

## Effects of regionalization decisions on an O/E index for the US national assessment

Lester L. Yuan<sup>1</sup>

National Center for Environmental Assessment, US Environmental Protection Agency,  
1200 Pennsylvania Ave, NW, 8623P, Washington, DC 20460 USA

Charles P. Hawkins<sup>2</sup>

Western Center for Monitoring and Assessment of Freshwater Ecosystems, Department of Watershed  
Sciences, and the Ecology Center, Utah State University, Logan, Utah 84322-5210 USA

John Van Sickle<sup>3</sup>

National Health and Environmental Effects Laboratory, US Environmental Protection Agency,  
200 SW 35<sup>th</sup> St., Corvallis, Oregon 97333 USA

**Abstract.** We examined the effects of different regionalization schemes on the performance of River InVertebrate Prediction and Classification System (RIVPACS)-type predictive models in assessing the biological conditions of streams of the US for the National Wadeable Streams Assessment (WSA). Three regionalization schemes were considered: a single national predictive model (MOD1), separate predictive models for each of the 9 WSA aggregated Omernik level III ecoregions (MOD9), and 3 predictive models roughly corresponding to the western US, the Appalachian Mountains, and the Central and Coastal Plains (MOD3). The goal of the WSA was to assess stream condition at the national scale and at the scale of WSA aggregated ecoregions, so we compared the performance of the ratio of the observed number of taxa to the expected number of taxa (O/E) index estimated using different regionalization schemes at both of these spatial scales. We assessed model performance with a randomized resampling procedure, in which we set aside 10% of the reference sites, calibrated the model with the remaining sites, and applied the model to the set-aside sites. Performance statistics for the set-aside reference sites were accumulated over 10 iterations. When summarized at the national scale, mean model predictions of O/E for set-aside reference sites from the 3 different regionalization schemes were all reasonably close to 1. When summarized by the 9 aggregated ecoregions, MOD1 and MOD3 predictions of O/E differed systematically from 1 in certain aggregated ecoregions. Over all 9 ecoregions, the magnitude of these differences was significantly greater than observed with MOD9 predictions. Results from our analysis suggest that O/E values at test sites should be interpreted with respect to mean and SD of O/E of reference sites from the same region to minimize the effects of systematic biases in the predictions. RIVPACS-type predictive models also should be calibrated at a spatial scale similar to the scale at which summary statistics are reported.

**Key words:** RIVPACS predictive model, systematic bias, regionalization, ecoregion.

The national Wadeable Streams Assessment (WSA) assessed the condition of streams in the conterminous US, covering a large geographical area (nearly 8 million km<sup>2</sup>) and a wide range of natural conditions (USEPA 2006). Analyzing the biological data in this

assessment presented a unique challenge because of the spatial scale of the assessment. In our paper, we discuss the issues associated with developing an observed/expected (O/E) index of stream biological condition on the basis of River InVertebrate Prediction and Classification System (RIVPACS)-type predictive models for such a large study area.

RIVPACS-type models assess the biological condition of a stream by comparing taxa that are observed

<sup>1</sup> E-mail addresses: yuan.lester@epa.gov

<sup>2</sup> chuck.hawkins@usu.edu

<sup>3</sup> vansickle.john@epa.gov

in a stream with taxa that would be expected under reference conditions. Model results typically are summarized as an O/E value, which is the ratio of the number of observed taxa (O) to the number of expected taxa (E). Expected taxa are predicted statistically from data collected at a set of reference sites that are assumed to be the least-disturbed streams in the study region. This modeling approach was pioneered in Great Britain (Moss et al. 1987) and has been applied in many different locations (e.g., California: Hawkins et al. 2000b; Australia: Simpson and Norris 2000; North Carolina and the mid-Atlantic US: Van Sickle et al. 2005, Hawkins 2006). Most previous models have been applied to relatively small areas. However, in Australia, RIVPACS-type models for individual states and territories were used to assess the biological condition of streams for most of the continent, a spatial scale comparable with the present study (Simpson and Norris 2000).

The O/E index for the WSA was estimated using 3 regional models for each of 3 large regions of the continental US (the West, the Plains, and the Appalachian Mountains). The decision to model these regions separately was driven by an interest in maximizing the precision and accuracy of the models, while maintaining a sufficiently large number of reference sites to permit the development of robust statistical relationships. Our study evaluates the robustness of the original O/E index with respect to choice of regionalization scheme. That is, we consider the effects of regionalization choices on the accuracy and precision of the resulting O/E index.

In most previous cases, a single RIVPACS-type model has been built for a particular reference data set, but O/E indices for some studies have been based upon a set of regional RIVPACS-type models. For example, Stoddard et al. (2006b) used an O/E index based on 5 distinct regional models to assess the biological condition of western US streams, and the current O/E index for California is based on 3 distinct regional models (Ode et al. 2008). In these cases, regional models were built because preliminary tests indicated that the performance of these regional models was superior to that of a single model applied to the entire assessment area. The O/E index in Australia also was based on separate regional models built for each of 8 states and territories and for different habitat types. Regional model results were then aggregated for a national assessment (Australia State of the Environment Committee 2001). In Australia, political considerations motivated the selection of a regionalization scheme (Davies 2000). In our study, we had no restrictions on how or whether to subdivide the data. So, our initial decision on how to

subdivide the data was driven only by the goals for the national assessment, which were to provide accurate assessments at both the national scale and at the scale of WSA aggregated Omernik level III ecoregions (USEPA 2006, Herlihy et al. 2008).

The reasons for subdividing data for a RIVPACS-type model differ from the reasons for subdividing data in preparation for developing multimetric indices. In the latter case, biological data are subdivided to partition natural heterogeneity in assemblage composition into regions with relatively similar biological assemblages (e.g., Simon and Lyons 1995). Then, expectations for different metric values are defined for each region on the basis of the distribution of the metric values within that region. RIVPACS-type models are designed to account explicitly for natural variations in assemblage composition, and therefore, regions with similar assemblages are not required. However, different regionalization schemes potentially can influence 2 aspects of predictions from a RIVPACS-type model: 1) they can introduce or control systematic spatial biases in model predictions, and 2) they can change the precision of those predictions (e.g., Stoddard et al. 2006b, Ode et al. 2008).

Spatial biases can originate from 3 possible sources. First, environmental factors that cause variations in assemblage structure might be omitted from the model, and these factors might vary systematically in space. Thus, predictions from the resulting model will be systematically biased depending on the value of the omitted factor at a particular location. Second, the functional form relating an environmental factor with assemblage structure might be specified incorrectly. Most RIVPACS models rely on a linear discriminant analysis for relating assemblage structure to environmental factors (Moss et al. 1987), and this linear approximation might cause systematic biases when assemblage composition must be modeled over a large range of environmental conditions. A 3<sup>rd</sup> source of spatial bias originates from systematic variations in the levels of human disturbance in reference sites used to build the model. As is typical for these types of models, reference sites are chosen to represent the least-disturbed conditions in the study area. Within this set of reference sites, one would expect that the least-disturbed sites identified from regions with pervasive human activity would be more disturbed than sites from elsewhere (Herlihy et al. 2008). However, RIVPACS models assume that human disturbance effects are very small in the entire reference data set, and use only variables describing natural gradients to calculate predictions. If the quality of reference sites is unevenly distributed among regions, predictions of E might be systematically

biased across a study area, depending on the relative degree of human disturbance within different regions.

Subdividing the data set into smaller regions can improve the degree to which predictor variables account for variance in the model and influence the precision of the model predictions. First, different predictor variables can be selected for different regions. For example, a variable for catchment limestone geology might be included in a model built for a particular region of the country in which limestone streams occur. Conversely, in a model built at a larger spatial scale, limestone streams might be relatively rare, and therefore, the variable might not be selected because other factors associated with large-scale variation in assemblage composition (e.g., climate) could obscure the regional effect of limestone geology. Hence, subdividing a larger model allows one to optimize the choice of predictor variables to those that are both available and appropriate. Second, subdividing the data set partitions the range of environmental variables such that any single model must account for only a portion of the overall range of a variable. Then, within this restricted range, linear models can more accurately represent inherently nonlinear relationships between assemblage composition and different environmental variables (ter Braak and Prentice 1988). Last, regions potentially can be specified within which the extent of human disturbance in reference sites is more comparable than across the entire study area.

Subdividing the data set has disadvantages as well. RIVPACS-type models are statistical models that predict assemblage composition at new sites on the basis of relationships observed in a reference data set, so the size of this reference data set is important. As with most statistical models, smaller sample sizes (i.e., fewer reference sites) allow fewer explanatory variables for modeling variations in the data (Harrell 2001). Hence, subdividing the data might yield regions in which too few reference sites are available to develop a predictive model. Also, RIVPACS-type models built with a relatively small number of reference sites can overestimate E systematically (Yuan 2006). A 2<sup>nd</sup> disadvantage of subdividing the data is that one must make an a priori decision as to which sites to group into subsets. The subdivisions that we examined for the WSA were based on ecoregions that imposed hard boundaries on environmental factors that varied continuously (Omernik 1987). In some analyses, these ecoregions are relatively ineffective in explaining variability in biological assemblages (Hawkins et al. 2000a). Thus, one could argue that a single well-constructed predictive model would explain variability in macroinvertebrate assemblages across

the country more effectively than several models that rely on a priori, and perhaps inappropriate, regionalization schemes.

We examine the effects of choice of regionalization schemes in the development of a national O/E indicator of stream biological integrity for the WSA. More specifically, we examine the hypothesis that subdividing a larger data set into smaller regions and building models for each of these regions yields more accurate and more precise models, and hence, a more nationally consistent indicator than does a single large-scale model.

## Methods

### *Data*

Physical habitat, chemical, and biological data were collected from sites on wadeable, 1<sup>st</sup>-through 5<sup>th</sup>-order streams in the eastern US in summer 2005. These data were combined with data previously collected in the western US from 2000 to 2003 (Stoddard et al. 2006b) to produce complete spatial coverage of the conterminous US. Sites in both of these surveys were selected by a probabilistic sampling design (see Olsen and Peck 2008) or were handpicked as potential reference sites. All data consisted of a single sample per site. Chemical, physical, and landscape characteristics of all sites were screened on the basis of predefined criteria to identify least-disturbed reference sites (Herlihy et al. 2008). Any probabilistically selected samples that failed  $\geq 1$  of the reference criteria were designated as test sites for later use in quantifying the sensitivity with which O/E values detect nonreference conditions (see below).

To increase the number of reference sites, reference filtering criteria (Herlihy et al. 2008) were applied to additional data collected from several previous regional surveys (New England Wadeable Streams Survey: USEPA 2001; Regional Environmental Monitoring and Assessment survey [R-EMAP] Region 7: Kansas DWP 2002; R-EMAP Region 6: New Mexico Environment Department 2004; Mid-Atlantic Integrated Assessment: Stoddard et al. 2006a). Samples that satisfied all reference criteria were added to the national reference data set. Reference data also were augmented with samples collected by the US Geological Survey (USGS) National Water Assessment Program (NAWQA; USGS 2005). Reference sites from NAWQA data were selected on the basis of criteria similar to those used for US Environmental Protection Agency (EPA) data (D. Carlisle, USGS NAWQA program, personal communication). Also, data from an extensive reference-site survey in the western US conducted by Utah State University (USU) were included. Reference sites were

TABLE 1. Number of stream kilometers, reference sites, and stream kilometers represented by each reference site in each National Wadeable Stream Assessment (WSA) aggregated Omernik level III ecoregion. Aggregated ecoregion codes are given in parentheses.

Aggregated ecoregion	Stream km	Number of reference sites	Stream km/reference site
Coastal Plain (CPL)	116,057	64	1813
Northern Appalachians (NAP)	157,542	138	1142
Northern Plains (NPL)	21,633	52	416
Southern Appalachians (SAP)	287,124	318	903
Southern Plains (SPL)	30,994	44	704
Temperate Plains (TPL)	162,314	82	1979
Upper Midwest (UMW)	58,804	27	2178
Western Mountains (WMT)	203,436	555	367
Xeric West (XER)	41,816	169	247

unevenly distributed across the country because of the many disparate sources of data (Table 1).

Field collection methods across all regional surveys were similar and are described in detail elsewhere (e.g., USEPA 2004, Hughes and Peck 2008), so we provide only a brief review here. In all EPA surveys, macroinvertebrates were collected with a D-frame kick net (~500- $\mu$ m mesh) from 0.09-m<sup>2</sup> sampling plots at locations on 11 equally spaced transects along the sampled reach (0.99 m<sup>2</sup> total area). Samples were composited, sieved, and preserved with 95% ethanol to a final concentration of ~70%. The number of individuals identified varied, but the target number of individuals was always  $\geq 300$ . Individuals were identified in the laboratory to the lowest practical taxonomic level (usually genus). NAWQA and USU sampling protocols differed somewhat from EPA protocols. USU data were collected in fast-water habitats (e.g., riffles) from eight 0.09-m<sup>2</sup> sampling plots, but in all other respects, the sampling method was identical to that used in the EPA surveys. In NAWQA, five 0.25-m<sup>2</sup> samples (1.25 m<sup>2</sup> total area) were collected preferentially from the richest habitats (usually riffles). The collection method (Slack samples, kick nets, etc.) varied depending upon the type of habitat sampled (Moulton et al. 2002). Herlihy et al. (2008) assessed the effects of differences between NAWQA and EPA sampling protocols on macroinvertebrate samples collected, and Carlisle and Hawkins (2008) describe the effects of sample type on O/E estimates. Both analyses revealed sampling effects, but overall, the 2 sampling protocols yielded reasonably comparable biological data.

Most candidate predictor variables could be extracted easily from maps. These variables included the geographical location of the site (decimal-degree latitude and longitude), watershed area (km<sup>2</sup>), and elevation (m). Long-term climatic variable summaries, including average annual precipitation (inches), average maximum and minimum air temperature (mean of

long-term monthly maxima and minima; °C), mean annual average air temperature, annual number of wet days, and average annual relative humidity at the site, were obtained from PRISM (<http://www.prism.oregonstate.edu/docs/przfact.html>; Daly et al. 2001). Sampling day of year also was included in the list of potential predictor variables (Table 2). Dummy variables indicating the dominant geological composition of the watershed (e.g., carbonate rocks or sedimentary in origin) also were included.

At some sites, field slope measurements were available, so we calculated a stream power index as the product of the local slope and the square root of the catchment area. This unitless index value is approximately correlated with the local unit stream power (P. Kaufmann, US EPA, personal communication).

#### Statistical analysis

*RIVPACS model development.*—We randomly resampled observed macroinvertebrate assemblages at all sites to 300 individuals. We then summarized taxa observed in the full national reference data set and the WSA test sites (see above for definition of test sites) in terms of nonambiguous operational taxonomic units (OTU). OTUs were required to ensure that individuals identified at different levels of taxonomic resolution were not double counted (Ostermiller and Hawkins 2004). For example, if individuals in a sample were identified at both the genus level (e.g., *Caenis*) and at the family level (e.g., *Caenidae*), one could potentially count a single occurrence of *Caenis* as 2 taxa (i.e., the genus and family). Specification of OTUs ensured that individual taxa were consistently counted at a single taxonomic level across all samples in the data set. In assigning OTUs, we tried to maximize the biological information across samples. For example, among all samples, individuals identified no further than the family *Caenidae* were recorded in only 16 samples, whereas  $\geq 1$  *Caenidae* genus (*Amercaenis*, *Brachycercus*, *Caenis*, or *Cercobranchys*) was recorded from 785

TABLE 2. Explanatory variables selected for models built for each National Wadeable Stream Assessment (WSA) aggregated Omernik level III ecoregion. See Table 1 for aggregated ecoregion names.

Variable	CPL	NAP	NPL	SAP	SPL	TPL	UMW	WMT	XER
Latitude		X		X	X		X	X	X
Longitude		X	X	X				X	X
Log <sub>10</sub> (watershed area)	X	X	X	X	X	X		X	X
Elevation	X	X		X		X		X	X
Day of year	X	X	X	X			X	X	X
Log <sub>10</sub> (precipitation)			X	X		X		X	X
Maximum air temperature	X	X			X	X		X	X
Number of wet days			X	X				X	X
Average relative humidity	X	X		X				X	X
Sedimentary geology	X	X						X	X
Carbonate geology				X					
Stream power index								X	X

samples. The choices here were to: 1) aggregate all genus-level identifications to the family level and use the family as the OTU or 2) exclude any individual identified as Caenidae and retain all genus-level occurrences. In this example, the 2<sup>nd</sup> option seemed likely to retain far more biologically relevant information than the 1<sup>st</sup> option. Choices for other taxa were not as clear, but we generally chose to coarsen taxonomic resolution in cases in which retaining the finer resolution would have forced us to exclude occurrences in twice as many samples as we retained. Most final OTU designations were genus-level identifications. We refer to these OTUs as taxa for the remainder of our paper.

RIVPACS-type predictive models were developed following methods described previously in Clarke et al. (2003) and Hawkins et al. (2000b). In brief, we clustered reference sites with the flexible  $\beta$  clustering technique ( $\beta = -0.5$ ) applied to a matrix of pairwise similarities in the presence/absence of macroinvertebrate taxa (Sørensen index) at different reference sites. We used linear discriminant function analysis to identify linear combinations of candidate explanatory variables that maximized differences between clusters of reference sites. We applied the resulting discriminant functions to new sites and used them to estimate the probability that each site was a member of each cluster. We used these membership probabilities to weight the frequencies of occurrence of each taxon within each cluster to determine an average probability of capture ( $p_c$ ) for that taxon at the new site. Following Hawkins et al. (2000b), we computed E at a site by summing all  $p_c \geq 0.5$ . We computed O as the number of taxa for which  $p_c \geq 0.5$  that were actually observed at a site. The value O/E is interpreted as a site-specific measure of the biological condition of the site. Values of O/E < 1.0 indicate that some expected

taxa were not captured at the site, and these taxon losses have been associated with human activities (e.g., Davies 2000, Hawkins et al. 2000b, Hemsley-Flint 2000).

We developed RIVPACS models using 3 different regionalization schemes: a single RIVPACS model for the entire country (MOD1), separate models for each of the 9 aggregated ecoregions (MOD9) (USEPA 2006, Herlihy et al. 2008), and an intermediate-scale scheme (MOD3) in which we combined Western Mountain (WMT) and Xeric (XER) aggregated ecoregions into a single model for the West, Northern (NAP) and Southern Appalachians (SAP) aggregated ecoregions into a single Appalachians model, and all remaining aggregated ecoregions (Coastal Plain [CPL], Temperate Plains [TPL], Southern Plains [SPL], Northern Plains [NPL], and Upper Midwest [UMW]) into a single Plains model.

*Randomized resampling.*—The number of reference sites was limited for many of the regional models, so we used a randomized resampling procedure to evaluate model performance. We first set aside a random 10% of reference sites. We calibrated the predictive model with the remaining 90% of sites, and computed predictions for the set-aside 10%. This procedure guaranteed that data used to assess the performance of the model were independent of data used to calibrate the model. We repeated this procedure 9 more times such that each reference site was set aside once. The resulting validation data set was composed of 1 O/E value for each reference site computed from 1 of 10 resampled models. We computed model performance statistics (mean and SD of O/E) from this ensemble of O/E values. Henceforth, the terms mean(O/E) and SD(O/E) of set-aside reference sites refer to statistics computed from this resampled ensemble.

The resampling calculation required 10 repetitions of the entire model-building process, so we automated 2 aspects of model building that usually are done by hand. First, we automatically set the pruning level for each clustering dendrogram by searching candidate pruning levels until we identified a level that maximized the number of clusters while keeping the minimum number of members per cluster to 5 sites. In CPL, for example, across 10 resampled iterations, the average number of sites/cluster was  $\sim 11$ , and ranged from a maximum of 19 to a minimum of 5 sites/cluster. Second, we modified the variable selection procedures to identify a single set of explanatory variables to use for all resampled iterations. For each reference data set, we built preliminary models on the basis of 100 data sets resampled with replacement from the original reference data set and used stepwise linear discriminant analysis to select variables from a master list of candidate explanatory variables. We then selected those variables that appeared in  $\geq 75\%$  of the resampled iterations. To guard against overfitting, we also restricted the number of explanatory variables for each model to  $< N/10$ , where  $N$  was the number of reference sites (Harrell 2001). We then fixed this final set of variables for each model, and ran the 10-fold resampling calculation with a single discriminant analysis based on the same set of variables for each iteration to accumulate statistics for assessing model performance.

*Model biases.*—Mean(O/E) and SD(O/E) in the set-aside reference data were computed for each regionalization scheme (MOD1, MOD3, and MOD9). These values were summarized and compared at the level of the 9 aggregated ecoregions and at the national level. Summary statistics at the national scale were computed by weighting O/E values within each aggregated ecoregion by the number of stream kilometers within that region divided by the number of reference sites to account for variation in reference-site density across the aggregated ecoregions (Table 1). The total length of streams within the target population of the survey within each ecoregion was based on 1:100,000 maps and was estimated as part of the analysis of the WSA data (Olsen and Peck 2008). This weighting is similar to the weighting that is used to compute population statistics from data collected with stratified random designs, such as the WSA probability samples (Olsen and Peck 2008, Stoddard et al. 2006b). Here, though, WSA probability weights could not be used to compute reference-site statistics because our reference data included both hand-picked and randomly selected samples.

Mean(O/E) values in reference-quality sites should be  $\approx 1$ . Therefore, we interpreted the differences

between mean(O/E) for the set-aside reference sites and 1 as potential evidence of systematic bias in the models. We considered 2 sources of bias in our assessment of mean(O/E) values: one that arises when an insufficient number of reference sites is used to predict E at a site (i.e., cluster-size bias) and one that arises from spatially systematic errors in model specification (i.e., spatial bias). Cluster-size bias occurs when the number of reference sites is relatively small and when  $p_c$  values are screened by a threshold value  $> 0$  (Yuan 2006). This systematic bias occurs because the set of taxa used to calculate E is constrained to taxa whose predicted  $p_c$  values are at least as large as the selected threshold value.  $p_c$  values are estimated from calibration data that represent only a single instance of a range of possible sampling outcomes associated with the true  $p_c$  of each taxon. Taxa are included or excluded on the basis of this single sampling event (i.e., one calibration data set), so taxa can be included whose true  $p_c$  values are actually less than the threshold value. Thus, E is overestimated and O/E is underestimated. The magnitude of the underestimation of O/E increases as the number of sites/cluster decreases. More specifically, the potential for bias in predicted  $p_c$  within a cluster increases as the cluster size decreases. The magnitude of the underestimation of O/E at a particular site depends on the combination of clusters used to calculate E at that site; if the estimate of  $p_c$  values at a site is heavily weighted toward small clusters, then E can be overestimated. Cluster-size bias can be quantified for a particular site by calculating a theoretical mean(O/E) that is based on the number of samples in each of the clusters specified in the calibration data, the probabilities that the site of interest is a member of each cluster, and an estimate of the distribution of true  $p_c$  values of the modeled taxa (Yuan 2006). We calculated theoretical mean(O/E) for each set-aside reference site, and adjusted each raw O/E value by subtracting the theoretical mean(O/E) and adding 1.

Spatial biases can arise from errors in model specification that vary systematically with location or from spatial variation in the level of human disturbance at reference sites. In the present analysis, we hypothesized that these spatial biases would be evident when predictions from a model calibrated at a large spatial scale (e.g., models built for the MOD1 or MOD3 regionalization schemes) were summarized at a smaller scale (i.e., the 9 aggregated ecoregions). More specifically, we expected that estimates of mean(O/E) for set-aside reference sites based on MOD1 and MOD3 regionalization schemes would differ systematically from 1 within the 9 aggregated ecoregions even after adjusting for cluster-size biases (see above).

In contrast, we expected that estimates of mean(O/E) based on the MOD9 regionalization scheme would be  $\approx 1$  within each of the aggregated ecoregions because the models were calibrated at the same spatial scale as that at which their results were summarized. We tested for a significant difference in spatial bias by applying a blocked ANOVA on the magnitude of the difference between mean(O/E) and 1 (i.e., the bias estimate), with treatments defined by the regionalization scheme (MOD1, MOD3, or MOD9) and by the aggregated ecoregion. We then used Tukey's honestly significant difference (HSD) test to determine if bias estimates for MOD1 and MOD3 differed significantly from those for MOD9.

*Model sensitivity.*—We calculated a different value of O/E at each test site with models built from each of 10 resampled data sets to yield 10 different O/E values for each test site. Within each resampled iteration, a test site was flagged as being outside the experience of the model if its discriminant function predictor values lay outside the joint predictor–variable distributions of all reference-site groups in a  $\chi^2$  test with  $p < 0.01$  (Moss et al. 1987). We calculated final test-site O/E values as the average over the ensemble of 10 iterations, and we declared test sites that had been flagged in  $>5$  of the iterations to be outside the experience of the aggregated ensemble of models. We declared test sites within the experience of the ensemble of models to be statistically different from reference if the mean(O/E) at the site over the 10 resampled iterations was less than a threshold value defined as  $\text{mean(O/E)} - 1.64(\text{SD}[\text{O/E}])$ , where O/E values in this relationship were values from resampled set-aside reference sites (Van Sickle et al. 2007). This threshold value approximately corresponds to a 95% probability that the O/E value at the test site was less than values in the reference population. Test sites used in this calculation did not satisfy  $\geq 1$  reference criteria; thus, they differed at least nominally from reference sites, and the proportion of these sites found by the model to differ significantly from reference provided a measure of the sensitivity of the model to disturbance (Hawkins 2006, Van Sickle et al. 2007).

For each aggregated ecoregion, we tested whether the proportion of test sites declared outside the model experience differed across the 3 regionalization schemes. We compared the proportions between 2 regionalization schemes, for the 3 possible pairings (MOD1 vs MOD3, MOD3 vs MOD9, MOD1 vs MOD9). We used an exact version of McNemar's test (Agresti 1990, SAS Institute 2004) because these proportions were all assessed on the same collection of sites and, hence, were dependent, and because some regions had near-0 counts of outside sites. For each

region, we carried out a family of 3 McNemar's tests and declared a significant difference between 2 models if the  $p$ -value was  $< 0.05/3 = 0.0167$ , applying a Bonferroni correction to maintain a familywise significance level of 0.05. We applied this same testing procedure to the proportions of sites declared to be different from reference by the 3 regionalization schemes. Each comparison for a pair of models included only those test sites that were inside the experience of both models.

## Results

Data were available at 1449 reference sites. These reference sites were unevenly distributed across the country (Table 1). The densest network of reference sites occurred in XER and WMT, with one reference site per 247 and 367 stream kilometers, respectively. The sparsest network of reference sites was available in UMW and TPL, with one reference site per 2178 and 1979 stream kilometers, respectively.

The selected predictor variables varied substantially across different models (only variables used in  $\geq 1$  of the final regional models are shown in Table 2). In the 9 aggregated ecoregional models developed for MOD9,  $\log_{10}(\text{watershed area})$  was the most commonly selected predictor variable, occurring in 8 of 9 ecoregion models. Carbonate geology was selected only once, in the model for SAP. Field measurements of stream slope were available only in XER and WMT, so the stream power index could be included only in models for those aggregated ecoregions. All predictor variables listed in Table 2 were selected and used in each of the models built for MOD1 and MOD3 regionalization schemes. In MOD3, the variable list for the West model also included the stream power index.

Mean(O/E) values calculated for set-aside reference sites were  $< 1$  for all ecoregions and all regionalization schemes, but the magnitude of the difference from 1 varied (Table 3). In most ecoregions mean(O/E) values computed from the MOD9 regionalization scheme were somewhat lower than values computed using MOD1 or MOD3 regionalization scheme. These differences were particularly evident in XER, UMW, CPL, and NPL. However, in SPL, mean(O/E) was closer to 1 with the MOD9 regionalization scheme than with the other 2 regionalization schemes. In the MOD9 regionalization scheme, mean(O/E) values in ecoregions with relatively few reference sites (i.e.,  $< 70$ ) were lower than the other ecoregions. These same differences across ecoregions were observed in the MOD3 and MOD1 regionalization schemes. Mean(O/E) values at the national scale were similar across the 3 regionalization schemes.

TABLE 3. Mean and SD of the observed/expected (O/E) taxa ratio estimated for each Wadeable Streams Assessment (WSA) aggregated Omernik level III ecoregion and nationally from models based on 3 different regionalization schemes (national: MOD1, 3 broad ecoregions: MOD3, and 9 aggregated ecoregions: MOD9). See Methods for explanation of MOD1, MOD3, and MOD9 regionalization schemes. See Table 1 for aggregated ecoregion names.

Aggregated ecoregion	Mean(O/E)			SD(O/E)		
	MOD1	MOD3	MOD9	MOD1	MOD3	MOD9
CPL	0.85	0.83	0.81	0.254	0.254	0.248
NAP	0.99	0.99	0.95	0.165	0.180	0.172
NPL	0.92	0.89	0.88	0.324	0.324	0.290
SAP	0.98	0.94	0.95	0.201	0.203	0.197
SPL	0.74	0.72	0.85	0.343	0.345	0.331
TPL	0.94	0.94	0.94	0.299	0.287	0.328
UMW	0.90	0.87	0.86	0.257	0.205	0.201
WMT	0.96	0.97	0.95	0.229	0.243	0.236
XER	0.93	0.92	0.86	0.274	0.236	0.239
National	0.94	0.93	0.92	0.244	0.243	0.243

SD(O/E) within most of the WSA ecoregions was very weakly affected by the spatial scale of the model (Table 3). SD(O/E) in UMW was substantially less in the MOD9 and MOD3 than in the MOD1 regionalization schemes. SD(O/E) in XER and NPL also decreased from MOD1 to MOD9 regionalization schemes. SD(O/E) values for the other aggregated ecoregions were similar across regionalization schemes. SD(O/E) values varied across aggregated ecoregions and ranged from  $\sim 0.17$  in NAP to  $>0.32$  in SPL and TPL (under the MOD9 regionalization scheme). These patterns in SD(O/E) across aggregated ecoregions were similar for O/E values calculated from all 3 regionalization schemes. SD(O/E) values at

TABLE 4. Mean observed/expected (O/E) taxa ratio adjusted by the theoretical mean(O/E) for each Wadeable Streams Assessment (WSA) aggregated Omernik level III ecoregion and nationally from models based on 3 different regionalization schemes (national: MOD1, 3 broad ecoregions: MOD3, and 9 aggregated ecoregions: MOD9). See Methods for explanation of MOD1, MOD3, and MOD9 regionalization schemes. See Table 1 for aggregated ecoregion names.

Aggregated ecoregion	MOD1	MOD3	MOD9
CPL	0.92	0.92	0.93
NAP	1.02	1.03	1.06
NPL	0.97	0.96	1.03
SAP	1.01	0.98	1.02
SPL	0.79	0.80	1.00
TPL	0.97	0.98	1.01
UMW	0.96	0.95	0.99
WMT	0.99	1.00	1.00
XER	0.97	0.95	1.00
National	0.98	0.98	1.01

the national scale were similar across the 3 regionalization schemes.

Adjusting O/E with theoretical mean(O/E) reduced the magnitude of the difference between mean(O/E) and 1 for most ecoregions and regionalization schemes (Table 4). However, mean(O/E) in CPL remained at  $\sim 0.93$  after adjustment by theoretical mean(O/E) for all regionalization schemes. Also, mean(O/E) in SPL remained at  $\sim 0.8$  for MOD1 and MOD3 regionalization schemes, but was  $\sim 1$  for the MOD9 regionalization scheme. A Tukey HSD test indicated that the magnitudes of the differences between adjusted mean(O/E) and 1 were greater for MOD1 and MOD3 than for MOD9 ( $p < 0.001$  for both comparisons).

The number of test sites that were outside the experience of the model increased markedly from MOD1 to MOD9 for 3 aggregated ecoregions (CPL, NAP, and XER), but changed little for the remaining 6 aggregated ecoregions (Table 5). For example, in CPL, only 2 sites were outside the model experience of MOD1, whereas 30 sites were outside model experience of MOD9. No significant differences across regionalization schemes were observed in the number of test sites declared to be different from reference, except for the increased number for MOD3 in XER. At the national scale, somewhat more sites were found to be different from reference under MOD1 and MOD3 regionalization schemes than under MOD9. More than  $\frac{1}{2}$  of these sites were reclassified under MOD9 as being outside model experience.

## Discussion

When modeling relationships in any data set, one must often consider whether subdividing the data set

TABLE 5. Numbers of test sites declared outside model experience and different from reference in models based on 3 different regionalization schemes (national: MOD1, 3 broad ecoregions: MOD3, and 9 aggregated ecoregions: MOD9). For each Wadeable Stream Assessment (WSA) aggregated level III ecoregion and analysis variable, fonts (regular or **bold**) are identical for regionalization schemes (MOD1, MOD3, MOD9) whose proportions were not declared significantly different by McNemar tests. Totals are not independent of regional counts and were not tested. See Methods for explanation of MOD1, MOD3, and MOD9 regionalization schemes. See Table 1 for aggregated ecoregion names.

Aggregated ecoregion	No. of test sites	Outside model experience			Different from reference		
		MOD1	MOD3	MOD9	MOD1	MOD3	MOD9
CPL	80	2	<b>19</b>	<b>30</b>	13	8	6
NAP	71	0	5	<b>23</b>	26	24	17
NPL	75	0	0	1	5	5	4
SAP	173	0	<b>9</b>	2	33	37	45
SPL	42	0	1	0	2	1	4
TPL	131	0	0	1	13	13	8
UMW	63	1	1	0	9	12	10
WMT	405	1	5	7	59	53	54
XER	143	6	6	<b>27</b>	35	<b>48</b>	24
Total	1183	10	46	91	195	201	172

will yield better models, and the RIVAPCS-type predictive models considered here are no different. The continental spatial scale of the present study area provided a unique opportunity to explore these regionalization questions at a very large spatial scale. In particular, the number of samples in this data set was large enough to allow us to consider 3 spatial scales of regionalization. Also, the spatial extent of the data allowed us to consider the potential errors inherent in modeling macroinvertebrate composition across very broad ranges of environmental conditions. We have considered the effects of 3 spatial scales of regionalization on the accuracy, precision, and sensitivity of a national O/E index of stream condition.

*Model accuracy*

We considered 2 aspects of model accuracy when evaluating the performance of different regionalization schemes: 1) the accuracy of the O/E values, relative to expectations, and 2) the accuracy of assessments of stream biological condition on the basis of these O/E values. Our expectation of accurate O/E values is that mean(O/E) will be close to 1 when a model is applied to reference sites drawn from the same population as the reference sites used to calibrate the model. In the present analysis, we found that different regionalization schemes can strongly influence this aspect of model accuracy by introducing or controlling cluster-size and spatial biases. We observed substantial cluster-size biases in several ecoregions, particularly in those with relatively few reference sites (Table 3). Also, as expected, the magnitude of the difference between mean(O/E) and 1 generally increased as the

spatial scale decreased (i.e., from MOD1 to MOD9) and the number of reference sites decreased. Surprisingly, in some aggregated ecoregions these biases persisted regardless of the spatial scale at which models were built (e.g., mean[O/E] = 0.92 in NPL with the MOD1 regionalization scheme). This persistence of biases suggests that within certain aggregated ecoregions, predictions from models built using any of the 3 regionalization schemes were based on clusters with relatively few reference sites. That is, even when the spatial extent of the model was very large (e.g., MOD1 or MOD3), predictions at certain sites were still calculated from a relatively small number of similar reference sites.

Our finding of significant spatial biases in O/E values predicted from the MOD1 and MOD3 regionalization schemes was not surprising and matched our initial hypothesis (Table 4). Spatial biases in RIVPACS models have been observed in a few other studies. For example, Ostermiller and Hawkins (2004) observed that O/E values at specific sites were systematically >1 or <1 over repeated samples, whereas Ode et al. (2008) observed that O/E values estimated from regional-scale models differed systematically from values estimated from models built at smaller spatial scales. In the broader application of habitat modeling, the potential for spatial biases has been discussed extensively (e.g., Fielding and Bell 1997, Cade et al. 2005, Barry and Elith 2006). In habitat models, spatial biases can arise when errors in model specification are associated with spatial location. RIVPACS-type predictive models are subject to the same issues, but 2 factors might make RIVPACS models even more susceptible to spatial biases. First, predictor variables

used in RIVPACS models are rarely, if ever, the proximate variables causing changes in assemblage composition. Instead, RIVPACS predictor variables are only correlated with the true causal factors, and it is possible that relationships between distal variables and assemblage composition vary more with spatial location than would a relationship based on a more proximate factor. For example, in the present analysis, mean(O/E) value in SPL were substantially  $<1$  under MOD1 and MOD3 regionalization schemes (Table 4). In these models, the relationships between key predictor variables (watershed area, latitude, and maximum air temperature) and assemblage composition estimated for SPL might differ substantially from the same relationship estimated from all the Plains aggregated ecoregions (in the case of MOD3 regionalization scheme) or from the entire country (in the case of the MOD1 regionalization scheme). A 2<sup>nd</sup> potential source of spatial bias unique to RIVPACS models is the assumption that the reference sites used to build the model have comparable levels of human disturbance. Systematic differences in the level of human disturbance in reference sites can cause systematic biases in model predictions. That is, least-disturbed reference sites in SPL might be substantially more disturbed by human activities than reference sites in the other Plains aggregated ecoregions; thus, mean(O/E) for SPL was  $<1$  when estimated from the larger-scale models (i.e., models built using MOD1 and MOD3 regionalization scheme).

The accuracy of condition assessments based on O/E values depends on the inherent accuracy of O/E values *and* on how we interpret those values. The effects of biases in the O/E values on condition assessments can be minimized effectively by interpreting model predictions at test sites relative to model predictions at reference sites. Thus, even if test-site predictions are biased, they will be compared with similarly biased predictions at reference sites, and the resulting condition assessment will still be accurate. This approach for interpreting RIVPACS-type output is common, and we used it to evaluate the sensitivity of the model predictions in the present analysis (Table 5). Optimally, when comparing test-site O/E to reference-site O/E statistics, reference-site statistics should be computed at the smallest spatial scale of assessment. In our study, mean(O/E) and SD(O/E) values varied strongly across aggregated ecoregions (Table 3). Thus, application of a single, national threshold to sites in all aggregated ecoregions might yield misleading assessments. After sites are assessed at the smallest scales, assessments at larger spatial scales can be calculated easily by aggregating smaller-scale assessments.

Other approaches for controlling the effects of biases are possible. For cluster-size bias, adjusting O/E values by theoretical predictions based on assumed binomial distributions is effective (Table 4; Yuan 2006). However, these theoretical predictions are only estimates, and their use to correct observed O/E values introduces another source of variability into the reference distribution of O/E values. This additional error increases SD(O/E) and reduces the sensitivity of the model. For this reason, theoretical adjustments might be useful only for quantifying the magnitude of cluster-size biases, and sites still should be assessed with unadjusted values as described above. To control spatial biases, models can be built at a spatial scale similar to the scale at which assessment results will be reported. As with most other regression models, one usually can assume that RIVPACS models provide unbiased predictions for samples from the same populations that were used to calibrate the models. Thus, we assumed in our analysis that models built at the scale of the 9 aggregated ecoregions would provide unbiased prediction when summarized at the same aggregated ecoregion scale (Table 4).

Our results indicate that assessing stream condition using fixed thresholds probably will lead to misleading assessments. For example, in some assessments, an O/E value  $< 0.5$  is interpreted as a loss of  $\geq 50\%$  of the expected taxa in a sample (USEPA 2006). In this mode of assessment, systematic differences in biases across ecoregions can lead to over- and underprediction of biological conditions within different ecoregions.

#### *Precision*

The precision of O/E index values, as quantified as SD(O/E), is partially determined by how accurately RIVPACS models account for variations in assemblage composition across the study area. As we reduced the spatial scale of our models across regionalization schemes, we expected that more appropriate predictor variables specific to each particular region would be selected and that the range of environmental conditions that the model would have to represent would be reduced. Thus, we expected that precision of O/E values would improve as we reduced the spatial scale of our models. Our comparisons across regionalization schemes yielded mixed support for these initial hypotheses because of several different possible factors (Table 3). First, the smallest spatial scale we considered (i.e., the aggregated ecoregions) was still large relative to the spatial extent of most other RIVPACS models, so in certain aggregated ecoregions (e.g., WMT and SAP), the reduction in spatial scale from MOD1 to MOD9 might not have reduced the range of environmental

conditions enough to influence  $SD(O/E)$ .  $SD(O/E)$  for models built at smaller spatial scales in these same geographic areas (e.g., Mid-Atlantic Highlands, Hawkins 2006; Wyoming, Hargrett et al. 2007) are generally less than observed in our study, and we might have observed decreases in  $SD(O/E)$  had we reduced spatial scale below that considered in MOD9. Indeed, RIVPACS models described in Stoddard et al. (2006b) subdivided the western US into 5 regions (cf. WMT and XER in our study) using a scheme to maximize within-group similarity of predictor variables explicitly, and  $SD(O/E)$  values from these models were lower than  $SD(O/E)$  in our study.

Second, the most appropriate variables for predicting variations in assemblage composition might not have been available in certain ecoregions, so the degree to which predictor variables could be refined with decreased spatial scale was limited. For example, some of the decrease in  $SD(O/E)$  we observed in XER from MOD1 to MOD3 might have been caused by the inclusion of stream power index as a predictor variable in the model built at a smaller spatial scale. In other ecoregions, key regionally specific predictor variables might not have been available for the present analysis.

Third, the degree to which available reference sites represented the range of stream types might have influenced model precision. A sufficiently representative set of reference sites helps ensure that predictions at a particular site are based on enough comparable reference sites, and can improve the predictive power of the model. One crude measure of representativeness is reference-site density. In NPL, where we observed a modest decrease in  $SD(O/E)$  from MOD1 to MOD9 (Table 3), the density of reference sites was relatively high (Table 1).

Values of  $SD(O/E)$  also are influenced by sampling and temporal variability. Observed patterns in  $SD(O/E)$  across aggregated ecoregions might partially reflect true differences in the temporal variability of macro-invertebrate assemblages across different stream types. For example, a more variable assemblage composition might be expected in streams with flashy highly variable flow regimes, whereas more stable assemblage composition might be expected in streams with relatively stable flow (e.g., Death and Winterbourne 1994). It seems likely that the temporal component of  $SD(O/E)$  would differ even across streams within a single ecoregion.

### *Sensitivity*

One disadvantage of reducing the spatial scale of different RIVPACS models by subdividing the data is that the range of applicability of each model is also

reduced. This reduction in the range of applicability even influences the degree to which the model can be applied to sites within the same ecoregion (Table 5). For example, in our study, it is likely that the reference sites that were available in CPL represented only a limited range of conditions because of widespread effects of human activities within this aggregated ecoregion. Incorporation of additional reference sites from other aggregated ecoregions (e.g., SPL) in larger-scale regionalizations extended the degree to which the reference sites represented the full range of conditions and reduced the number of sites found to be outside the experience of the model (Table 5). Furthermore, these additional reference sites generally did not decrease the precision of the model. Therefore, regionalization schemes that incorporate larger geographical areas might provide a means to extend the range of applicability of a RIVPACS model. Ultimately, identifying and sampling more reference sites that span the entire range of natural conditions within each ecoregion still should provide the best models (e.g., Carlisle et al. 2008).

The regionalization schemes considered in our analysis did not influence model sensitivity strongly. This finding was not surprising given that  $SD(O/E)$  did not differ substantially across regionalization schemes. However, we did observe a concomitant increase in the number of sites found to be different from reference in a few aggregated ecoregions where  $SD(O/E)$  decreased across regionalization schemes. For example, in XER,  $SD(O/E)$  decreased substantially from MOD1 to MOD3 (Table 3), and the number of test sites different from reference increased from 35 to 48 (Table 5). However, the relationship between model sensitivity and  $SD(O/E)$  was not consistent, and some possible explanations for this lack consistency might include the fact that our assessments of sensitivity were partially confounded by the decrease in the range of model applicability. For example,  $SD(O/E)$  in XER was lower for MOD9 than for MOD1, but fewer test sites were different from reference in MOD9 than in MOD1. It seems likely that some sites that were classified as different from reference in MOD1 were reclassified as outside model experience in XER. In other aggregated ecoregions (e.g., NPL),  $SD(O/E)$  values were so high in MOD1 and MOD3 that the modest decrease in MOD9 might not have been enough to increase the number of sites found to be different from reference.

### *Implications*

Our original WSA analysis subdivided the national data set into 3 models and used the same regionalization scheme as considered here with MOD3. The

results from our present study suggest that some of the original interpretations of O/E values as taxon loss might have been in error. We have found with our present analysis that, in certain aggregated ecoregions, predictions of O/E under the MOD3 regionalization scheme were systematically  $<1$  because of cluster-size effects or because of spatial biases. Thus, estimates of the proportions of taxa lost in different aggregated ecoregions, in which O/E values were interpreted using fixed thresholds at O/E = 0.9, 0.8, and 0.5, might have been in error. In the WSA report, test sites also were classified as being significantly different from reference by comparing test-site values of O/E with reference distributions. These classifications were then used to assess the relative risk of different stressors (Van Sickle and Paulsen 2008). On the basis of our analysis, we can be reasonably confident that these assessments were accurate. Indeed, given that our estimates of O/E were relatively imprecise, our original designations of the number of degraded sites (i.e., sites significantly different from reference) probably underestimated the true extent of biological degradation in streams across the US.

The regionalization schemes considered in our analysis originally were defined to partition the variability in macroinvertebrate assemblage composition in support of the development of multimetric indices (Herlihy et al. 2008), and these regions were not necessarily the best regions for optimizing the performance of RIVPACS models. Little guidance is available for methods that optimally subdivide data for RIVPACS models, and more research in this area is needed. On the basis of our analysis, it seems that one should consider the spatial scale of the assessment questions, the range of key environmental parameters, and the relative quality of reference sites within different potential regions when making regionalization decisions.

The resampling method introduced here provided a useful way to assess the performance of our models with independent reference sites. Independent validation data is critical for accurate estimates of mean(O/E), but reference-site databases are rarely large enough to split into calibration and validation data. We used mean(O/E) and SD(O/E) statistics calculated from an ensemble of set-aside reference sites to compare different regionalization schemes, but the relationship between these statistics (each estimated from an ensemble of 10 different models) and the statistics one would calculate from a single model and a single validation data set is uncertain. More work is needed to assess the degree to which the ensemble of resampled models used here represents the performance of a single model.

## Acknowledgements

The authors acknowledge the data collection efforts of the many WSA sampling teams, and the funding for the national stream survey provided by US EPA Office of Water and Office of Research and Development. Views expressed in this paper are those of the authors and do not reflect the views of the US EPA.

## Literature Cited

- AGRESTI, A. 1990. *Categorical data analysis*. 2<sup>nd</sup> edition. John Wiley and Sons, New York.
- AUSTRALIA STATE OF THE ENVIRONMENT COMMITTEE. 2001. Australia State of the Environment 2001. Independent report to the Commonwealth Minister for the Environment and Heritage. CSIRO Publishing on behalf of the Department of the Environment and Heritage, Collingwood, Victoria, Australia. (Available from: <http://www.environment.gov.au/soe/2001/index.html>)
- BARRY, S., AND J. ELITH. 2006. Error and uncertainty in habitat models. *Journal of Applied Ecology* 43:413–423.
- CADE, B. S., B. R. NOON, AND C. H. FLATHER. 2005. Quantile regression reveals hidden bias and uncertainty in habitat models. *Ecology* 86:786–800.
- CARLISLE, D. M., AND C. P. HAWKINS. 2008. Land use and the structure of western US stream invertebrate assemblages: predictive models and ecological traits. *Journal of the North American Benthological Society* 27:986–999.
- CARLISLE, D. M., C. P. HAWKINS, M. R. MEADOR, M. POTAPOVA, AND J. FALCONE. 2008. Biological assessments of Appalachian streams derived from predictive models for fish, macroinvertebrate, and diatom assemblages. *Journal of the North American Benthological Society* 27:16–37.
- CLARKE, R. T., J. F. WRIGHT, AND M. T. FURSE. 2003. RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. *Ecological Modelling* 160:219–233.
- DALY, C., G. H. TAYLOR, W. P. GIBSON, T. W. PARZYBOK, G. L. JOHNSON, AND P. PASTERIS. 2001. High-quality spatial climate data sets for the United States and beyond. *Transactions of the American Society of Agricultural Engineers* 43:1957–1962.
- DAVIES, P. E. 2000. Development of a national river bioassessment system (AUSRIVAS) for Australia. Pages 113–124 in J. F. Wright, D. W. Sutcliffe, and M. T. Furse (editors). *Assessing the biological quality of fresh waters: RIVPACS and other techniques*. Freshwater Biological Association, Ambleside, UK.
- DEATH, R. G., AND M. J. WINTERBOURNE. 1994. Environmental stability and community persistence: a multivariate perspective. *Journal of the North American Benthological Society* 13:125–139.
- 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.
- HARGETT, E., J. ZUMBERGE, C. P. HAWKINS, AND J. OLSON. 2007. Development of a RIVPACS-type predictive model for

- bioassessment of wadeable streams in Wyoming. *Ecological Indicators* 7:807–826.
- HARRELL, F. E. 2001. *Regression modeling strategies*. Springer-Verlag, New York.
- HAWKINS, C. P. 2006. Quantifying biological integrity by taxonomic completeness: evaluation of a potential indicator for use in regional- and global-scale assessments. *Ecological Applications* 16:1251–1266.
- HAWKINS, C. P., R. H. NORRIS, J. GERRITSEN, R. M. HUGHES, S. K. JACKSON, R. K. JOHNSON, AND R. J. STEVENSON. 2000a. Evaluation of the use of landscape classification for the prediction of freshwater biota: synthesis and recommendations. *Journal of the North American Benthological Society* 19:541–556.
- HAWKINS, C. P., R. H. NORRIS, J. N. HOGUE, AND J. W. FEMINELLA. 2000b. Development and evaluation of predictive models for measuring the biological integrity of streams. *Ecological Applications* 10:1456–1477.
- HEMSLEY-FLINT, B. 2000. Classification of the biological quality of rivers in England and Wales. Pages 55–70 in J. F. Wright, D. W. Sutcliffe, and M. T. Furse (editors). *Assessing the biological quality of fresh waters: RIVPACS and other techniques*. Freshwater Biological Association, Ambleside, UK.
- HERLIHY, A. T., S. G. PAULSEN, J. VAN SICKLE, J. L. STODDARD, C. P. HAWKINS, AND L. L. YUAN. 2008. Striving for consistency in a national assessment: the challenges of applying a reference condition approach at a continental scale. *Journal of the North American Benthological Society* 27:860–877.
- HUGHES, R. M., AND D. V. PECK. 2008. Acquiring data for large aquatic resource surveys: the art of compromise among science, logistics, and reality. *Journal of the North American Benthological Society* 27:837–859.
- KANSAS DWP (KANSAS DEPARTMENT OF WILDLIFE AND PARKS). 2002. Measuring the status and trends of biological resources in Kansas using Environmental Monitoring and Assessment Program probability based sampling design (R-EMAP). Kansas Department of Wildlife and Parks, Pratt, Kansas. (Available from: <http://www.epa.gov/emap/remap/html/docs/kansas.html>)
- MOSS, D. M., M. T. FURSE, J. F. WRIGHT, AND P. D. ARMITAGE. 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., J. G. KENNEN, R. M. GOLDSTEIN, AND J. A. HAMBROOK. 2002. Revised protocols for sampling algal, invertebrate, and fish communities as part of the National Water-Quality Assessment Program. U.S. Geological Survey Open File Report 02–150. US Geological Survey, Reston, Virginia.
- NEW MEXICO ENVIRONMENT DEPARTMENT. 2004. Water quality survey summary for the Lower Rio Chama watershed. Surface Water Quality Bureau, New Mexico Environment Department, Santa Fe, New Mexico. (Available from: <http://www.nmenv.state.nm.us/swqb/Surveys/LowerChama1999.pdf>)
- ODE, P. R., C. P. HAWKINS, AND R. D. MAZOR. 2008. Comparability of biological assessments derived from predictive models and multimetric indices of increasing geographic scope. *Journal of the North American Benthological Society* 27:967–985.
- OLSEN, A., AND D. PECK. 2008. Survey design and extent estimates for the Wadeable Streams Assessment. *Journal of the North American Benthological Society* 27:822–836.
- OMERNIK, J. M. 1987. Ecoregions of the conterminous United States. *Annals of the Association of American Geographers* 77:118–125.
- OSTERMILLER, J. D., AND C. P. HAWKINS. 2004. Effects of sampling error on bioassessments of stream ecosystems: applications to RIVPACS-type models. *Journal of the North American Benthological Society* 23:363–382.
- SAS INSTITUTE. 2004. SAS OnlineDoc 9.1.3. SAS Institute, Cary, North Carolina. (Available from: <http://support.sas.com/onlinedoc/913/docMainpage.jsp>)
- SIMON, T. P., AND J. LYONS. 1995. Application of the Index of Biotic Integrity to evaluate water resource integrity in freshwater ecosystems. Pages 245–262 in W. S. Davis and T. P. Simon (editors). *Biological assessment and criteria: tools for water resource planning and decision making*. CRC Press, Boca Raton, Florida.
- SIMPSON, J. C., AND R. H. NORRIS. 2000. Biological assessment of river quality: development of AUSRIVAS models and outputs. Pages 125–142 in J. F. Wright, D. W. Sutcliffe, and M. T. Furse (editors). *Assessing the biological quality of fresh waters: RIVPACS and other techniques*. Freshwater Biological Association, Ambleside, UK.
- STODDARD, J. L., A. T. HERLIHY, B. H. HILL, R. M. HUGHES, P. R. KAUFMANN, D. J. KLEMM, J. M. LAZORCHAK, F. H. MCCORMICK, D. V. PECK, S. G. PAULSEN, A. R. OLSEN, D. P. LARSEN, J. VAN SICKLE, AND T. R. WHITTIER. 2006a. Mid-Atlantic Integrated Assessment (MAIA) state of the flowing waters report. EPA/620/R-06/001. Office of Research and Development, US Environmental Protection Agency, Washington, DC.
- STODDARD, J. L., D. V. PECK, A. R. OLSEN, D. P. LARSEN, J. VAN SICKLE, C. P. HAWKINS, R. M. HUGHES, T. R. WHITTIER, G. LOMNICKY, A. T. HERLIHY, P. R. KAUFMANN, S. A. PETERSON, P. L. RINGOLD, S. G. PAULSEN, AND R. BLAIR. 2006b. Environmental Monitoring and Assessment (EMAP) western streams and rivers statistical summary. EPA/620/R-05/006. Office of Research and Development, US Environmental Protection Agency, Washington, DC.
- TER BRAAK, C. J. F., AND I. C. PRENTICE. 1988. A theory of gradient analysis. *Advances in Ecological Research* 18:271–317.
- USEPA (US ENVIRONMENTAL PROTECTION AGENCY). 2001. New England Wadeable Streams (NEWS). Region 1, US Environmental Protection Agency, Boston, Massachusetts. (Available from: <http://epa.gov/Region1/lab/reportsdocuments/wadeable/index.html>)
- USEPA (US ENVIRONMENTAL PROTECTION AGENCY). 2004. Wadeable Streams Assessment field operations manual. EPA841-B-04-004. Office of Water and Office of Research and Development, US Environmental Protection Agency, Washington, DC.
- USEPA (US ENVIRONMENTAL PROTECTION AGENCY). 2006. Wadeable Streams Assessment: a collaborative survey of the nation's streams. EPA 841-B-06-002. Office of Research

- and Development and Office of Water, US Environmental Protection Agency, Washington, DC.
- USGS (US GEOLOGICAL SURVEY). 2005. Design of the National Water-Quality Assessment Program: occurrence and distribution of water-quality conditions. U.S. Geological Survey Circular 1112. US Geological Survey, Reston, Virginia.
- VAN SICKLE, J., C. P. HAWKINS, AND D. P. LARSEN. 2005. A null model for the expected macroinvertebrate assemblage in streams. *Journal of the North American Benthological Society* 24:178–191.
- VAN SICKLE, J., D. P. LARSEN, AND C. P. HAWKINS. 2007. Exclusion of rare taxa affects performance of the O/E index in bioassessments. *Journal of the North American Benthological Society* 26:319–331.
- VAN SICKLE, J., AND S. G. PAULSEN. 2008. Assessing the attributable risks, relative risks, and regional extents of aquatic stressors. *Journal of the North American Benthological Society* 27:920–931.
- YUAN, L. L. 2006. Theoretical predictions of observed to expected ratios in RIVPACS-type predictive model assessments of stream biological conditions. *Journal of the North American Benthological Society* 25:841–850.

*Received: 21 November 2007*

*Accepted: 25 July 2008*