Predicting the biological condition of streams: use of geospatial indicators of natural and anthropogenic characteristics of watersheds

**Daren M. Carlisle · James Falcone ·
Michael R. Meador**

Abstract We developed and evaluated empirical models to predict biological condition of wadeable streams in a large portion of the eastern USA, with the ultimate goal of prediction for unsampled basins. Previous work had classified (i.e., altered vs. unaltered) the biological condition of 920 streams based on a biological assessment of macroinvertebrate assemblages. Predictor variables were limited to widely available geospatial data, which included land cover, topography, climate, soils, societal infrastructure, and potential hydrologic modification. We compared the accuracy of predictions of biological condition class based on models with continuous and binary responses. We also evaluated the relative importance of specific groups and individual predictor variables, as well as the relationships between the most important predictors and biological condition. Prediction accuracy and the relative importance of predictor variables were different for two subregions for which models were created. Predictive accuracy in the highlands region improved by including predictors that represented both natural and human activities. Riparian land cover and road-stream intersections were the most important predictors. In contrast, predictive accuracy in the lowlands region was best for models limited to predictors representing natural factors, including basin topography and soil properties. Partial dependence plots revealed complex and nonlinear relationships between specific predictors and the probability of biological alteration. We demonstrate a potential application of the model by predicting biological condition in 552 unsampled basins across an ecoregion in southeastern Wisconsin (USA). Estimates of the likelihood of biological condition of unsampled streams could be a valuable tool for screening large numbers of basins to focus targeted monitoring of potentially unaltered or altered stream segments.

Keywords Biological assessment · Predictive models · Random forests · Geospatial data · Spatial scale

D. M. Carlisle (✉) · J. Falcone · M. R. Meador
National Water-Quality Assessment Program,
US Geological Survey,
Reston, VA 20192, USA
e-mail: dcarlisle@usgs.gov

Introduction

Information about the ecological quality of rivers and streams is critical to understanding the state of our environment (Heinz Center 2002). Statistical sampling methods provide unbiased estimates of the proportion of stream miles within a population (e.g., wadeable streams) in different states of ecological condition (Paulsen et al. 1998). These estimates address questions about the extent and severity of ecological degradation, which are often the goal of regional (e.g., USEPA 2000) and national (e.g., USEPA 2006a) assessments. With the knowledge that a portion of streams are ecologically impaired,

resource managers often need to prioritize which unsampled waters require targeted monitoring and assessment, and this requires tools that predict impairment using readily measured (e.g., map-based) predictors. Empirical models that use geospatial variables to spatially predict (i.e., predict beyond sampled systems) the biological quality (sensu Wright et al. 2000) of stream ecosystems may be such a tool.

The literature linking landscape attributes to stream ecosystems is voluminous, but few have attempted to predict biological quality. The importance of landscape attributes to stream ecosystems is well known (review by Allan 2004), due in large part to empirical models (e.g., multiple linear regression) of associations between landscape characteristics and biological communities (e.g., Strayer et al. 2003; King et al. 2005). Few studies, however, have extended empirical models to prediction (Nilsson et al. 2003), particularly for independent datasets. Although select biological indicators have been projected using geospatial variables (Potter et al. 2004; Van Sickle et al. 2004), there have been few attempts to predict overall biological condition (but see Vólstad et al. 2003), which normalizes the biological response to a deviation from an expected condition (Wright et al. 2000). Such biological endpoints can be widely applied and interpreted (Davies and Jackson 2006; Hawkins 2006) and are therefore desirable for modeling across large spatial scales.

Several statistical challenges complicate empirical modeling of biological condition. Geographic information systems can produce scores of geospatial variables, many of which exhibit highly skewed distributions and complex covariance structures (King et al. 2005). In addition, the desire to model interactions and nonlinear effects, which are theoretically important (Strayer et al. 2003), leads to a dizzying number of potential predictor variables. Analyses are therefore encumbered by statistical pitfalls in variable selection (Hastie et al. 2001; Venables and Ripley 2002) and substantial uncertainty associated with model selection (Burnham and Anderson 2002; Van Sickle et al. 2004). Because these statistical challenges are difficult to overcome with traditional linear modeling, alternative approaches may be more productive.

Breiman (2001) introduced "random forests" (hereafter RF) as an improvement on classification and regression trees. Tree-based methods have proved a useful aid in decision making and data analysis (Venables and Ripley 2002) because they efficiently handle messy (i.e., skewed distributions, nonlinearities) data and high-order interactions (De'ath and Fabricius 2000; Urban 2002; Parmenter et al. 2003). By combining the robustness of tree-based methods and the strengths of model averaging (Breiman 2001; Hastie et al. 2001; Burnham and Anderson 2002), RF appear to be more accurate than linear models, classification trees, general additive models, and artificial neural networks (Svetnik et al. 2003; Garzón et al. 2006; Lawler et al. 2006; Cutler et al. 2007). RF may therefore be capable of predicting stream biological condition using basin-scale geospatial variables that are easily produced. The objective of this paper was to develop and evaluate an empirical model that uses geospatial variables to predict stream biological condition across a large geographic area. We also evaluated the relative importance of predictors and the specific relationships between the most important predictors and biological condition. Finally, we illustrate how such models might be used to screen large numbers of basins for possible biological alteration.

Materials and methods

Study area and data collection

We used biological data collected in 24 major river basins by the US Geological Survey, National Water-Quality Assessment (NAWQA) Program from 1993–2004. This analysis was limited to data from streams in 920 small- to moderate-sized (<1,000 km^2) basins east of the 100th meridian (Fig. 1). Within most major river basins, wadeable streams were targeted for sampling based on prominent land uses (urban, agriculture, undeveloped) within their individual watersheds (Gilliom et al. 1995). Additional streams in six of the major river basins were targeted along gradients of urban land use (Tate et al. 2005). Benthic macroinvertebrates were collected using consistent methods (Cuffney et al. 1993; Moulton et al. 2002). When present, riffles were targeted for macroinvertebrate sampling. Five discrete collections (each 0.25 m^2) were made using a 425–500 μm mesh net and combined into a composite sample. In low-gradient streams with fine-grained substrates, woody snags along the margins or within the channel were
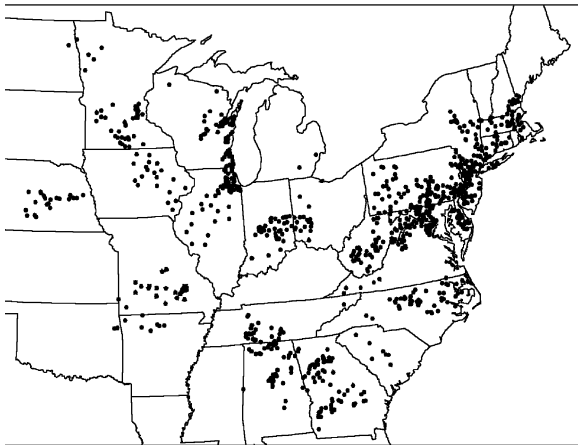
Fig. 1 Locations of 920 sites used for empirical modeling of biological condition

sampled by making collections of least two snags from five separate locations. All invertebrate samples were processed at the USGS National Water Quality Laboratory in Denver, Colorado, using the methods of Moulton et al. (2000). Taxa were identified to the lowest practical taxonomic unit and quality-assurance procedures were maintained (Moulton et al. 2000).

Assessing biological condition

The models described in this study were developed to predict an indicator of biological condition (sensu Wright et al. 2000), which was itself derived from an independent modeling effort. We previously assessed biological condition using an indicator of taxonomic completeness (sensu Hawkins 2006) of macroinvertebrate assemblages primarily at the genus-level of taxonomic resolution (Carlisle and Meador 2007). Taxonomic completeness is defined as the ratio of observed (O) and expected (E) assemblage composition, where O/E <1 indicates that expected taxa were not observed in a sample. Although O is measured directly, E is estimated using empirical models that relate taxonomic composition to natural environmental gradients (Clark et al. 2003). Details of model philosophy (Wright et al. 2000; Hawkins 2006) and construction (Wright et al. 2000; Hawkins et al. 2000; Hawkins and Carlisle 2001; Clark et al. 2003) have been exhaustively described elsewhere. We developed (Carlisle and Meador 2007) a RIVPACS-type (sensu Hawkins et al. 2000) predictive model to estimate E for NAWQA sites within 24 major river basins in the eastern US. Eight predictors (latitude, longitude,

drainage area, mean annual temperature, mean annual precipitation, soil clay content, soil silt content, and sampling day of year) produced estimates of E with lowest bias (mean O/E=0.97) and highest precision (SD of O/E=0.11) when applied to an independent set of least-disturbed sites (Carlisle and Meador 2007). A RIVPACS-type model with these predictors was therefore applied to all non-reference sites for the estimation of E, from which O/E was derived. All of the reference sites used to develop the predictive model and most (>90%) of the test sites were also used in this study.

Although O/E is a continuous measure, our objective was to predict whether or not biological alteration was evident because it approximates the classifications used by many monitoring entities (i.e., supporting vs. not-supporting designated beneficial uses). We defined a threshold of biological alteration as O/E <0.80, which we believed balances statistical and biological significance. We selected this threshold because it was the 10th percentile of O/E values among reference sites (Carlisle and Meador 2007), and therefore approximated the degree of change that was most likely to be statistically detectable while minimizing the misclassification rate of reference sites (i.e., 10%). We also judged that a 20% loss of expected taxa was biologically sufficient to classify a site as altered. Others have also used this or similar thresholds for classifying biological condition (Clark et al. 1996; Ostermiller and Hawkins 2004; Van Sickle et al. 2005).

Although biological condition of all sites was assessed with a single RIVPACS-type model, for this study we stratified our modeling effort into two geographic sub-regions of the eastern USA. We reasoned that models for increasingly homogenous (with respect to major landforms and climate) regions would be more successful at predicting biological condition using basin-scale indicators of anthropogenic stressors. We stratified the eastern USA into one region that included the Appalachian and northeastern highlands (n=281 unaltered and 74 altered sites), and a second region that included the northern/central plains and southeastern lowlands (n=374 unaltered and 191 altered sites). This regionalization is coarse, but represents a balance between increasing spatial resolution (i.e., homogenization of landscape features) and retaining sufficient numbers of observations for modeling purposes, and has been used in other large-scale assessments (USEPA 2006a).

Geospatial data and predictor selection

One hundred fifty-one variables derived from Geographic Information System (GIS) processing were screened as potential predictors of biological condition (Appendix). Geospatial predictors represented broad factors influenced by human activities such as land cover, infrastructure, pesticide and nutrient application, water impoundment, flow diversion, and point source pollution. We also included predictors representing largely natural factors such as climate, topography, hydrology, and soils.

Basin boundaries were acquired as part of the USGS NAWQA Program (USGS 2006a). Land-cover data were derived from a NAWQA-enhanced version of the National Land Cover Dataset 1992 (USGS 2006b), a 21 class, 30 m resolution raster dataset derived from Landsat imagery of the period 1990–1992. Land-cover data were calculated as both basin percentages and riparian zone percentages. Riparian zones were defined as the buffer area 100 m from the stream centerline for all streams in the basin, based on National Hydrography Data (NHD) 100k streams (USGS 2006c). Population density was derived from Census 1990 data (GeoLytics 2001). Road density and road-stream intersections were based on Census 2000 roads (GeoLytics 2001) and NHD 100k streams. Pesticide and nutrient application were derived from coverages developed by the NAWQA Program (Ruddy et al. 2006; USGS 2007). Water impoundment data were obtained from the National Inventory of Dams (USACE 2006). Flow diversion data were obtained from NHDPlus (Horizon Data Systems Corp. 2006). Data for point-source discharges were obtained from USEPA National Pollutant Discharge Elimination System (USEPA 2006b). Mean air temperature and precipitation statistics were derived from 1km resolution Daymet model data (Daymet 2006), which represented 18 year (1980–1997) temperature and precipitation averages obtained from terrain-adjusted daily climatological observations. Topographic variables were derived from 100 m data based on the USGS National Elevation Dataset (USGS 2006d). Hydrologic variables were based on 1 km resolution grids developed from USGS sources (Wolock and McCabe 1995; Wolock et al. 1997). Soils data were based on 100 m weighted averages of US Department of Agriculture State Soil Geographic Data (USDA 2006).

We retained 66 of the original variables for use as predictors. Variables that were highly redundant (i.e., Spearman rank $|\rho| > 0.80$) were removed from consideration, including all basin-scale land-cover variables, which were highly correlated with land cover in the riparian buffer. Because many (28) of the available predictors represent natural environmental gradients and were also used in the RIVPACS-type predictive models from which biological condition was assessed (Carlisle and Meador 2007), we evaluated whether the inclusion of these variables improved model predictions (see "Statistical modeling").

Statistical modeling

Random forests (RF) extend classification and regression trees (Prasad et al. 2007; Cutler et al. 2007) by using an ensemble of trees to make predictions (Breiman 2001). RF builds many classification/regression trees, each with a bootstrapped sub-sample of the original observations. In addition, nodes are created by selecting the best predictor variable from a randomly selected subset of all predictors. The largest tree possible is grown but not pruned. The observations not selected in the bootstrap sample are then passed through each tree. This process is repeated, creating a "forest" of individual trees. The final classification (or value in regression) for each observation is made by selecting the class to which it was most frequently classified (average predictions for regression) across all trees where that observation was excluded from the bootstrap sample. Additional details of RF are given elsewhere (Breiman 2001; Prasad et al. 2007; Cutler et al. 2007). We constructed RF models with 2000 trees using an implementation written for the R statistical system (R Development Core Team 2006) by Liaw and Wiener (2002).

We compared the accuracy of modeling biological condition as a categorical and continuous variable. We classified each observation using the threshold described above (O/E=0.80), then used RF in classification mode to predict the class of each observation. We also used RF in regression mode to predict O/E as a continuous variable, and then subsequently classified each observation as altered (predicted O/E<0.80) or unaltered (predicted O/E>0.80). Because the final prediction for each observation is computed over the trees where it was excluded from RF construction, these estimates are equivalent to that of a cross-

validated procedure (Cutler et al. 2007). Prediction accuracy and error rates were evaluated by computing the overall percentage of correctly classified observations (PCC), sensitivity (percentage of altered sites correctly classified), specificity (percentage of unaltered site correctly classified), kappa (measure of agreement between observed and predicted classes corrected for agreement due to chance). We also used the area under the receiver-operating characteristic curve (AUC) to measure the performance of classification models because it provides a single measure of overall accuracy that is independent of thresholds used for binary classification (Fielding and Bell 1997). AUC values range from 0.5, which indicates the classification performed no better than chance, to a maximum of 1.0, which indicates the classification was perfectly accurate. We evaluated the performance of each RF model when constructed with subsets and the entire set of 66 predictor variables. For both the RF regression and classification models, we compared prediction accuracy for models that included: 28 natural covariates as predictors, 38 predictors that represent human activities, and all 66 predictor variables.

Our secondary objective was to evaluate the relative importance of predictor variables for each model. Variable importance was measured with observations not selected in the bootstrap samples used to construct each tree (so-called "out of bag" or OOB). First, the OOB observations are passed through each tree and classified. Then, values of each predictor variable are randomly permuted in turn and the corresponding decrease in prediction accuracy (or mean squared error for regression) of the OOB observations is recorded for each tree. The decrease in accuracy for each predictor is averaged and standardized (divided by the standard error) across all trees (Liaw and Wiener 2002; Cutler et al. 2007). Intuitively, the relative decrease in prediction accuracy when a predictor variable is permuted is related to its importance in the classification.

We examined the relationships between individual predictors and biological condition using partial dependence plots, which is a tool to visualize the effects of a single variable on the outcomes of classification and regression models (Hastie et al. 2001; Cutler et al. 2007). For a given value of the predictor under examination, the expected prediction is estimated by averaging the predictions over all the combinations of observed values of the other predictors in the data set. In essence, partial dependence functions represent the

effects of the examined predictor variable after accounting for the average effects of all other predictors (Hastie et al. 2001). We examined partial dependence for a subset of the most important predictors for each model. Partial dependence was computed for each of 500 equally spaced points over the range of each predictor selected for examination (Liaw and Wiener 2002).

Finally, we illustrate how models that predict biological condition may be applied to unsampled locations as a tool to screen broad regions for the existence of biologically altered streams. Predictor variables were generated for 552 unsampled basins in the Southeastern Wisconsin Till Plains ecoregion that had similar size range (10–150 km^2) as the 56 sampled basins also located in that ecoregion (Fig. 1). Basin boundaries for unsampled sites were derived from 30 m National Elevation Data (USGS 2006d). We first applied the procedures described by Moss et al. (1987) and Clark et al. (2003) to determine if the predictor variable values of each unsampled basin belonged to the same statistical population as the reference sites used to develop the RIVPACS-type model. The Mahalanobis squared distances between the locations of test-site values in predictor space and the centroids of each classification group are distributed as a χ^2 function. Any unsampled basin was regarded as outside the range of the model if the smaller of these values exceeded the critical χ^2 value for $\alpha = 0.01$ and N degrees of freedom, where N equals the number of discriminant functions. We applied the RF classification model to unsampled basins and aggregated the votes (i.e., proportion of OOB instances where each observation was classified as altered). We then mapped the unsampled basins and illustrated the proportion of votes for which each was classified as having altered biological condition.

Results and discussion

The type (i.e., continuous vs. categorical) of biological response variable did not appear to substantially influence prediction accuracy. By most measures, binary predictions from RF classification and a posteriori classifications from RF regression were similarly accurate (Tables 1, 2). Both approaches produced highly accurate (mean sensitivity: classification=91%,

Table 1 Accuracy measures for predictions of biological condition for eastern highlands region

Accuracy measure	Model					
	Classification			Regression		
	Natural	Human	Full	Natural	Human	Full
PCC	83	86	87	85	85	86
Specificity	41	55	51	42	55	51
Sensitivity	94	95	96	96	93	96
Kappa	0.405	0.451	0.542	0.452	0.510	0.535

PCC is the percentage of observations correctly classified. Specificity is the percentage of altered streams correctly classified. Sensitivity is percentage of unaltered streams correctly classified. Kappa is a measure of agreement between the predicted and observed corrected for the agreement due to chance alone

regression=90%) predictions of unaltered sites, less accurate (mean specificity: classification=51%, regression=52%) predictions of altered sites, and shared similar performance relative to random chance (mean kappa: classification=0.441, regression=0.460). Given this similarity, all further results and discussion will be limited to classification models.

Prediction accuracy was modestly influenced by the subset of predictors included in models, but this effect varied by region. In the highlands (Table 1), models that included only human-activity predictors improved model accuracy over that obtained from models limited to natural covariates as predictors. The model with both types of predictors performed best overall, but only by a modest amount (i.e., kappa increased from 0.405 to 0.542). In contrast, the addition of human-activity predictors to a model that included natural covariates did not improve predictions of biological condition in the lowlands region (Table 2). Moreover, the model limited to human-activity predictors performed worse than the model with only natural predictors.

The interactive effect of natural covariates and human-activity variables on predictions of biological condition (i.e., O/E) in the highlands region could arise for at least two reasons. One possibility is that some variation in E due to natural factors remained unaccounted for by the model used to assess biological condition (Carlisle and Meador 2007). This would result in deviations of O from E that vary

systematically by some natural factor (e.g., basin size). In theory, we were able to at least partially "remove" the effect of natural covariates because climate, soil properties, basin size, topography, and geographic area were used in the RIVPACS-type model (Carlisle and Meador 2007) to estimate E. However, RF may more successfully model variability than linear methods under some conditions (Cutler et al. 2007), and linear discriminant analysis underlies RIVPACS-type models (Hawkins et al. 2000). We therefore evaluated in a post hoc fashion whether RF could explain residual variation in O/E that had not been accounted for by the RIVPACS-type model. We constructed an RF-regression model that predicted O/E values at reference sites (n=273) using the 28 natural covariates included in this study. We limited the analysis to reference sites because, in theory, the only variation in E among reference sites (excluding model and sampling error) should be due to natural variability. Observed O/E values from this model explained <10% of the variation in predicted O/E values, which strongly suggests that there was no substantive residual variation in E due to natural factors that was unexplained by the original RIVPACS-type model. Another explanation for the importance of natural covariates in predictive accuracy of highland streams is that environmental settings influence the response (i.e., O/E) of stream invertebrate communities to human activities. Many studies have shown that the vulnerability or ecological

Table 2 Accuracy measures for predictions of biological condition for eastern lowlands region

| Accuracy measure	Model					
	Classification			Regression		
	Natural	Human	Full	Natural	Human	Full
PCC	77	73	76	77	73	76
Specificity	58	43	57	59	47	58
Sensitivity	87	88	87	86	87	85
Kappa	0.460	0.337	0.451	0.461	0.359	0.442

PCC is the percentage of observations correctly classified. Specificity is the percentage of altered streams correctly classified. Sensitivity is percentage of unaltered streams correctly classified. Kappa is a measure of agreement between the predicted and observed corrected for the agreement due to chance alone

response of streams to human activities is influenced by a variety of natural factors (e.g., channel gradient, stream size) (Allan 2004; Yates and Bailey 2006). We think this is the mostly likely explanation for our own observations.

Models were slightly less accurate in the lowlands than in the highlands, and also did not improve with the addition of stressor predictor variables (Tables 1 and 2). One possible explanation is due to the pervasiveness of human activity in the lowlands

region. Relative to the highlands region, reference sites used to assess biological condition in the lowlands were from basins experiencing high levels of human activity. Reference sites in the lowlands were considered to be the least disturbed sites within the region, but were often selected based on expert knowledge that local-scale conditions (e.g., habitat) was sufficient to support relatively intact biological communities (Carlisle and Meador 2007). As a consequence, it is likely that our basin-scale stressor

Fig. 2 Top ranked variables from random forests classification for predicting biological condition in highlands of the eastern USA. Mean decrease in accuracy for a variable is the normalized difference of the classification accuracy for the observations excluded from model calibration, and the classification accuracy for the same observations when values of the predictor are randomly permuted. Higher values of the mean decrease in accuracy indicate a predictor is more important to the classification
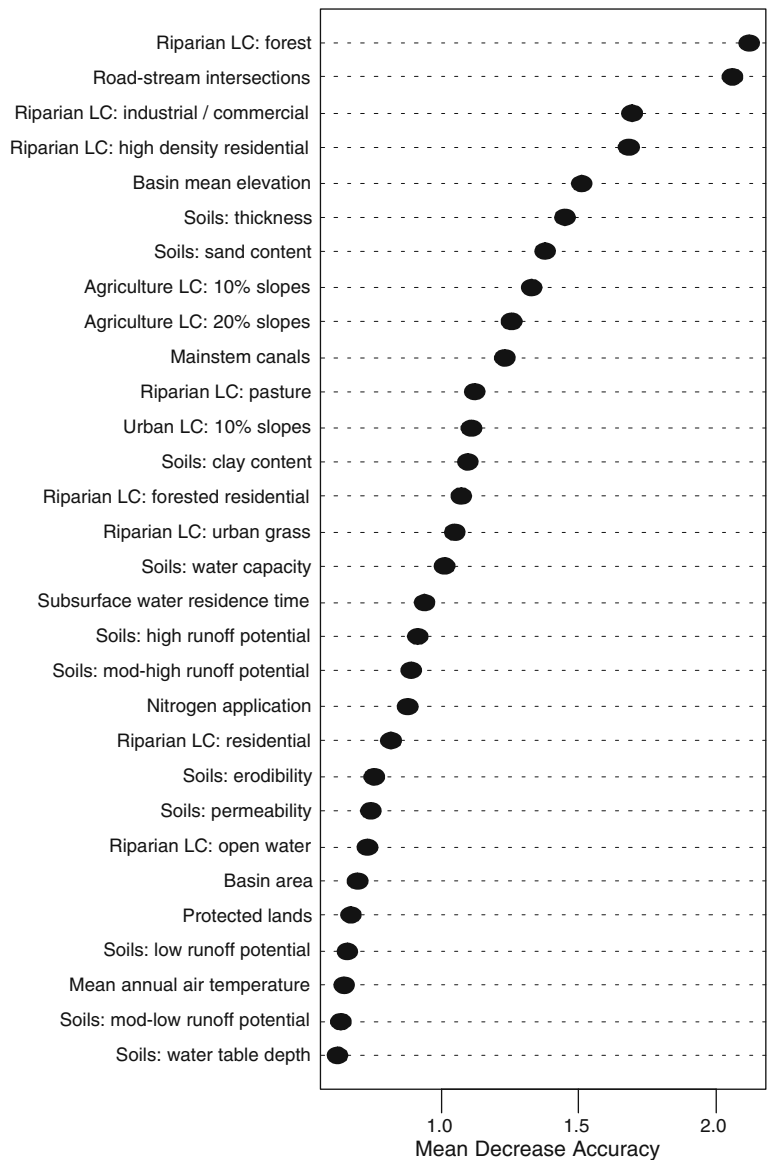
predictors did not vary substantively between lowland basins with altered and unaltered biological communities. Instead, natural factors that probably covary with unmeasured stressors were sufficient to discriminate between altered and unaltered sites.

The best predictors of biological condition differed between the lowlands and highlands (Fig. 2). In the highlands, predictor variables with the greatest effect were, in decreasing order of importance, riparian forest cover, road-stream intersections, and riparian urban land cover, followed by basin elevation and soil properties. In contrast, soil properties and basin elevation were the most important predictors in the lowlands, followed by riparian urban land cover and several additional natural covariates related to soils and hydrologic properties (Fig. 3). Our results for the

Fig. 3 Top ranked variables from random forests classification for predicting biological condition in lowlands of the eastern USA. Mean decrease in accuracy for a variable is the normalized difference of the classification accuracy for the observations excluded from model calibration, and the classification accuracy for the same observations when values of the predictor are randomly permuted. Higher values of the mean decrease in accuracy indicate a predictor is more important to the classification
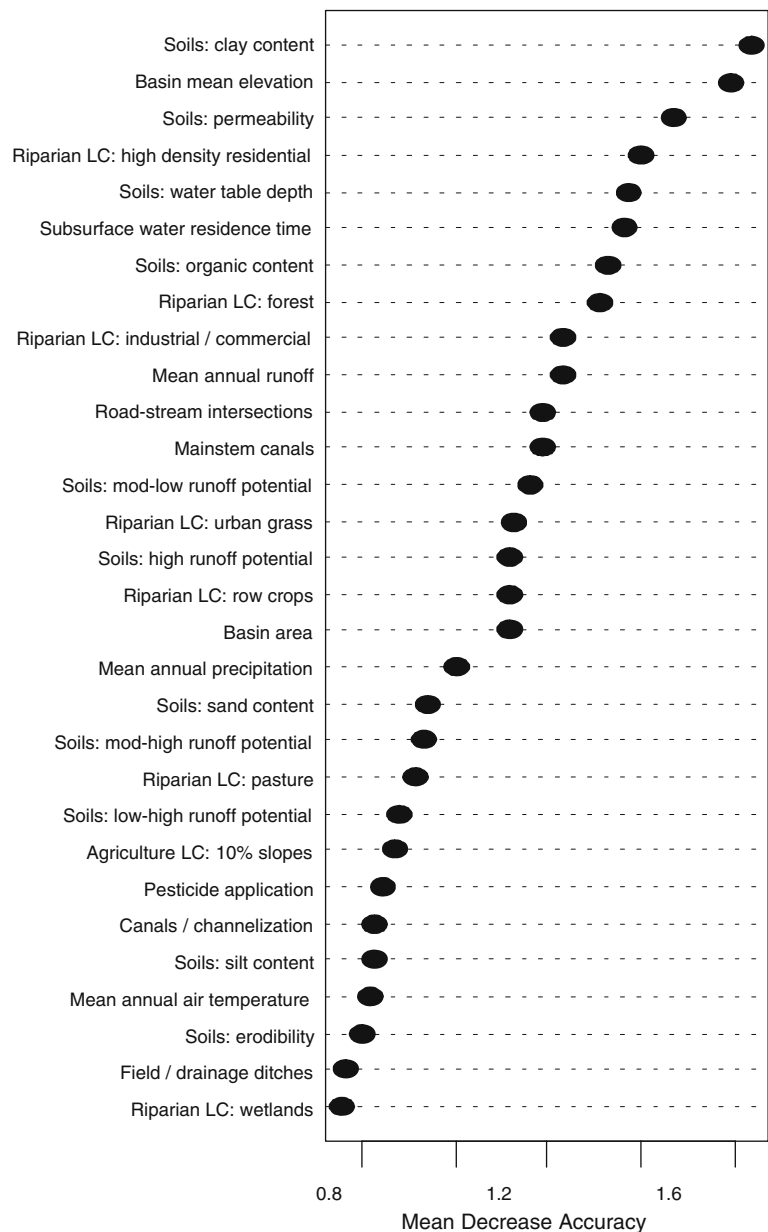
highlands support many studies that showed riparian forest cover is an important predictor of biological communities and influences stream ecosystems in natural (Wallace et al. 1997), agricultural (Allan 2004; Moore and Palmer 2005), and urbanized basins (Paul and Meyer 2001; Allan 2004; Moore and Palmer 2005; Walsh et al. 2005). In contrast, our results for the lowlands suggest that the stressor predictors were

not useful for discriminating between biologically altered and unaltered sites (Table 2), for reasons described above (i.e., poor quality reference sites) or because the factors responsible for biological alteration covaried with basin soil properties. Nevertheless, urbanized riparian land cover appeared to be more important than any indicator of agricultural use intensity (e.g., riparian land cover in row crops) in the

Fig. 4 Partial dependence plots for selected predictors for random forests predictions of the presence of biological alteration in highlands streams of the eastern USA. Partial dependence is the dependence of the probability that a stream was classified as altered based on a single predictor variable after averaging out the effects of all other predictors in the model
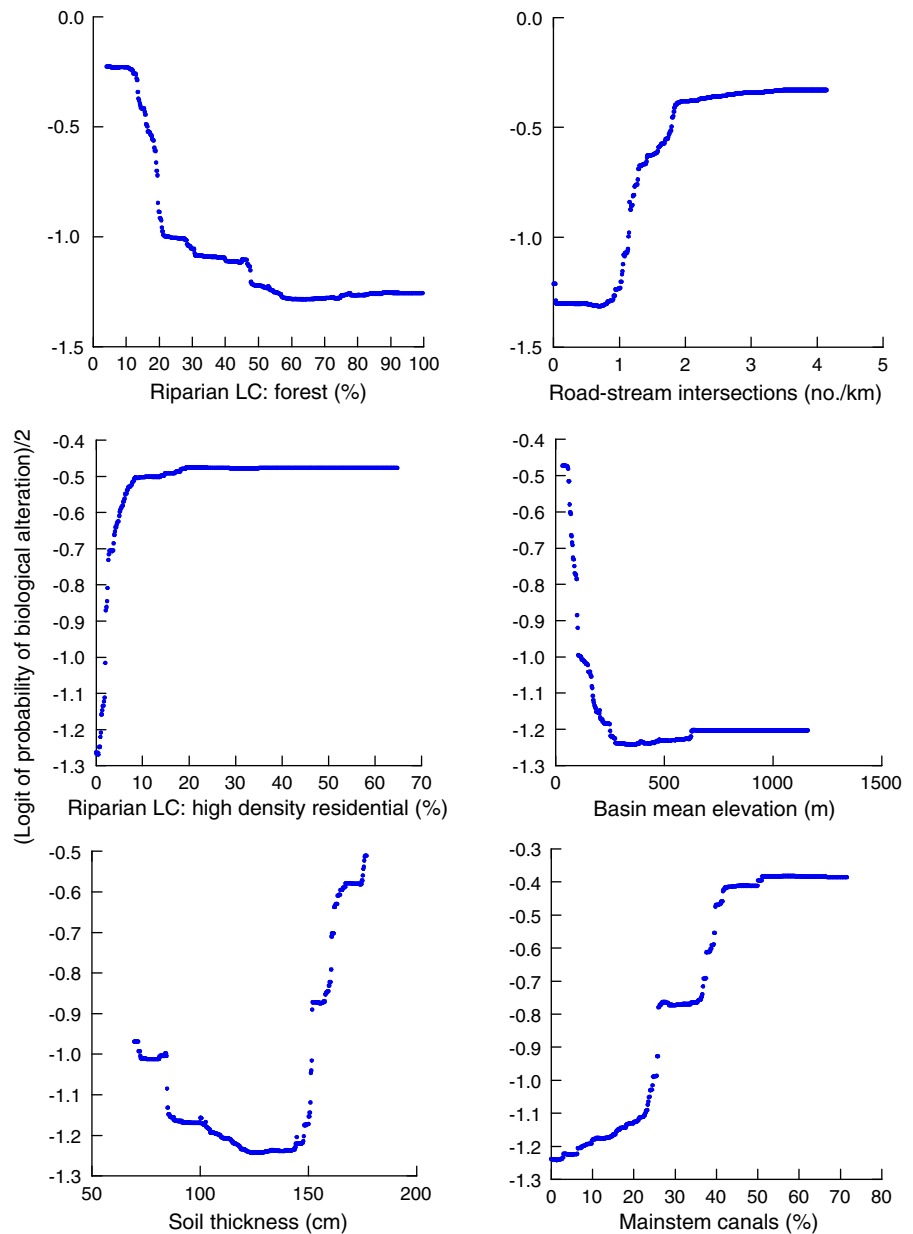
Fig. 5 Partial dependence plots for selected predictors for random forests predictions of the presence of biological alteration in lowlands streams of the eastern USA. Partial dependence is the dependence of the probability that a stream was classified as altered based on a single predictor variable after averaging out the effects of all other predictors in the model
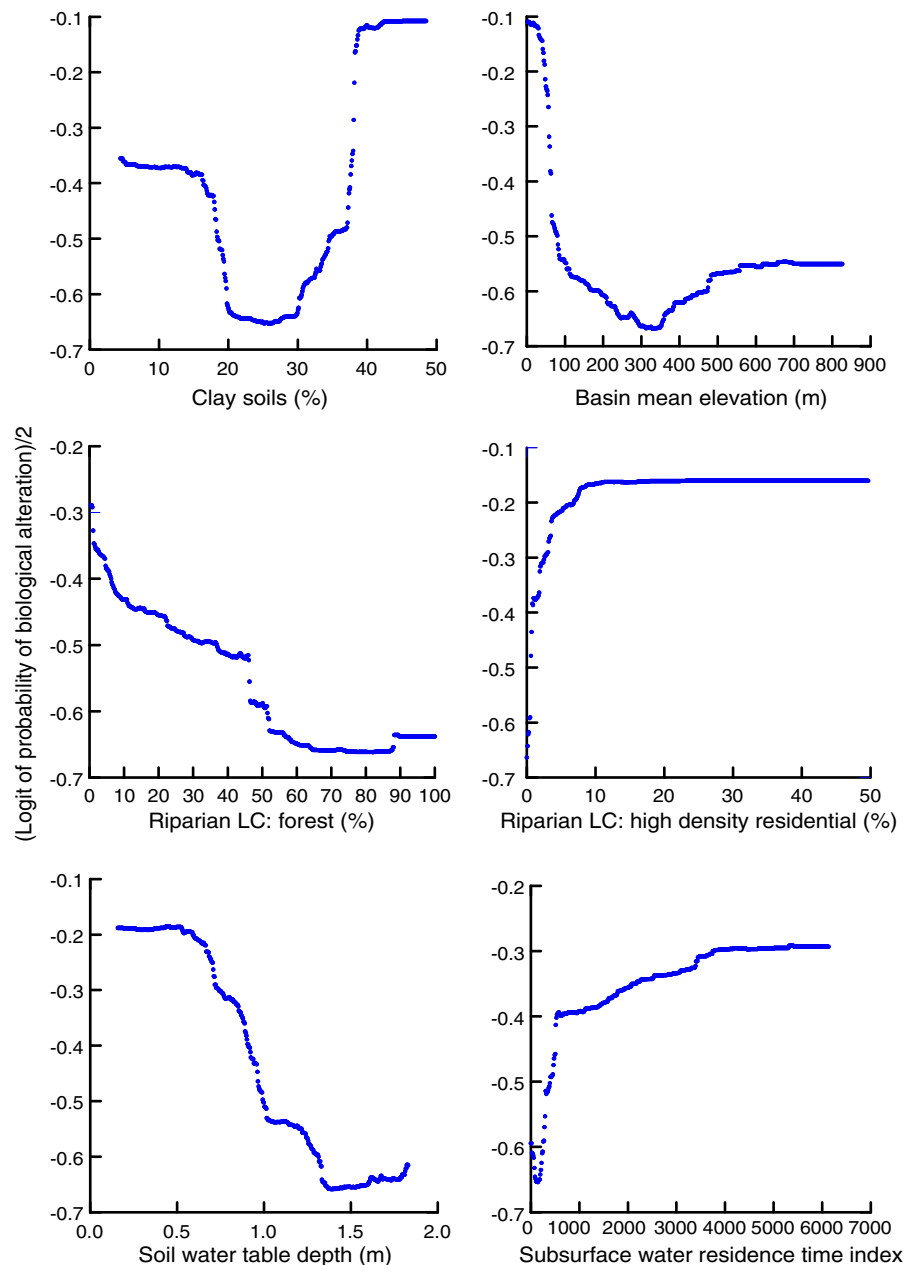
lowlands despite the pervasiveness of agricultural activity in this region. This observation can be attributed, in part, to the overwhelming influence of urbanization on stream biota in landscapes that have previously been disturbed by significant agricultural activity (Moore and Palmer 2005).

Partial dependence plots revealed the presence of potential thresholds in relationships between biological alteration and indicators of disturbance (Figs. 4 and 5). After averaging the effects of all other predictor variables, the probability that a stream was biologically altered increased dramatically as urban

land cover in the riparian zone reached 10%, then changed little thereafter. There was also an apparent threshold of road density in the highlands region, where abrupt increases in the probability of biological alteration occurred between one to two crossings per kilometers of stream length. Thresholds of impervious or urban land cover related to biological responses have been widely reported and debated (Paul and Meyer 2001; Walsh et al. 2005). The apparent threshold that we observed in both regions is evidence that significant biological degradation is associated with minimal urbanization in a variety of natural settings, but our methods can neither be used to identify specific threshold values and associated uncertainty, nor isolate the actual causal factors that underlie this phenomenon (Paul and Meyer 2001). The probability of biological alteration in both regions tended to increase as riparian forest land cover decreased, but only below values of about 60%. The probability of biological alteration was unrelated to riparian forests above values >60% after averaging the effects of all other predictors, which suggests that significant buffers ameliorate the effects of other human activities on the landscape (Roy et al. 2006; Carline and Walsh 2007).

Biological condition appeared to be associated with some natural factors in complex ways. Steep declines in the probability of alteration were observed in both regions as mean basin elevation increased to approximately 200 m, followed by relatively modest or no changes in probability at elevations >500 m. This result may have emerged from the tendency of watersheds with more topographic relief (and presumably higher mean elevation) to have generally less landscape alteration than watersheds in valleys and plains. The effect of soil properties on the probability of biological alteration in both regions was "U"-shaped, in that biological alteration decreased from low to moderate values of soil thickness and percent clays, but then increased precipitously from moderate to high values. Interactions of basin soil characteristics and other environmental factors in structuring stream biological communities have been reported (Richards et al. 1993), including the influence of basin soil characteristics on reach-scale stream habitat (Davies et al. 2000). Soil characteristics may also be surrogates for basin geology, which is known to influence stream

biota through effects on geochemistry and stream habitat (Pyne et al. 2007).

RF classification models generate model-averaged estimates of the evidence that each observation belongs to each class, which have potential use in decision making. Every observation is classified in RF models by aggregating predictions across trees where it was excluded from model development. The class in which the observation was most frequently classified is selected, with tie "votes" (i.e., proportion of trees in which the observation was classified into each category) broken randomly. The proportion of trees in which each observation was classified as altered or unaltered is conceptually similar to evidence obtained by model averaging (sensu Burnham and Anderson 2002), and therefore more informative than the simple binary prediction produced by the model. The proportions convey information about uncertainty in the sampling, assessment, and predictive modeling processes, and therefore indicate the level of confidence in each prediction. We would likely have more confidence in the prediction for a basin that was classified as altered in 90% of trees (0.90 proportion) than one classified as such in only 50% of trees. This quantification of uncertainty in a simple and understandable way allows decision making in the face of imperfect information because predictions and priorities can be weighted accordingly. For example, sites classified with the highest certainty (e.g., >90% or <10% of votes for alteration) could reliably be considered as altered or unaltered and therefore selected for targeted monitoring. This simple approach could be used to screen a large population of sites across a large geographic area in order to prioritize targeted monitoring and assessment efforts.

We illustrate how the votes for biological alteration for unsampled Wisconsin basins can be applied in decision making (Fig. 6). Chi-squared tests revealed that environmental conditions in >98% of the 552 unsampled basins were similar to those of the reference sites used to assess biological condition. Figure 6 illustrates the proportion of votes across 2,000 classification trees that the stream at each basin terminus was classified as biologically altered (i.e., loss of >20% of the expected taxa). Most basins with a high proportion of votes were within or adjacent to
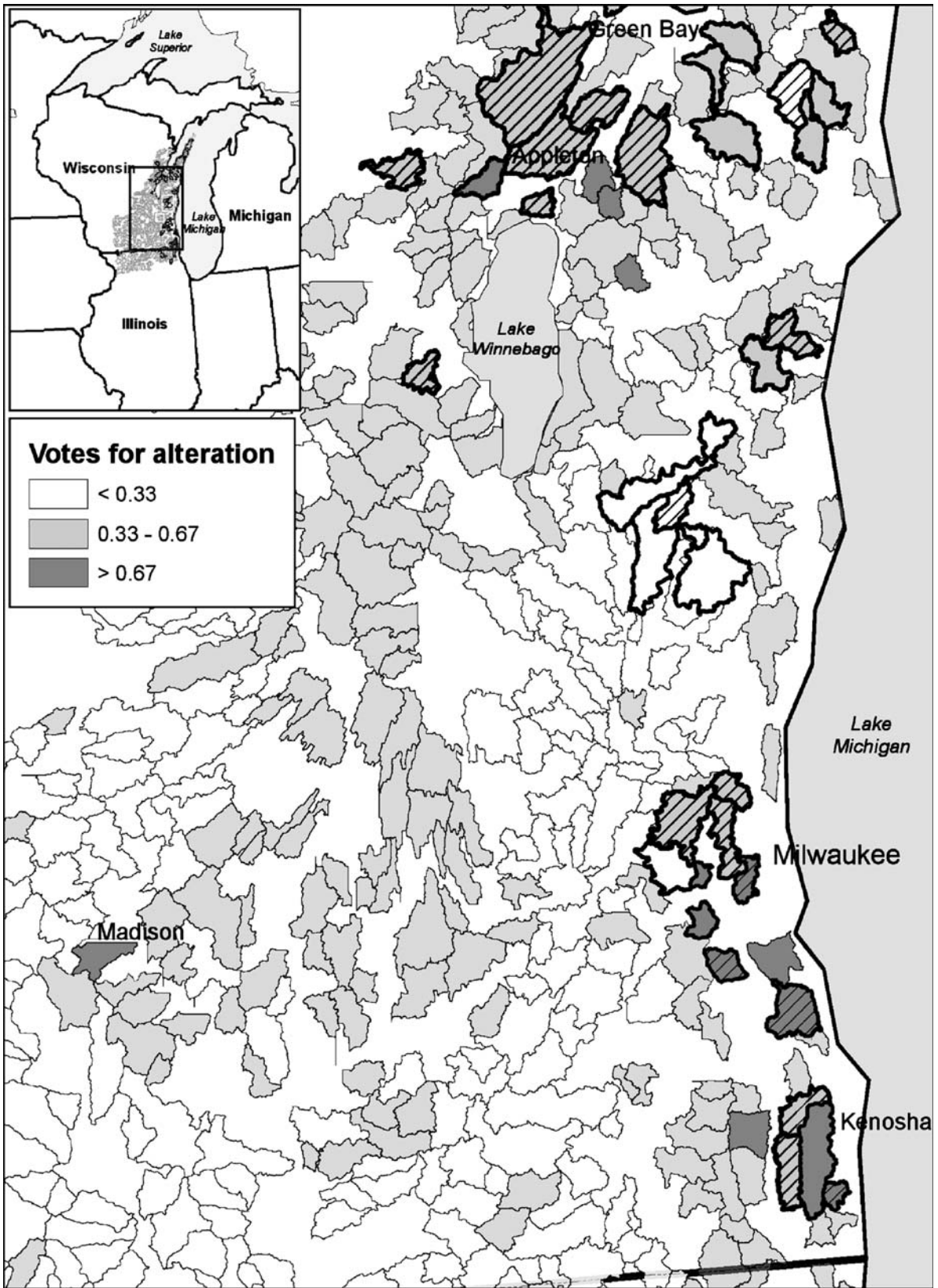
urban areas. Several basins outside urban centers, however, were also likely to be biologically altered, which suggests that other anthropogenic activities may be the cause of degradation in these areas. There was reasonably good agreement between the model's predicted evidence for biological alteration (i.e., votes) and the biological condition of streams actually sampled in Wisconsin (Fig. 6). In general, sampling revealed biological alteration in basins with a moderate to high modeled potential for alteration, and sampling revealed unaltered biological condition in basins with low modeled potential for alteration. However, the model failed to detect biological alteration in two (6%) basins, where the votes for alteration were low (<0.33) but actual sampling revealed otherwise. Also, the model predicted a high probability of alteration in four (12%) basins where sampling revealed biologically unaltered conditions.

We offer several considerations about the broad application of this or similar models. First, on an absolute scale, RF model predictions were an improvement over random classifications using prior probabilities of group membership, as evidenced by AUC values for the highlands (0.86) and lowlands (0.82). These values indicate that in 86 or 82% of trials, the predicted probability of being altered will be greater at a randomly-chosen altered site than at a randomly-chosen unaltered site. We therefore argue that the application of such a model improves the information available to managers, especially if model-averaged votes are retained and considered in decision making. Second, we emphasize that the data used to construct the model were from targeted samples so their representativeness of some larger population of eastern US streams is unknown. Nonetheless, we believe that the modeling effort benefited from targeted sampling designed to represent a large gradient of human disturbance. Third, our definition of biological alteration is limited to this study, largely because currently there are not consistent definitions among jurisdictions that conduct bioassessments (USGAO 2002; Davies and Jackson 2006). Our measure also contained

inherent error (as do all approaches used in bio-assessments), and surely increased the error with which condition could be modeled in the present study. Nevertheless, we believe that our definition is biologically interpretable (>20% loss of native taxa relative to reference conditions) and therefore broadly applicable. Finally, we do not advocate replacing biological sampling with models or using model predictions for management decisions. Rather, we emphasize that empirical models like those we produce can be useful for making projections as a tool to screen large numbers of basins for targeted monitoring aimed at selecting reference-quality or severely degraded sites.

Conclusions

In the absence of biological data, models may provide sufficient evidence for informed predictions about stream ecological conditions, and are therefore useful for screening large numbers of potential basins and prioritizing resources for targeted sampling and assessment. Our results demonstrate that the occurrence of biologically impoverished streams can be predicted with reasonable accuracy using widely available geospatial variables. Machine learning methods such as random forests were developed to extract information from large amounts of data, and are potentially powerful predictive tools for addressing objectives related to environmental monitoring and decision making. The model produced site-specific, model-averaged estimates that approximate the degree of certainty in predictions, which can be integrated in the decision-making process. The generality of the present model to other eastern US streams is not certain, but evidence suggests that it has broad applicability within certain environmental limits.

Appendix

Table 3 Geospatial variables considered and used to predict biological condition in 920 wadeable streams in the eastern USA

Variable	Source (units)	Variable	Source (units)
Latitude of sampling site[a]	USGS 2006a (degrees)	Land cover: urban grass[a]	USGS 2006b (%)
Longitude of sampling site[a]	USGS 2006a (degrees)	Land cover: woody herbaceous wetlands[a]	USGS 2006b (%)
Drainage area[a]	USGS 2006a (km²)	Sum of urban classes	USGS 2006b (%)
Elevation at sampling site[a]	USGS 2006d (meters)	Sum of agriculture classes	USGS 2006b (%)
1990 Census population	GeoLytics 2001 (no./km²)	Mean annual precipitation[a]	Daymet 2006 (cm)
Land cover: open water[a,b]	USGS 2006b (%)	Mean annual temperature[a]	Daymet 2006 (degrees)
Land cover: low intensity residential[a]	USGS 2006b (%)	Mean monthly precipitation: January–December[a]	Daymet 2006 (cm)
Land cover: high intensity residential[a]	USGS 2006b (%)	Mean monthly temperature: January–December[a]	Daymet 2006 (degrees)
Land cover: commercial/industrial[a]	USGS 2006b (%)	Road density	GeoLytics 2001 (km/km²)
Land cover: forested residential[a]	USGS 2006b (%)	Stream density[a]	GeoLytics 2001 (km/km²)
Land cover: bare[a]	USGS 2006b (%)	Road/stream intersections[a]	GeoLytics 2001 (no./km²)
Land cover: quarry/mining[a]	USGS 2006b (%)	Average runoff 1971–2000[a]	Wolock and MacCabe 1995 (mm)
Land cover: transitional[a]	USGS 2006b (%)	Percent base flow of total flow[a]	Wolock and MacCabe 1995 (%)
Land cover: deciduous evergreen mixed forest[a]	USGS 2006b (%)	Percent Dunne overland flow[a]	Wolock and MacCabe 1995 (%)
Land cover: shrubland[a]	USGS 2006b (%)	Percent Horton overland flow[a]	Wolock and MacCabe 1995 (%)
Land cover: orchards[a]	USGS 2006b (%)	Topographic wetness index[a]	Wolock and MacCabe 1995 (%)
Land cover: grasslands[a]	USGS 2006b (%)	Ground water residence time index[a]	Wolock and MacCabe 1995 (%)
Land cover: pasture/hay[a]	USGS 2006b (%)	Mean basin elevation[a]	USGS 2006d (meters)
Land cover: row crops[a]	USGS 2006d (%)	Mean basin slope	USGS 2006d (%)
Mean basin east aspect[a]	USGS 2006d (%)	Coarse soils (5 mm sieve) by weight[a]	USDA 2006 (%)
Mean basin north aspect[a]	USGS 2006d (%)	Fine soils (0.074 mm sieve) by weight[a]	USDA 2006 (%)
Index of basin shape compactness[a]	USGS 2006d (unitless)	Medium soils (2 mm sieve) by weight[a]	USDA 2006 (%)
Soils in STATSGO hydrologic Group A[a]	USDA 2006 (%)	Soils – percent silt[a]	USDA 2006 (%)
Soils in STATSGO Hydrologic Group B[a]	USDA 2006 (%)	Soils – percent sand[a]	USDA 2006 (%)
Soils in STATSGO Hydrologic Group C[a]	USDA 2006 (%)	Soil erodibility factor (USL equation)[a]	USDA 2006 (%)
Soils in STATSGO Hydrologic Group D[a]	USDA 2006 (%)	Soil rainfall/runoff factor (USL equation)[a]	USDA 2006 (%)
Soils in STATSGO Hydrologic Group AD[a]	USDA 2006 (%)		

Table 3 (continued)

Variable	Source (units)	Variable	Source (units)
Soils in STATSGO Hydrologic Group BD[a]	USDA 2006 (%)	Protected lands, most protected	USGS 2006b (%)
Soils in STATSGO Hydrologic Group CD[a]	USDA 2006 (%)	Protected lands, protected	USGS 2006b (%)
Soil permeability[a]	USDA 2006 (%)	Protected lands, somewhat protected	USGS 2006b (%)
Soil available water capacity[a]	USDA 2006 (%)	Mainstem stream classified as "Canal", "Ditch", "Pipeline" or "Artificial" in NHDPlus[a]	Horizon Data Systems Corp. 2006 (%)
[a]Soil bulk density	USDA 2006 (%)	Dams in the mainstem buffer[a]	USACE, 2006 (no./km)
Soil organic matter[a]	USDA 2006 (%)	Streams classified as "Canal", "Ditch", or "Pipeline" in NHDPlus[a]	Horizon Data Systems Corp. 2006 (%)
Soil depth to water table[a]	USDA 2006 (%)	Streams classified as "Artificial" in NHDPlus[a]	Horizon Data Systems Corp. 2006 (%)
Soil rock depth[a]	USDA 2006 (%)	NPDES "major" point locations in basin	USEPA 2006b (no./km^2)
Soils – percent clay[a]	USDA 2006 (%)	Urban (agriculture) land cover on slopes ≥5%[a]	USGS 2006a (%)
Straightline distance of sampling site to nearest major NPDES point in basin[a]	USEPA 2006b (m)	Urban (agriculture) land cover on slopes ≥10%[a]	USGS 2006a (%)
Land area using surface drainage, field ditches (NRI code 607)[a]	USGS 2006b (%)	Urban (agriculture) land cover on slopes ≥20%[a]	USGS 2006a (%)
Straightline distance of sampling site to nearest major dam in basin[a]	USACE 2006 (m)	Urban (agriculture) land cover on slopes ≥30%[a]	USGS 2006a (%)
Average straightline distance of sampling site to all major dams in basin[a]	USACE 2006 (m)	Total nitrogen and phosphorus application[a]	Ruddy et al. 2006 (kg/km^2)
		Total pesticide application[a]	USGS 2007 (kg/km^2)

[a] Geospatial variables considered and used

[b] Land cover classes were computed for entire basins and for riparian corridor

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.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer.

Carline, R. F., & Walsh, M. C. (2007). Responses to riparian restoration in the spring creek watershed, central Pennsylvania. *Restoration Ecology*, 15, 731–742.

Carlisle, D. M., & Meador, M. R. (2007). A predictive model for the biological condition of macroinvertebrate assemblages in eastern U.S. streams. *Journal of the American Water Resources Association*, 43, 1194–1207.

Clark, R. T., Furse, M. T., Wright, J. F., & Moss, D. (1996). Derivation of a biological quality index for river sites: comparison of the observed with the expected fauna. *Journal of Applied Statistics*, 23, 311–332.

Clark, R. T., Wright, J. F., & Furse, M. T. (2003). RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. *Ecological Modelling*, 160, 219–233.

Cuffney, T. F., Gurtz, M. E., & Meador, M. R. (1993). Methods for collecting benthic macroinvertebrate samples as part of the National Water-Quality Assessment Program. Open File Report 93-406, US Geological Survey.

Cutler, D. R., Edwards, T. C., Jr., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., et al. (2007). Random forests for classification in ecology. *Ecology*, 88, 2783–2792.

Davies, S. P., & Jackson, S. K. (2006). The Biological Condition Gradient: a conceptual model for interpreting detrimental change in aquatic ecosystems. *Ecological Applications*, 16, 1251–1266.

Davies, N. M., Norris, R. H., & Thoms, M. C. (2000). Predication and assessment of local stream habitat features using large-scale catchment characteristics. *Freshwater Biology*, 45, 343–369.

Daymet 2006. Numerical Terradynamic Simulation Group: University of Montana. Retrieved from www.daymet.org.

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

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

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

GeoLytics (2001). CensusCD 2000 and StreetCD 2000 CDROM. GeoLytics, Inc., East Brunswick: New Jersey.

Gilliom, R. J., Alley, W. A., & Gurtz, M. E. (1995). Design of the National Water-Quality Assessment Program: Occurrence and distribution of water-quality conditions. US Geological Survey Circular 1112, Sacramento, California.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer.

Hawkins, C. P. (2006). Quantifying biological integrity by taxonomic completeness: its utility in regional and global assessments. *Ecological Applications*, 16, 1277–1294.

Hawkins, C. P., & Carlisle, D. M. (2001). Use of predictive models for assessing the biological integrity of wetlands and other aquatic habitats. In R. B. Rader, D. P. Batzer, & S. A. Wissinger (Eds.), *Bioassessment and management of North American wetlands* (pp. 59–83). New York: Wiley.

Hawkins, C. P., Norris, R. H., Hogue, J. N., & Feminella, J. W. (2000). Development and evaluation of predictive models for measuring the biological integrity of streams. *Ecological Applications*, 10, 1456–1477.

Heinz Center (2002). The state of the nation's ecosystems: Measuring the lands, waters, and living resources of the United States. The H. John Heinz III Center for Science, Economics and the Environment, 1001 Pennsylvania Ave, NW Suite 735 South, Washington, DC.

Horizon Systems Corporation (2006). National Hydrography Dataset Plus (NHDPlus) Home: Horizon Systems Corporation. Retrieved August 2006 from http://www.horizon-systems.com/nhdplus/.

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, 137–153.

Lawler, J. J., White, D., Neilson, R. P., & Blaustein, A. R. (2006). Predicting climate-induced range shifts: Model differences and model reliability. *Global Climate Change Biology*, 12, 1568–1584.

Liaw, A., & Wiener, M. (2002). Classification and regression by random Forest. *R News*, 2/3, 18–22.

Moore, A. A., & Palmer, M. A. (2005). Invertebrate biodiversity in agricultural and urban headwater streams: Implications for conservation and management. *Ecological Applications*, 15, 1169–1177.

Moss, D., Furse, M. T., Wright, J. F., & Armitage, P. D. (1987). The prediction of the macro-invertebrate fauna of unpolluted running-water sites in Great Britain using environmental data. *Freshwater Biology*, 17, 41–52.

Moulton, S. R., II, Carter, J. L., Grotheer, S. A., Cuffney, T. F., & Short, T. M. (2000). Methods of analysis by the US Geological Survey National Water Quality Laboratory: Processing, taxonomy, and quality control of benthic macroinvertebrate samples. Open File Report 00-212, US Geological Survey.

Moulton, S. R., II, Kennen, J. G., Goldstein, R. M., & Hambrook, J. A. (2002). Revised Protocols for Sampling Algal, Invertebrate, and Fish Communities as Part of the National Water-Quality Assessment Program. Open-file Report 02-150, US Geological Survey.

Nilsson, C., Pizzuto, J. E., Moglen, G. E., Palmer, M. A., Stanley, E. H., Bockstael, N. E., et al. (2003). Ecological forecasting and the urbanization of stream ecosystems: challenges for economists, hydrologists, geomorphologists, and ecologists. *Ecosystems*, 6, 659–674.

Ostermiller, J. D., & Hawkins, C. P. (2004). Effects of sampling error on bioassessments of stream ecosystems: applica-

tions to RIVPACS-type models. *Journal of the North American Benthological Society*, 23, 363–382.

Parmenter, A. W., Hansen, A., Kennedy, R. E., Cohen, W., Langer, U., Lawrence, R., et al. (2003). Land use and land cover change in the Greater Yellowstone ecosystem: 1975–1995. *Ecological Applications*, 13, 687–703.

Paul, M. J., & Meyer, J. L. (2001). Streams in the urban landscape. *Annual Review of Ecology and Systematics*, 32, 333–365.

Paulsen, S. G., Hughes, R. M., & Larson, D. P. (1998). Critical elements in describing and understanding our nation's aquatic resources. *Journal of the American Water Resources Association*, 34, 995–1005.

Potter, K. M., Cubbage, F. W., Blank, G. B., & Schaeberg, R. H. (2004). A watershed-scale model for predicting nonpoint pollution risk in North Carolina. *Environmental Management*, 34, 62–74.

Prasad, A. M., Iverson, L. R., & Liaw, A. (2007). Newer classification and regression tree techniques: bagging and random forests for ecological predictions. *Ecosystems*, 9, 181–199.

Pyne, M. I., Rader, R. R., & Christensen, W. F. (2007). Predicting local biological characteristics in streams: a comparison of landscape classifications. *Freshwater Biology*, 52, 1302–1321.

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

Richards, C., Host, G. E., & Arthur, J. W. (1993). Identification of predominant environmental factors structuring stream macroinvertebrate communities within a large agricultural catchment. *Freshwater Biology*, 29, 285–294.

Roy, A. H., Freeman, M. C., Freeman, B. J., Wenger, S. J., Meyer, J. L., & Ensign, W. E. (2006). Importance of riparian forests in urban catchments contingent on sediment and hydrologic regimes. *Environmental Management*, 37, 523–539.

Ruddy, B. C., Lorenz, D. L., Mueller, D. K. (2006). County-level estimates of nutrient inputs to the land surface of the conterminous United States, 1982–2001: US Geological Survey Scientific Investigations Report 2006-5012.

Strayer, D. L., Beighley, R. E., Thompson, L. C., Brooks, S., Nilsson, C., Pinay, G., et al. (2003). Effects of land cover on stream ecosystems: Roles of empirical models and scaling issues. *Ecosystems*, 6, 407–423.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Science*, 43, 1947–1958.

Tate, C. M., Cuffney, T. F., McMahon, G., Giddings, E. M. P., & Zappia, H. (2005). Use of an urban intensity index to assess urban effects on streams in three contrasting environmental settings. In L. R. Brown, R. H. Gray, R. M. Hughes, & M. R. Meador (Eds.), *Effects of urbanization on stream ecosystems* (pp. 291–316). Maryland: American Fisheries Society Symposium 47.

Urban, D. L. (2002). Classification and regression trees. In B. McCune & J. B. Grace (Eds.), *Analysis of ecological communities* (pp. 222–232). Oregon: MjM Software Design.

US Army Corps of Engineers (2006). National Inventory of Dams: U.S. Army Corps of Engineers. Retrieved July 2006 from http://crunch.tec.army.mil/nidpublic/webpages/nid.cfm.

US Department of Agriculture (2006). US General Soil Map (STATSGO)|NRCS NCGC. Retrieved from http://www.ncgc.nrcs.usda.gov/products/datasets/statsgo/.

US Environmental Protection Agency (2000). Mid-Atlantic highlands stream assessment. EPA/903/R-00/015. U.S. Environmental Protection Agency, Philadelphia, Pennsylvania.

US Environmental Protection Agency (2006a). Draft wadeable streams assessment: A collaborative survey of the Nation's streams. EPA 841-B-06-002. Office of Water, Washington DC.

US Environmental Protection Agency (2006b). National Pollutant Discharge Elimination System (NPDES): US Environmental Protection Agency. Retrieved June 2006 from http://cfpub.epa.gov/npdes/.

US General Accounting Office (GAO) (2002). Water quality: Inconsistent state approaches complicate nation's efforts to identify its most polluted waters. GAO-02-186. United State General Accounting Office, 441 G. Street NW, Washington, DC.

US Geological Survey (2006a). USGS National Water-Quality Assessment Program (NAWQA). Retrieved from http://water.usgs.gov/nawqa/.

US Geological Survey (2006b). MRLC consortium: National Land Cover Dataset (NLCD). Retrieved from http://www.mrlc.gov/.

US Geological Survey (2006c). National Hydrography Dataset (NHD) Home Page. Retrieved from http://nhd.usgs.gov/.

US Geological Survey (2006d). National Elevation Dataset (NED). Retrieved from http://ned.usgs.gov.

US Geological Survey (2007). Grids of agricultural pesticide use in the conterminous United States, 1997: US Geological Survey. Retrieved June 2007 from http://water.usgs.gov/GIS/metadata/usgswrd/XML/agpest97grd.xml.

Van Sickle, J., Baker, J., Herlihy, A., Bayley, P., Gregory, S., Haggerty, P., et al. (2004). Projecting the biological condition of streams under alternative scenarios of human land use. *Ecological Applications*, 14, 368–380.

Van Sickle, J., Hawkins, C. P., Larsen, D. P., & Herlihy, A. T. (2005). A null model for the expected macroinvertebrate assemblage in streams. *Journal of the North American Benthological Society*, 24, 178–191.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. New York: Springer.

Vølstad, J. H., Roth, N. E., Mercurio, G., Southerland, M. T., & Strebel, D. E. (2003). Using environmental stressor information to predict the ecological status of Maryland non-tidal streams as measured by biological indicators. *Environmental Monitoring and Assessment*, 84, 219–242.

Wallace, J. B., Eggert, S. L., Meyer, J. L., & Webster, J. R. (1997). Multiple trophic levels of a stream linked to terrestrial litter inputs. *Science*, 277, 102–104.

Walsh, C. J., Roy, A. H., Feminella, J. W., Cottingham, P. D., Groffman, P. M., & Morgan, R. P., II (2005). The urban stream syndrome: current knowledge and the search for a cure. *Journal of the North American Benthological Society*, 24, 706–723.

Wolock, D. M., Fan, J., & Lawrence, G. B. (1997). Effects of basin size on low-flow stream chemistry and subsurface contact time in the Neversink River watershed, New York. *Hydrological Processes*, *11*, 1273–1286.

Wolock, D. M., & McCabe, G. J. (1995). Comparison of single and multiple flow-direction algorithms for computing topographic parameters in TOPMODEL. *Water Resources Research*, *31*, 1315–1324.

Wright, J. F., Sutcliffe, D. W., & Furse, M. T. (Eds.) (2000). *Assessing the biological quality of fresh waters: RIVPACS and other techniques*. United Kingdom: Freshwater Biological Association.

Yates, A. G., & Bailey, R. C. (2006). The stream and its altered valley: Integrating landscape ecology into environmental assessments of agro-ecosystems. *Environmental Monitoring and Assessment*, *114*, 257–271.