.3	14	78.6	81.0
Blackside darter	16	81.3	24	79.2	80.3
Spotfin shiner	4	100.0	5	60.0	80.0
Northern hog sucker	11	90.9	16	68.8	79.8
Green sunfish	14	78.6	18	77.8	78.2
Largemouth bass	17	70.6	7	85.7	78.1
Western blacknose dace	19	89.5	17	64.7	77.1
Bluegill	15	73.3	10	80.0	76.7
Hornyhead chub	9	66.7	14	85.7	76.2
White sucker	32	75.0	37	75.7	75.4
Rainbow trout	14	71.4	13	76.9	74.2
Brook trout	17	64.7	24	83.3	74.0
Smallmouth bass	12	75.0	10	70.0	72.5
Mottled sculpin	24	75.0	19	68.4	71.7
Yellow perch	12	66.7	4	75.0	70.8
Central mudminnow	22	77.2	23	60.8	69.0
Northern logperch	10	80.0	7	57.1	68.6
Johnny darter	21	71.4	32	65.6	68.5
Common shiner	15	60.0	36	72.2	66.1
Brown trout	19	63.2	21	66.7	64.9
Creek chub	27	63.0	33	60.1	61.6
Rainbow darter	13	53.8	11	63.3	58.6
Rosyface shiner	5	60.0	8	50.0	55.0

temperature was in about 30%, these two variables were only in 8 (18.6%) and 9 (20.9%), respectively, of the 43 RA models. Interestingly, both gradient and downstream link moved from the bottom of the PA list to near the top of the RA list (Table 4).

We looked for patterns in the RA trees by examining splits of the most frequent variables. Although the effect of catchment area and gradient were ambiguous, a decrease in LINKDCATCH resulted in greater fish abundance in 9 of the 10 RA models that included this variable, and an increase in total predicted phosphorus was associated with an increase in fish abundance for 12 of the 14 relevant RA models. Also, an increase in the coarse surficial geology variable resulted in lower abundance for 10 of 12 RA models, and an increase in 90% exceedence flow yield resulted in lower abundance for all 11 relevant RA models.

Distribution Maps

Using the predictions generated from the models, we created either PA or absence–RA statewide distribution maps. We give an example of a map that combines the PA and RA models to classify rock bass as absent, present in low abundance, or present in high abundance within each Michigan stream (Figure 3). In this example, rock bass are predicted to be at low densities throughout the larger rivers of the Upper Peninsula and northern Lower Peninsula. The highest density of rock bass is predicted for the south-central portion of the Lower Peninsula throughout the upper portions of the Saginaw, Grand, Kalamazoo, and St. Joseph River watersheds. These predictions were tested against both PA independent data and a 20% validation sample that was withheld from the abundance training data (Figure 3).

All species maps are available upon request to the corresponding author. Also available are interactive

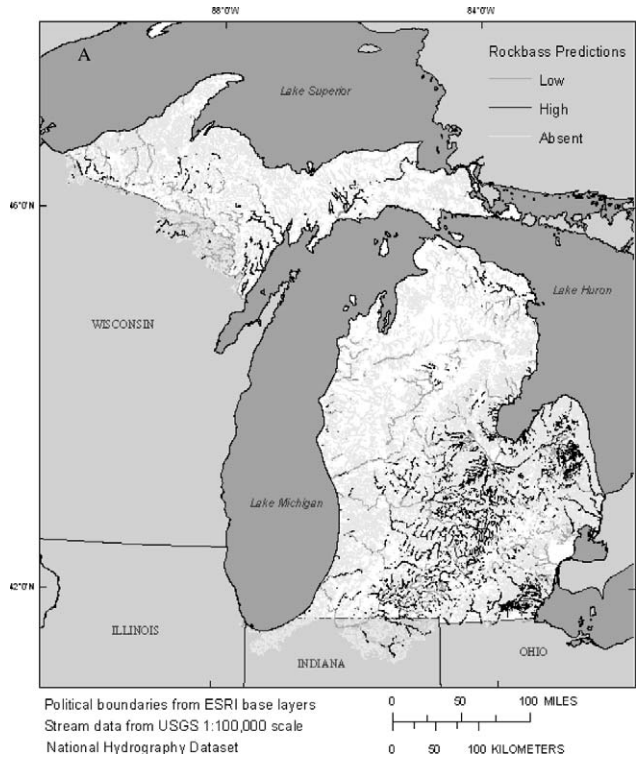


FIGURE 3.—**(A)** Map of rock bass distribution in Michigan that combines predictions from presence–absence (PA) and relative abundance (RA) models; **(B)** map of PA data that were used to test the prediction accuracy of the combined model (presence = 73.8% correct, absence = 75.4% correct); and **(C)** map of RA data (two categories) that were used to test the combined model (low abundance = 100% correct, high abundance = 76.9% correct). If the PA model predicted that a species was absent from a stream reach, then the final prediction was "absent," regardless of the RA model result. Political boundaries are from ArcGIS base layers (Environmental Systems Research Institute, Inc., Redlands, California); stream data are from the U.S. Geological Survey's National Hydrography Dataset (1:100,000 scale).

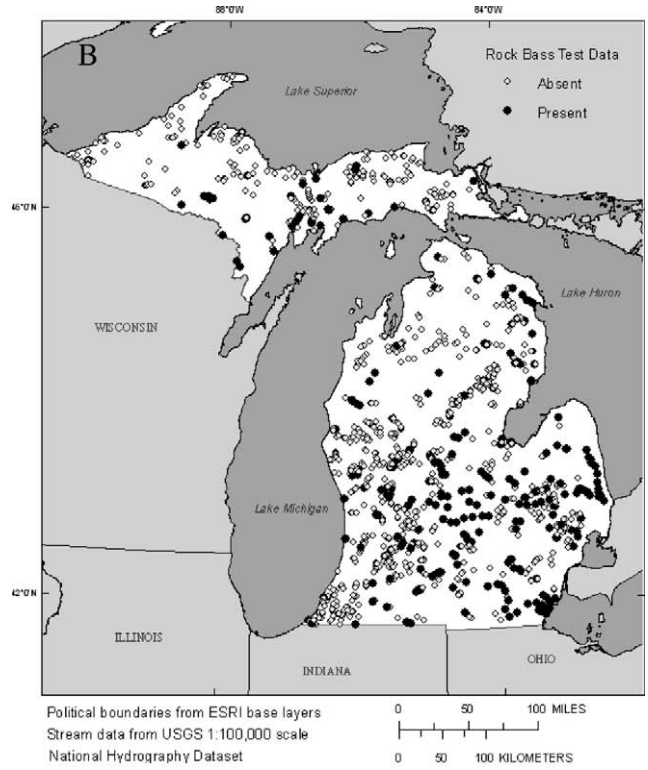
maps that run in the free downloadable program, ArcReader (ESRI). This program allows a user to query specific streams in the GIS to obtain observed and predicted fish information and the habitat variables used in the models.

Discussion

Habitat Variable Choice

We created PA models for 93 fish species typically found in Michigan streams, and we developed RA models for 46 of the 93 species. Of every 10 predictions, about 7 were accurate for the PA models, about 6 were accurate for the three-category RA models, and about 8 were accurate for the two-category RA models. This suggests that landscape-scale factors alone can be used to predict overall occurrence and abundance of most fish species in Michigan rivers when site-specific data are not available.

Optimally, we would be able to create models based on both landscape- and local-scale variables (Wiley et al. 1997). Habitat conditions at the site scale (e.g., channel morphology, substrate, and cover conditions) can have very strong effects on localized fish abundance patterns in streams. Because many of our landscape-scale variables affect local-scale mechanisms, we indirectly modeled some aspects of local-scale control. However, without direct measurement of local-scale variables, we were unable to capture all of the variation in these variables. Also, because the fish data were based on a single sample from each stream, it was impossible to detect how temporal variation could change species presence and abundance (Wiley et al. 1997). Additionally, research has shown that biological variables like competition are important for determining species occurrence and abundance (Larson and Moore 1985; Flecker and Townsend 1994; Stoks and McPeek 2003). For these reasons, we did not expect model accuracies much higher than those obtained with this model set, and errors in our predictions were expected.



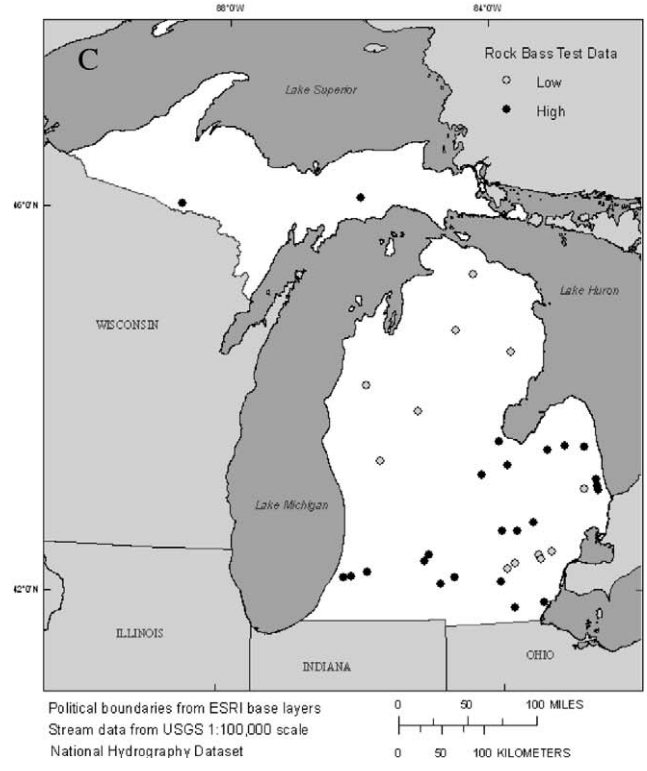


FIGURE 3.—Continued.

However, using local-scale variables to build models like those in this study would be impossible; that is, obtaining small-scale data throughout an area as large as Michigan would require prohibitive amounts of time and money. Given that research in landscape ecology has indicated that large-scale variables are as or more important for fish distribution than small-scale variables and often correlate strongly with small-scale variables, we believe our use of large-scale variables was justified and was the best approach for meeting the goals of the study (Schlosser 1991; Wiley et al. 1997; Fausch et al. 2002; Allan 2004).

Presence–Absence Model Summary

With about 70% prediction accuracy against a test data set, the PA models performed very well overall. Large-river fish, such as redhorses and channel catfish, were modeled accurately, indicating that large-scale processes determine their distribution. Coldwater species presence and absence were also predicted well. Centrarchids were typically modeled with moderate accuracy (~65–75%), indicating that landscape-scale habitat and characteristics were important; however, there are other factors determining their distribution that we were not able to detect with these models. For example, including temporal variation in fish populations would probably have increased model accuracy.

We found notable variation in model accuracy between different species, and some models performed either barely better or worse than random predictions. There are a variety of ways to explain why some fish were modeled poorly. Misidentification of fish during the data collection phase could have played a role in poor model performance, because some of the less-accurate models were built on fish species that are difficult to identify quickly in the field (e.g., silver, northern brook, and American brook lampreys). It is possible that the stream habitat data were not causally linked to the distribution of the lake species that were found in rivers, resulting in poor predictions for some of these species (e.g., burbot and freshwater drum). Some species were found virtually everywhere (e.g., white sucker and blacknose dace), and so the models were not able to distinguish between streams in which the fish were present and those in which they were absent. Unfortunately, the presence and absence of many rare species were also predicted poorly (e.g., blacknose shiner, creek chubsucker, and eastern sand darter); these rare species were historically widespread but their current distributions are much narrower due to pollution and siltation (Trautman 1981; Roberts et al. 2005). Although the predictive models for rare species did not accurately identify the current distributions, the models may be useful for indicating the potential distributions of these species.

Zorn et al. (1998) used low-flow yield (an index of water temperature) and catchment area as primary ordination axes in separating clusters of fish assemblages and explained that these two variables can reliably be used to determine which species may reside in a particular stream section. Unsurprisingly, the two most important variables in our PA models were also water temperature and catchment area. Numerous other studies have found water temperature to be key in the classification of fish (Fausch et al. 1988; Matthews and Robison 1988; Lyons 1992; Hinz and Wiley 1997; Zorn et al. 2002; Wehrly et al. 2003; Steen et al. 2006), and there is also a long history of studies on how stream changes depend on the stream's position in the catchment (Hawkes 1975; Vannote et al. 1980; Wiley et al. 1990; Smith and Kraft 2005).

Many of our GIS-based habitat variables served as surrogates for site-scale habitat variables. These variables require a conceptual leap from site-based to landscape-based modeling, and their importance in the models emphasizes the linkages between the two scales of data. Catchment area is one such variable; it is a measure of the amount of land draining to the stream and therefore is convenient for indicating a stream's approximate discharge, width, depth, and gradient (Vannote et al. 1980). These stream characteristics are highly correlated with site-scale habitat values, such as velocity, channel substrate, and dissolved nutrients (Vannote et al. 1980; Wiley et al. 1990; Rahel and Hubert 1991; Lyons 1996). Based on our models, many Michigan fish species seemed to preferentially occupy streams with larger catchment areas, indicating that larger streams with low gradient, high discharge, and warm summer water temperatures tended to favor the greatest number of species. Larger streams also have greater habitat complexity, providing space for a variety of fish species with different habitat requirements. The importance of catchment area has also been seen in previous work on fish classification and ordination (Zorn et al. 2002).

Stream yield and specific power variables are GIS-derived surrogates for stream discharge, stream velocity, substrate, erosive force, and sediment transport capability. On average, these variables were contained within about 18% of the models; thus, although they are not integral to every model, they still have important effects. For example, the models predicted correctly that black crappies, bowfins, northern pike, and black bullheads would be absent from streams with high stream power, indicating a preference for low-velocity, lentic conditions. Bluegills were present in streams with a low 10% yield; this species avoids

streams with high peak flows. Slimy sculpin were often absent from streams with a low 90% yield, showing a tendency toward occupation of groundwater-driven streams with consistent flow rather than flashy, runoff-driven streams.

The connectivity variables (e.g., distance from one of the Great Lakes, a pond, or a larger river) were included in only in about 10% of the models; however, these variables were very important in the modeling of several species. For coho salmon and Chinook salmon models, the first split in the classification tree was the variable describing distance from the Great Lakes. Both models indicated that these species are very unlikely to be found more than 122 km from one of the Great Lakes. Removing this variable from either model resulted in predictions that were only slightly better than random guessing; therefore, this variable was integral for successful prediction.

The distance from the stream to the closest of the Great Lakes also indicated whether a stream was disconnected from the Great Lakes due to a dam or waterfall. While this aspect of the variable was unimportant (and unexpectedly so) in the coho salmon and Chinook salmon models, it was important in the rainbow trout model. The rainbow trout model indicated that it was unlikely, though not impossible, for this species to be found in a stream above a dam or waterfall. This result was entirely logical given the life history of migrating steelhead (anadromous rainbow trout; because of uncertainty in the sampling database, no distinction was made between steelhead and rainbow trout during model development).

Distance from a pond or lake and distance from a large river were also key variables for several species. For example, largemouth bass, smallmouth bass, and yellow perch were more likely to be found at sites that were within 20, 8, and 6 km, respectively, of a pond or lake. The bowfin model predicted that this species would be found within 150 m of a stream's confluence with a larger river (i.e., one with a catchment area at least 10% greater than that of the stream). This variable was also important for brown bullheads (21 km) and longnose suckers (23 km). Once again, it was entirely logical that the models included these variables, because these species are typically found in lakes or slow-moving backwaters but also live in stream environments.

Presence–Absence Error Analysis

In our PA models, false positive errors occurred more frequently than false negative errors by a ratio of 8:1. False negatives are typically seen as more severe than false positives (McKenna et al. 2006); a false negative is more likely to be caused by an error in the model rather than by a failure to detect a given species during sampling. In addition, false negative errors have a severe impact on conservation work based on models. For example, if a rare species is predicted to be absent from some streams in which it actually exists, those streams might not be given the level of protection needed to conserve the species.

When distribution models are used for conservation work, false positive errors tend to be safer errors. If we do not know whether a species is present in a stream, it is safer to assume that the fish is present (i.e., erring in favor of conservation). In contrast to false negative errors, a false positive error does not necessarily indicate a flaw in the model, but rather indicates insufficient sampling, incorrect identification, or the potential for a fish to live in the stream (McKenna et al. 2006).

False positive errors may also have been caused by quality discrepancies between the training and testing data. Overall, we had a higher degree of confidence in the fish identification accuracy and catch efficiency of the MRI training data. As a result, the FCS test data probably had a higher proportion of (1) fish that were improperly identified and (2) errors due to fish that were considered absent but should have been caught during sampling. The end result of this discrepancy was a higher number of false presence errors when the models were used to make predictions about test data. In other words, the model may have correctly said that the fish should have been present in a certain stream, but the FCS data may not have been comprehensive enough to show that the species was there. Therefore, the number of false presence errors in the test data may be inflated and may underestimate the accuracy of the models, especially for hard-to-identify species.

To check this hypothesis, we compared the average false presence error rate for game fish, which are easily identified (brook trout, brown trout, smallmouth bass, largemouth bass, Chinook salmon, coho salmon, walleye, and yellow perch), against the average false presence error rate for the modeled cyprinids (chubs, daces, and minnows), which are typically harder to identify. The average false presence error rate was 19.2% for game fish and 27.2% for cyprinids. The difference between the two was not as large as we had anticipated (independent t-test: $t = -1.5$, df $= 26$, $P = 0.16$), so this species categorization method probably does not fully explain the abundance of false presence predictions. However, it is possible that the discrepancy between the data sets could account for some of the false presence errors.

Several of our habitat predictor variables were correlated to the number of false positive errors made at a stream reach. Water temperature was most strongly

correlated with false positive errors. As temperature increased, models tended to overestimate the number of species in a stream. Because warmwater streams have a higher diversity of species, sampling efforts probably missed some of the species in these streams, which would cause false positive errors in test data. Another cause of these errors may be the bias introduced into the models through the disproportionate amount of coldwater versus warmwater stream samples in our training data; predictions for coldwater streams would be more accurate since such streams contributed more data to model development.

Relative Abundance Model Summary

When using abundance categories in modeling, delineation of category boundaries is a difficult problem and usually results in incorrect predictions for observations that do not clearly belong to one category. Due to this, we were only able to create 10 species models that had test data accuracy better than that of random guessing (every abundance category $\geq$33.3%). To develop RA models for the other species, we removed the middle density category to obtain a clear distinction between the high and low categories.

Of the 46 RA models created, 10 had three categories of predicted abundance (low, medium, and high) and 36 had two categories (low and high). Interestingly, two-category models performed well and were typically more accurate than the PA models in making predictions for the test data. This implies a greater stream habitat difference between low- and high-abundance streams than between presence and absence streams. For example, a fish is considered present in a stream whether 1 individual or 1,000 individuals are captured. The PA classification tree will have difficulty in distinguishing between a marginal stream containing one individual and a stream where the species is truly absent, resulting in misclassified observations. On the other hand, when abundance categories (e.g., 1 fish = low; 1,000 fish = high) are used instead of presence, the classification tree is able to separate them with greater accuracy because there are greater habitat differences between these streams than between the marginal stream and the stream from which the species is absent.

Although most of the common species in Michigan were modeled for RA, the low number of species modeled means that the RA results may not apply to all Michigan fish. Water temperature was an unimportant variable for most of the RA models and was more important for determining PA of a species. Zorn et al. (2004) observed the same phenomenon with temperature when developing landscape-based multiple regression models. Gradient, coarse surficial geology, and 90% exceedence flow were more important in the RA models than in the PA models. Increases in these variables were associated with decreases in abundance for several species that are known to prefer streams with low slope and more variable flows (e.g., black bullhead, bluntnose minnow, largemouth bass, white sucker, and yellow perch). Given that these flow characteristics were correlated with water temperature, their importance may explain the apparent unimportance of water temperature; the exclusion of temperature from the RA classification trees may have occurred because the variation in the data was already captured.

In the PA models, probability of presence increased with increasing total predicted phosphorus; similarly, the RA models showed that abundance increased with increasing phosphorus. This is a logical result (though its frequency in the models is somewhat surprising) because phosphorus can cause a bottom-up effect, increasing productivity in every trophic level (Vanni 1987; Vanni et al. 1997). High phosphorus levels are rare in Michigan streams and therefore were not modeled; high levels cause eutrophication and anoxic conditions, which would effectively alter a fish population. For this reason, the fish PA and RA results cannot be extrapolated beyond the phosphorus range in our data.

Other General Model Limitations

Overall, these models performed well in predicting PA and RA, but their limitations should be recognized. Users of these models should be aware of these issues, and researchers constructing similar models in the future will minimize model error by addressing these problems.

Data quality is always an issue when dealing with large data sets. Brenden et al. (2006) addressed specific limitations in the NHD and the quality of GIS-derived environmental variables. In short, because some of these variables were obtained from low-resolution maps (e.g., surficial geology, 1:250,000 scale), they do not have the desired accuracy as would be obtained from the NHD with a resolution of 1:100,000. In our models, coarse surficial geology occurred relatively often (18.3% of PA models, 27.9% of RA models), and it is possible that the scaling issue increased our model error slightly.

We used several habitat variables that were built from models and then predicted across the state to produce a value for each stream reach (e.g., water temperature, total phosphorus, and flow variables). Because these habitat models contained error, it is logical to expect the error to trickle down to species models, thus decreasing model accuracy. This problem is also known as propagation of error. As these habitat

models are improved in the future, the fish predictions will become more accurate.

The fish data were of good quality overall, but the fish were sampled over a long period, by different people, and for different purposes, so it was impossible to determine which samples were poorly counted or implemented. The samplers may have misidentified or failed to catch some fish, particularly those that are hard to identify, rare, or small. Training a model on flawed data can confound the training process and produce a species model that is inaccurate, especially if the predictor variables are correlated with the likelihood of failing to detect a species in a survey. Although this issue is indeed a problem, to minimize this error we included as many sites of absence as possible in the training data for each species. By pooling sites where a species was absent, we obtained replicate information on the probability of absence as indexed by the data. Since a fish could potentially be missed at any particular site, we attempted to include several sites with the same type of habitat where the fish would not be missed. This process may not produce absolute truth for every site, but the overall distribution should be correct. The errors in the training data are reflected in the accuracy measurements; the models are not perfect but should be good enough for the use for which they were intended.

A major problem throughout this study was the difficulty in developing statewide abundance predictions. We tried several methods (regression, regression trees, and classification trees with different category boundaries), but none performed to our satisfaction. In our final product, we were only able to produce accurate models by excluding data points so that a clear distinction could be made between high- and low-abundance streams. Although this procedure did produce models that were accurate in determining high versus low abundance, the culling of data is not to be taken lightly. However, given the choice of having no RA model or having working RA models with some problems, the latter is the right decision because these models have a place in a management and conservation context.

Model Application

The models in this study were developed primarily for two large conservation projects, the GLGAP and CIAUMR. The GLGAP will use these models and similar models from New York and Wisconsin to identify fish diversity hot spots (i.e., targets for conservation or restoration work). The CIAUMR will employ the models and algal and macroinvertebrate models to provide regional assessment of stream condition (Brenden et al. 2006).

The models have utility that go beyond the scope of these projects. Models built at a landscape scale are decision-making tools that can be used in a variety of management and conservation applications. When a scientist has little information and needs a starting point or confirmation of an idea, these models excel in providing baseline data. However, they should not be used to justify management decisions without outside confirmation of the results (e.g., additional sampling).

Fisheries managers could use these models and associated data in a variety of ways. At the most basic level, these models predict the amount and location of habitat that is suitable for each fish in Michigan. Inventory information is a vital component to fisheries management and species conservation, and the modeling described here is a good way to apply this data on a large geographic scale.

For some species, a manager can rule out the presence of a fish based on a single factor. We found that trout species were unlikely to be found in streams with mean July water temperatures exceeding a particular value (19.4°C for brook trout, 20.2°C for brown trout, and 19.6°C for rainbow trout). This information in combination with the ability to access water temperature in GIS would be very useful to managers deciding whether to manage marginal streams for trout.

Managers can use the models as aids in their fish sampling and stream assessment work. The models can be used to identify candidate high-quality reference streams and low-quality impaired sites. Of course, it would be necessary to confirm this exploratory work with site visits and additional sampling.

The models can be used identify streams that should be sampled for rare species or species of concern. Other than looking at streams where a species has been found in the past, it is difficult to determine which additional streams could be part of that species' distribution. These predictive models provide tested explanations for why a species inhabits one stream and not another.

These models have the potential to be used in various decision-making processes with the goal of protecting watersheds or influencing political decisions. For example, the models can be used to build "what-if" scenarios that predict future fish distributions as influenced by climate change or land use change (e.g., urban sprawl or deforestation). They can also be used to identify streams that have a good restoration potential. For example, one could predict whether adding a forest buffer strip would have (1) a positive effect on the fish community of a stream or (2) little effect because the stream's overall potential is low regardless of land use management.

Acknowledgments

A lot of work went on behind the scenes in the development of these models, and we thank all those who participated: Steve Aichele, Ed Bissell, Arthur Cooper, and Allain Rasolofoson developed the GIS and compiled fish data; Leon Hinz developed the flow models; Li Wang, Travis Brendan, and Kevin Wehrly worked on the temperature model (publication forthcoming); Jim McKenna, Dora Passino-Reader, Jana Stewart, Li Wang, and Mike Wiley helped develop the conceptual ideas behind this project; and Jacqueline Savino, Kevin Wehrly, Solomon David, and two anonymous reviewers provided suggestions that greatly improved the manuscript. This article is Contribution 1448 of the U.S. Geological Survey Great Lakes Science Center. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

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.

Allouche, O., A. Tsoar, and R. Kadmon. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology 43:1223–1232.

Bailey, J., and A. Alanara. 2006. Effect of feed portion size on growth of rainbow trout, *Oncorhynchus mykiss* (Walbaum), reared at different temperatures. Aquaculture 253:728–730.

Bailey, R. M., W. C. Latta, R. A. Simpson, and G. R. Smith. 2000. Distribution maps of Michigan fishes. University of Michigan, Museum of Zoology, Ann Arbor.

Bailey, R. M., and G. R. Smith. 2002. Names of Michigan fishes. Michigan Department of Natural Resources, Fisheries Division, Ann Arbor.

Baker, C., R. Lawrence, C. Montague, and D. Patten. 2006. Mapping wetlands and riparian areas using Landsat ETM + imagery and decision tree-based models. Wetlands 26:465–474.

Bell, J. F. 1999. Tree-based methods. Pages 89–105 *in* A. H. Fielding, editor. Machine learning methods for ecological applications. Kluwer, Boston.

Bent, P. C. 1971. Influence of surface glacial deposits on streamflow characteristics. U.S. Ecological Survey, Water Resources Division, Open-file Report, Lansing, Michigan.

Breiman, L. F. J., R. Olshen, and C. Stone. 1984. Classification and regression trees. Chapman and Hall, New York.

Brenden, T. O., R. D. Clark, A. R. Cooper, P. W. Seelbach, L. Wang, S. Aichele, E. G. Bissell, and J. S. Stewart. 2006. A GIS framework for collecting, managing, and analyzing multiscale landscape variables across large regions for river conservation and management. Pages 49–74 *in* R. M. Hughes, L. Wang, and P. W. Seelbach, editors. Landscape influences on stream habitat and biological assemblages. American Fisheries Society, Symposium 48, Bethesda, Maryland.

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

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.

Fausch, K. D., C. L. Hawkes, and M. G. Parsons. 1988. Models that predict standing crop of stream fish from habitat variables: 1950–1985. U.S. Forest Service General Technical Report PNW-GTR-213.

Fausch, K. D., C. E. Torgersen, C. V. Baxter, and H. W. Li. 2002. Landscapes to riverscapes: bridging the gap between research and conservation of stream fish. BioScience 52:483–498.

Fielding, A. H., and J. F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation 24:38–49.

Flecker, A. S., and C. R. Townsend. 1994. Community-wide consequences of trout introduction in New Zealand streams. Ecological Applications 4:798–807.

GLSC (Great Lakes Science Center). 2006. The Great Lakes Aquatic GAP Project. Available: www.glsc.usgs.gov. (September 2007).

Hawkes, H. A. 1975. River zonation and classification. Pages 312–374 *in* B. A. Whitton, editor. River ecology. University of California Press, Berkeley.

Hinz, L. C., and M. J. Wiley. 1997. Growth and production of juvenile trout in Michigan streams: influence of temperature. Michigan Department of Natural Resources, Fisheries Research Report 2041, Ann Arbor.

Holland, G. J., S. P. R. Greenstreet, H. M. Fraser, and M. R. Robertson. 2005. Identifying sandeel *Ammodytes marinus* sediment habitat preferences in the marine environment. Marine Ecology Progress Series 303:269–282.

Larson, G. L., and S. E. Moore. 1985. Encroachment of exotic rainbow trout into stream populations of native brook trout in the southern Appalachian Mountains. Transactions of the American Fisheries Society 114:195–203.

Leftwich, K. N., P. L. Angermeier, and C. A. Dolloff. 1997. Factors influencing behavior and transferability of habitat models for a benthic stream fish. Transactions of the American Fisheries Society 126:725–734.

Levin, S. A. 1992. The problem of pattern and scale in ecology. Ecology 73:1943–1967.

Lyons, J. 1992. Using the index of biotic integrity (IBI) to measure environmental quality in warmwater streams of Wisconsin. U.S. Forest Service General Technical Report NC-149.

Lyons, J. 1996. Patterns in the species composition of fish assemblages among Wisconsin streams. Environmental Biology of Fishes 45:329–341.

Maret, T. R., C. T. Robinson, and G. W. Minshall. 1997. Fish assemblages and environmental correlates in least-disturbed streams of the upper Snake River basin. Transactions of the American Fisheries Society 126:200–216.

Matthews, W. J., and H. R. Robison. 1988. The distribution of the fishes of Arkansas: a multivariate analysis. Copeia 1988:358–374.

McKenna, J. E., R. P. McDonald, C. Castiglione, S. S. Morrison, K. P. Kowalski, and D. Passino-Reader. 2006. A broadscale fish-habitat model development process: Genesee basin, New York. Pages 533–554 *in* R. M. Hughes, L. Wang, and P. W. Seelbach, editors. Influence of landscapes on stream habitats and biological assemblages. American Fisheries Society, Symposium 48, Bethesda, Maryland.

Olden, J. D. 2001. A species-specific approach to modeling biological communities and its potential for conservation. Conservation Biology 17:854–863.

Olden, J. D., and D. A. Jackson. 2002. A comparison of statistical approaches for modelling fish species distributions. Freshwater Biology 47:1976–1995.

Poff, N. L., and J. D. Allan. 1995. Functional organization of stream fish assemblages in relation to hydrological variability. Ecology 76:606–627.

Poff, N. L., J. D. Allan, M. B. Bain, J. R. Karr, K. L. Prestegaard, B. D. Richter, R. E. Sparks, and J. C. Stromberg. 1997. The natural flow regime. BioScience 47:769–784.

Rahel, F. J., and W. A. Hubert. 1991. Fish assemblages and habitat gradients in a Rocky Mountain–Great Plains stream: biotic zonation and additive patterns of community change. Transactions of the American Fisheries Society 120:319–332.

Rand, P. S., S. G. Hinch, J. Morrison, M. G. G. Foreman, M. J. MacNutt, J. S. MacDonald, M. C. Healey, A. P. Farrell, and D. A. Higgs. 2006. Effects of river discharge, temperature, and future climates on energetics and mortality of adult migrating Fraser River sockeye salmon. Transactions of the American Fisheries Society 135:655–667.

Rathert, D., D. White, J. C. Sifneos, and R. M. Hughes. 1999. Environmental correlates of species richness for native freshwater fish in Oregon. Journal of Biogeography 26:257–273.

Richards, C. L., L. B. Johnson, and G. E. Host. 1996. Landscape-scale influences on stream habitats and biota. Canadian Journal of Fisheries and Aquatic Sciences 53(Supplement 1):295–311.

Roberts, M. E., B. M. Burr, M. R. Whiles, and V. J. Santucci. 2005. Reproductive ecology and food habits of the blacknose shiner, *Notropis heterolepis*, in northern Illinois. American Midland Naturalist 155:70–83.

Schlosser, I. J. 1991. Stream fish ecology: a landscape perspective. BioScience 41:704–712.

Seelbach, P. W., and M. J. Wiley. 1997. Overview of the Michigan Rivers Inventory (MRI) project. Michigan Department of Natural Resources, Fisheries Technical Report 97-3, Ann Arbor.

Seelbach, P. W., M. J. Wiley, P. Sorranno, and M. Bremigan. 2002. Aquatic conservation planning: using landscape maps to predict ecological reference conditions for specific water. Pages 454–477 *in* K. J. Gutzwiller, editor. Applying landscape ecology in biological conservation. Springer, New York.

Smale, M. A., and C. F. Rabeni. 1995. Hypoxia and hyperthermia tolerances of headwater stream fishes. Transactions of the American Fisheries Society 124:698–710.

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

Snyder, C. D., J. A. Young, R. Villella, and D. P. Lemarie. 2003. Influences of upland and riparian land use patterns on stream biotic integrity. Landscape Ecology 18:647–664.

Steen, P. J., D. Passino-Reader, and M. J. Wiley. 2006. Modeling brook trout presence and absence from landscape variables using four different analytical methods. Pages 513–531 *in* R. M. Hughes, L. Wang, and P. W. Seelbach, editors. Influence of landscapes on stream habitats and biological assemblages. American Fisheries Society, Symposium 48, Bethesda, Maryland.

Steinberg, D., and P. Colla. 1997. CART—Classification and regression trees. Salford Systems, San Diego, California.

Stoks, R., and M. A. McPeek. 2003. Predators and life histories shape Lestes damselfly assemblages along a freshwater habitat gradient. Ecology 84:1576–1587.

Taverna, K., D. L. Urban, and R. I. McDonald. 2005. Modeling landscape vegetation pattern in response to historic land-use: a hypothesis-driven approach for the North Carolina Piedmont, USA. Landscape Ecology 20:689–702.

Trautman, M. B. 1981. The fishes of Ohio. Ohio State University Press, Columbus.

UM (University of Michigan). 2006. Ecological classification of rivers for environmental assessment. Available: sitemaker.umich.edu/riverclassproject. (September 2007).

USGS (U.S. Geological Survey). 2006. National hydrography dataset. Available: nhd.usgs.gov. (September 2007).

Usio, N., H. Nakajima, R. Kamiyama, I. Wakana, S. Hiruta, and N. Takamura. 2006. Predicting the distribution of invasive crayfish (*Pacifastacus leniusculus*) in a Kusiro Moor marsh (Japan) using classification and regression trees. Ecological Research 21:271–277.

Vanni, M. J. 1987. Effects of nutrients and zooplankton size on the structure of a phytoplankton community. Ecology 68:624–635.

Vanni, M. J., C. D. Layne, and S. E. Arnott. 1997. 'Top-down' trophic interactions in lakes: effects of fish on nutrient dynamics. Ecology 78:1–20.

Vannote, R. L., G. W. Minshall, K. W. Cummins, J. R. Sedell, and C. E. Cushing. 1980. The river continuum concept. Canadian Journal of Fisheries and Aquatic Sciences 37:131–137.

Vaughan, I. P., and S. J. Ormerod. 2003. Improving the quality of distribution models for conservation by addressing shortcomings in the field collection of training data. Conservation Biology 17:1601–1611.

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

Wang, L., J. Lyons, P. Kanehl, and R. Gatti. 1997. Influences of watershed land use on habitat quality and biotic integrity in Wisconsin streams. Fisheries 22(6):6–12.

Wang, L., J. Lyons, P. W. Rasmussen, and P. W. Seelbach. 2003. Watershed, reach, and riparian influences on stream fish assemblages in the Northern Lakes and

Forest Ecoregion, U.S.A. Canadian Journal of Fisheries and Aquatic Sciences 60:491–505.

Wehrly, K. E., M. J. Wiley, and P. W. Seelbach. 2003. Classifying regional variation in thermal regime based on stream fish community patterns. Transactions of the American Fisheries Society 132:18–37.

Wehrly, K. E., M. J. Wiley, and P. W. Seelbach. 2006. Influence of landscape features on summer water temperatures in lower Michigan streams. Pages 113–127 *in* R. M. Hughes, L. Wang, and P. W. Seelbach, editors. Influence of landscapes on stream habitats and biological assemblages. American Fisheries Society, Symposium 48, Bethesda, Maryland.

Wiley, M. J., S. L. Kohler, and P. W. Seelbach. 1997. Reconciling landscape and local views of aquatic communities: lessons from Michigan trout streams. Freshwater Biology 37:133–148.

Wiley, M. J., L. L. Osborne, and R. W. Larimore. 1990. Longitudinal structure of an agricultural prairie river system and its relationship to current stream ecosystem theory. Canadian Journal of Fisheries and Aquatic Sciences 47:373–384.

Zorn, T. G., P. W. Seelbach, and M. J. Wiley. 1998. Patterns in the distributions of stream fishes in Michigan's Lower Peninsula. Michigan Department of Natural Resources, Fisheries Research Report 2035, Ann Arbor.

Zorn, T. G., P. W. Seelbach, and M. J. Wiley. 2002. Distributions of stream fishes and their relationship to stream size and hydrology in Michigan's lower peninsula. Transactions of the American Fisheries Society 131:70–85.

Zorn, T. G., P. W. Seelbach, and M. J. Wiley. 2004. Utility of species-specific, multiple linear regression models for predictions of fish assemblages in rivers of Michigan's Lower Peninsula. Michigan Department of Natural Resources, Fisheries Research Report 2072, Ann Arbor.