2007

Modeling Natural Environmental Gradients Improves the Accuracy and Precision of Diatom-Based Indicators

Yong Cao

Charles P. Hawkins

John Olson
California State University, Monterey Bay, joolson@csumb.edu

J. N. Am. Benthol. Soc., 2007, 26(3):566–585

Modeling natural environmental gradients improves the accuracy and precision of diatom-based indicators

Yong Cao[1], Charles P. Hawkins[2], AND John Olson[3]

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

Mary A. Kosterman[4]

Surface Water Quality, Idaho Department of Environmental Quality, Boise, Idaho 83706 USA

Abstract. Diatom-based indicators can contribute significantly to comprehensive assessments of stream biological conditions. We used modeling to develop, evaluate, and compare 2 types of diatom-based indicators for Idaho streams: an observed/expected (O/E) ratio of taxon loss derived from a model similar to the River InVertebrate Prediction And Classification System (RIVPACS) and a multimetric index (MMI). Modeling the effects of natural environmental gradients on assemblage composition is a key component of RIVPACS, but modeling has seldom been used for MMI development. Diatom assemblage structure varied substantially among reference-site samples, but neither ecoregion nor bioregion accounted for a significant portion of that variation. Therefore, we used Classification and Regression Trees (CART) to model the variation of individual metrics with natural gradients. For both CART and RIVPACS modeling, we restricted predictors to natural variables unaffected by or resistant to human disturbances. On average, 46% of the total variance in 32 metrics could be explained by CART models, but the predictor variables differed among the metrics and often showed evidence of interacting with one another. The use of CART residuals (i.e., metric values adjusted for the effect of natural environmental gradients) affected whether or how strongly many metrics discriminated between reference and test sites. We used cluster analysis to examine redundancies among candidate metrics and then selected the metric with the highest discrimination efficiency from each cluster. This step was applied to both unadjusted and adjusted metrics and led to inclusion of 7 metrics in MMIs. Adjusted MMIs were more precise than unadjusted ones (coefficient of variation ~50% lower). Adjusted and unadjusted MMIs rated similar proportions of the test sites as being in nonreference condition but disagreed on the assessment of many individual test sites. Use of unadjusted MMIs probably resulted in higher rates of both Type I and Type II errors than use of adjusted metrics, a logical consequence of the inability of unadjusted metrics to distinguish the confounding effects of natural environmental factors from those associated with human-caused stress. The RIVPACS-type model for diatom assemblages performed similarly to models developed for invertebrate assemblages. The O/E ratio was as precise as the adjusted MMI, but rated a lower proportion of test sites as being in nonreference condition, implying that taxon loss was less severe than changes in overall diatom assemblage structure. As previously demonstrated for O/E measures, modeling appears to be an effective means of developing more accurate and precise MMIs. Furthermore, modeling enabled us to develop a single MMI for use throughout an environmentally heterogeneous region.

Key words: bioassessment, diatom assemblages, biotic indicators, multimetric indices, natural variability, predictive models, CART, RIVPACS.

Human activities have caused a range of alterations to the biota of freshwater ecosystems. Water resource managers need precise and accurate tools with which to measure biological condition to maintain or restore the biological integrity of these ecosystems. However, they cannot rely on indicators based on single assemblages to provide comprehensive assessments of biological integrity because different types of assemblages often respond to different types of disturbances (e.g., O'Connor et al. 2000, Soininen and

[1] E-mail addresses: yong.cao@usu.edu
[2] chuck.hawkins@usu.edu
[3] jrolson@cc.usu.edu
[4] mary-anne.kosterman@deq.idaho.gov

Kononen 2004, Griffith et al. 2005, Newall et al. 2006, D. M. Carlisle [USGS], CPH, M. R. Meador [USGS], M. Potapova [Academy of Natural Sciences], and J. Falcone [USGS], unpublished data). In the USA, most aquatic biological indicators are based on fish and macroinvertebrate assemblages (e.g., Barbour et al. 1999, Hawkins 2006). Much less effort has been devoted to development of indicators based on algal assemblages (Hill et al. 2000, Fore and Grafe 2003, Wang et al. 2005).

Work on algal assemblages generally has pursued development of 3 types of indicators (multimetric indices [MMI], tolerance indices, and multivariate predictive models), most of which are based on benthic diatoms, a subset of algal assemblages. Several researchers have developed MMIs to measure the overall condition of algal assemblages (e.g., Bahls 1993, KY DOW 1993, Hill et al. 2000, 2003, Fore and Grafe 2003, Coles et al. 2004, Griffith et al. 2005, Wang et al. 2005). Others have focused on tolerance indices that measure effects associated with specific types of stressors, such as nutrient addition or pH alteration (e.g., Dixit and Smol 1994, Pan and Stevenson 1996, Pan et al. 1996, 1999, Sonneman et al. 2001, Naymik and Pan 2005, Ponader et al. 2007, Potapova and Charles 2007). However, few researchers have explored the use of multivariate predictive models, such as River InVertebrate Prediction And Classification System (RIVPACS) (Moss et al. 1987), to assess alterations in algal assemblage composition (Chessman et al. 1999, Mazor et al. 2006, D. M. Carlisle, CPH, M. R. Meador, M. Potapova, and J. Falcone, unpublished data). No clear consensus exists regarding the utility or limitations of these 3 types of algal indicators, and additional work is needed to document and improve their performance.

We explored the performance (precision, bias, and responsiveness) of 2 types of indicators of the general biological condition of streams applied to benthic diatom assemblages: MMIs and an observed/expected (O/E) measure of taxonomic completeness derived from a RIVPACS-type model. These 2 approaches differ in how biotic data are summarized into an index, but both rely on a comparison of observed data with expectations derived from reference sites (Norris and Hawkins 2000, Stoddard et al. 2006). Reference sites often vary considerably among and within regions in naturally occurring features (e.g., climate, hydrology, geomorphology, and biogeochemistry) that can influence aquatic assemblages (Hawkins et al. 2000a, Leland and Porter 2000, Potapova and Charles 2002, Soininen et al. 2004). Therefore, accurate and precise biological assessments require a way to account for this natural variation when developing and applying biological indicators. RIVPACS-type models were developed specifically to predict assemblage composition under different naturally occurring environmental conditions, thereby allowing site-specific assessments of biological alteration (e.g., Wright et al. 2000). However, a priori spatial classifications (e.g., ecoregions, bioregions, stream order) typically are used to reduce natural variation in MMIs (Barbour et al. 1999).

More effective ways to partition the effects of natural gradients on the biotic metrics used in MMIs clearly are needed because the power of a priori classifications is often relatively weak (e.g., Hawkins et al. 2000a, Herlihy et al. 2006). Two recent studies used linear or logistic regression to account for natural variation in individual metrics with some success (Baker et al. 2005, Pont et al. 2006). However, the assumptions of both linear and logistic regression models often are violated in ecological analyses, and other statistical methods may provide more flexible and accurate models. One of these methods is Classification and Regression Trees (CART) (Breiman et al. 1984). CART models are often more precise and accurate than other types of models in predicting species occurrences and assemblage attributes (e.g., Rejwan et al. 1999, De'ath and Fabricius 2000, Karels et al. 2004, Bourg et al. 2005). Therefore, they might be effective in improving both the accuracy and precision of MMIs by partitioning the component of variability in biotic metrics that is associated with natural environmental factors.

We address 2 questions relevant to the use of diatom assemblages in biological assessments: 1) Can CART models be used to improve the performance of diatom MMIs? 2) How well does a RIVPACS-type model perform for diatoms in an environmentally heterogeneous region?

Data and Methods

Sample availability and comparability

We had access to 256 diatom samples from reference-quality sites. However, initial analyses of these samples showed that the taxonomies applied by the 3 different laboratories that processed the samples were not comparable. Therefore, we used only samples that had been processed by a single laboratory (n = 149) to develop indicators. We will report on the issue of taxonomic comparability in a separate publication because this issue has significant implications for indicator development and the reliability of subsequent indicator application.

Sampling design

Selecting reference sites.—The Idaho Department of Environmental Quality (ID DEQ) selected reference

sites using a multistep approach (ID DEQ 2004a). First, a list of candidate high-quality areas was compiled for each of 3 bioregions in Idaho (Northern Mountains, Central and Southern Mountains, and Basin) that were aggregations of different Level III ecoregions (Omernik 1987). Candidate areas had to satisfy a series of selection criteria (e.g., no known point discharges, no spills or other incidents, and low human population). A sampling site was selected in each candidate area, based on a combination of best professional judgment, inspection of satellite images, and landuse data. Local conditions were scored on several reach-level criteria (e.g., distance from roads, riparian vegetation, and stream channel morphology) during a field visit to each candidate site, and these data were used to screen the candidate sites further.

In total, 149 diatom samples (including site replicates) were collected from these reference sites during 1999 to 2003. We examined these reference samples for spatial independence. We delineated the watershed of each reference site, estimated the proportion of watershed area shared by adjacent sites, and identified any tributaries between them. We considered adjacent sites different if their watersheds overlapped by $<80\%$, tributaries occurred between them, or they were located >300 m from each other. These criteria yielded 88 sites that we regarded as different (Fig. 1), and we excluded the remaining sites. When replicate samples existed for individual sites, we randomly chose 1 sample for use in indicator development. Sixty-nine of the 88 samples had counts of 700 to 800 valves. We used these 69 samples to develop indicators and refer to them as reference calibration (RC) samples. The other 19 samples had lower counts (mean = 213 valves), and we set them aside for potential use in indicator validation, in case only nonrichness metrics were selected for the MMI. We did not use these reference validation (RV) samples for testing the RIVPACS-type model because O/E, the RIVPACS indicator, is sensitive to sample count.

Changes in assemblage structure over sampling years can confound development and interpretation of indicators. Neither nonmetric multidimensional scaling nor cluster analysis revealed a significant effect of sampling year on patterns of variation in diatom assemblages among sites. However, seasonal variation in assemblages was evident, so we used sampling date as a predictor for both MMIs and the RIVPACS-type model (Table 1).

Nonreference sites.—We used nonreference sites to calibrate and evaluate indicator performance. During 1999 to 2003, 155 samples were collected from sites that were noticeably disturbed by point-source discharges, grazing, or other stressors (test sites). Sampled

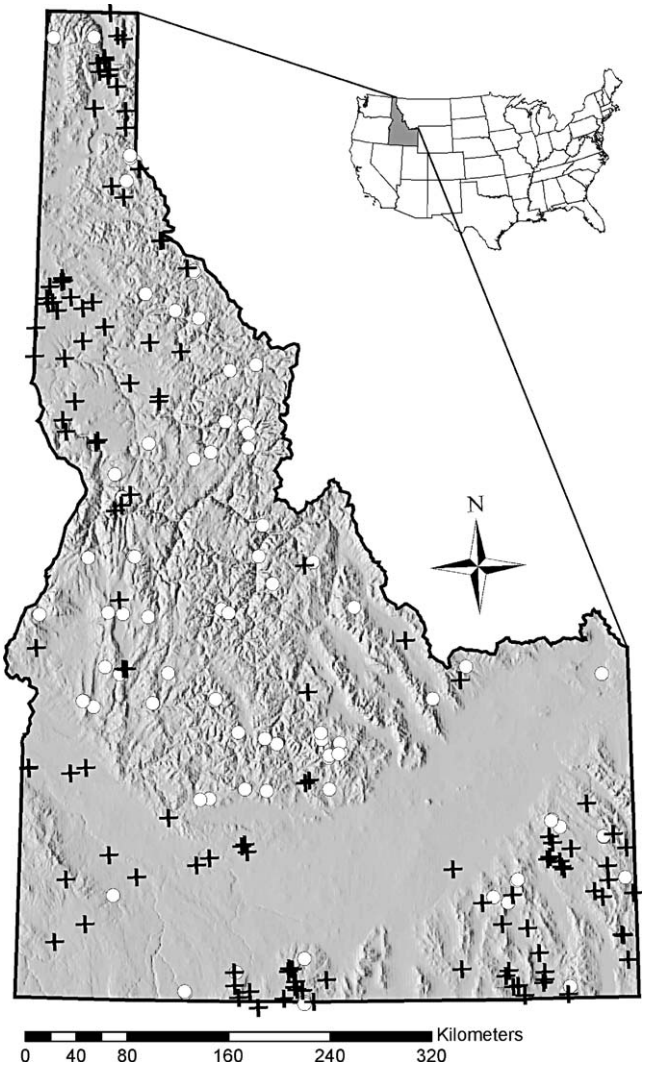

FIG. 1. Locations of 88 reference sites (white circles) and 125 test sites (crosses) in Idaho.

reaches were up to 30 channel-widths but not <100-m long. Screening for spatial independence and data completeness resulted in 125 distinct and valid test samples (Fig. 1). We randomly drew 100 of these samples to use for calibration of MMIs and refer to them as test calibration (TC) samples. We used the other 25 samples as test validation (TV) samples for index validation. We used all 125 samples to assess RIVPACS performance because the RIVPACS-type model does not rely on test sites for calibration.

Field sampling and laboratory processing.—Diatom samples were collected, and a range of habitat variables was measured at each sampling site with a standard field procedure (ID DEQ 2004b). Three riffle habitat units were randomly selected within the study reach. A stone was randomly chosen from each unit, and diatoms were removed from its surface with a

TABLE 1. Maximum (max), minimum (min), and mean values of 17 environmental variables used to partition natural variability of biotic metrics. RC = reference sites used for model calibration, RV = reference sites used for model validation, TC = test sites used for model calibration, TV = test sites used for model validation.

Predictive variable	Definition	Sources	RC (n = 69)			RV (n = 19)			TC and TV (n = 125)		
			Max	Min	Mean	Max	Min	Mean	Max	Min	Mean
DAY-OF-YEAR	Calendar day of sampling (1–365)	Field record	275	182	222.5	268	181	229	287	181	209.1
ELEV	Elevation (m)	Map	2451	640	1488	2489	697	1569	2096	329	1293
FIRST-32	Average calendar day of the first freeze	Map	276	223	248	263	230	246	277	234	256
GRADIENT	Channel slope at sampling sites (%)	Field record	15.9	0.40	3.3	15.0	0.5	4.6	12.0	0.2	2.1
HUMIDITY	Average annual relative humidity (%)	Map	71	51	61	69	54	62	69	46	60
LAST-32	Average calendar day of the last freeze	Map	201	132	168	196	143	171	184	127	154
LATITUDE	Decimal degrees	Map	48.823	41.980	44.597	48.809	42.840	44.635	48.994	41.946	44.591
LONGITUDE	Decimal degrees	Map	–111.313	–116.904	–114.810	–111.142	–116.957	–114.586	–111.072	–117.038	–114.697
LOG-CATCH	$\log_{10}$ (catchment area in acres)	Map	5.06	2.97	4.02	5.20	2.87	4.03	5.95	2.18	4.01
PRECIP	Mean annual precipitation (mm)	Map	1352	255	756	1035	386	658	1339	175	598
PRED-COND	Predicted conductivity (μS/cm)	Model	221	38	74	223	38	75	223	34	83
PRED-ALK	Predicted alkalinity (mg/L $CaCO_3$)	Model	331	39	96	334	39	97	334	32	111
ROCK-HARD	Predicted rock hardness (1–5)	Model	3.97	1.00	1.75	3.71	1.00	1.90	5.00	1.00	2.63
TMAX	Mean maximum monthly temperature (°C)	Map	16.9	5.6	12.0	16.2	8.5	11.8	19.4	8.3	13.5
TMEAN	Mean average monthly temperature (°C)	Map	9.9	–1.0	5.1	9.1	1.0	4.8	11.4	2.9	6.7
TMIN	Mean minimum monthly temperature (°C)	Map	3.8	–7.8	–1.8	2.1	–6.7	–2.2	4.0	–4.3	0.0
WETD	Days of measurable precipitation	Map	158	59	109	144	70	100	164	49	100

modified 30-mL syringe and a small, stiff-bristled brush. If no stone was available, a piece of submerged wood, debris, or other hard surface was used. The sample was placed in a 10-mL vial, and 2 drops of formalin were added to preserve it. Samples were sent to the Patrick Center for Environmental Research for processing. In the laboratory, samples were processed with the US Geological Survey (USGS) protocol (Charles et al. 2002). All individuals were identified and counted if <800 valves were available; otherwise an 800-valve subsample was taken. All valves were identified to the finest taxonomic level, usually species or variety. Any valve that could not be unambiguously identified was excluded from further analyses. In total, 622 taxa were recorded.

Field crews also measured or assessed a range of habitat variables, including flow, slope, water temperature, conductivity, proportions of pools and riffle habitat, riparian vegetation, and bank condition. However, no water-chemistry samples were collected. The lack of water-chemistry data was not critical to index development because water chemistry can easily be modified by human activities and thus, is not a robust predictor of the biota expected under reference conditions (Chessman et al. 1999, Hawkins 2006). We used values of alkalinity and conductivity predicted to occur under reference conditions (see *Deriving map and GIS data*) for index development and application.

Deriving map and GIS data.—We derived map and Geographic Information System (GIS) data for the watershed above each sampling site and included latitude, longitude, elevation, ecoregion, watershed size, watershed geology (predicted streamwater alkalinity and conductivity, and watershed rock hardness), watershed climate (temperature, precipitation, and humidity), and site climate (temperature, precipitation, and humidity). We obtained climate data from the Parameter-elevation Regressions on Independent Slopes Model and the 1961 to 1990 record (Daly and Taylor 2000). We estimated reference alkalinity and conductivity from models that related variation in these 2 variables to the CaO content of watershed geology. We estimated average rock hardness within each watershed from the uniaxial compressive strength of each lithology type within the watershed (J. R. Olson, Utah State University, unpublished models). The predicted values of alkalinity and conductivity are those expected at summer base flow under reference conditions and, thus, are potentially useful predictors of the biota expected under such conditions. Bedrock hardness is an important determinant of streambed stability, sediment yields, and channel morphology (e.g., Tooth et al. 2002, Sable and Wohl 2006), and we considered it potentially more useful for prediction than substrate characteristics measured on site, which also can be significantly altered by human disturbances.

MMI development

Candidate metrics.—We explored the utility of a large number of biotic metrics that previously were considered for diatom-based MMIs (Appendix). Many of these metrics were listed in Fore and Grafe (2003) and Wang et al. (2005). Some metrics appeared to be applicable only to specific regions (e.g., some in the Kentucky diatom index) and were excluded from our analysis. We evaluated 43 of the remaining metrics (Appendix), including 28 environmental tolerance or autecology metrics derived from those identified by Porter (2005; Algal Attribute Table, AlgalAttributes [version 7], USGS, Colorado), which have been used in several previous studies (e.g., Peterson and Porter 2002, Coles et al. 2004, S. D. Porter, Texas State University, personal communication). We also examined 5 assemblage-attribute metrics, 6 metrics based on specific taxa, and 4 metrics derived from Indicator Species Analysis (Dufrene and Legendre 1997), which identified those taxa that were overrepresented at either reference or test sites. One of the 5 assemblage-attribute metrics is a statistical estimator of total species richness (ACE; Chao and Lee 1992), which predicts the number of species occurring at a site from the abundances of species in individual samples. We also considered the expected direction of response to environmental stress as part of our metric evaluation process.

Using ecoregions and bioregions to partition natural biotic variability.—The reference sites fell into 8 of the 9 Level III ecoregions of Idaho (ID DEQ 2004a). We combined these ecoregions into 3 aggregate bioregions previously used in development of macroinvertebrate-based indicators (Jessup and Gerritsen 2000). We then estimated the classification strengths (CS) (Van Sickle 1997) of both regionalizations. We used Bray–Curtis similarities calculated from $\log(x + 1)$ data for these tests. Both classifications were statistically significant ($p < 0.001$), but the CS values were low (0.03 for ecoregions and 0.025 for bioregions), indicating that neither classification accounted for useful proportions of the natural variability in assemblage composition. Therefore, we did not pursue the use of these 2 classifications in developing MMIs.

Partitioning natural biotic variability with CART models.—CART models account for variation in a dependent variable by progressively splitting samples into 2 bins that best partition the total variation among samples. This process forms a prediction tree based on a series of binary splits in the data. The first split

TABLE 2. Summary of the Classification and Regression Tree (CART) models used to associate variation in biotic metrics with natural environmental features (predictors). Number of nodes and number of variables are measures of the complexity of the models. Pseudo-R^2 values measure the strength of the association between the metric and the predictors. Metric descriptions and abbreviations are in the Appendix. See Table 1 for explanation of predictor abbreviations.

Metric	Nodes	Pseudo-R^2	Predictors
A/AN-I	3	0.42	PRECIP, LATITUDE
A/AN-T	5	0.73	WETD, TMIN, HUMIDITY
ACID-I	3	0.46	PRECIP, WETD
ACID-T	3	0.63	PRECIP, ELEV
ALK-I	3	0.28	LATITUDE, ELEV
CYMB-I	6	0.48	ELEV, GRADIENT, PRED-COND, PRECIP, LONGITUDE
DOM	2	0.15	HUMIDITY
HIGH-O_2-I	3	0.50	PRECIP, TMAX
HIGH-O_2-T	5	0.67	PRECIP, ELEV, PRED-COND
LOW-O_2-I	5	0.21	TMEAN, GRADIENT, LONGITUDE, TMAX
LOW-O_2-T	3	0.34	TMAX, FIRST-32
MOB-I	4	0.41	TMEAN, DAY-OF-YEAR
MOB-T	7	0.60	TMAX, DAY-OF-YEAR, ELEV, LONGITUDE
MPSAP-SP	2	0.20	TMAX
MPSAP-T	2	0.45	LONGITUDE
NAVIC-T	8	0.68	LONGITUDE, ROCK-HARD, LOG-CATCH, TMEAN, WETD, TMAX, ELEV
NF-I	3	0.27	LATITUDE, ELEV
NHETER-I	4	0.49	TMAX, LONGITUDE, WETD
NHETER-T	3	0.86	TMAX, ELEV
OSAP-T	2	0.37	PRECIP
RICHNESS	3	0.30	ELEV, TMAX
REF-I	4	0.56	LATITUDE, LAST-32
REF-T	5	0.37	LONGITUDE, FIRST-32, TMEAN
SENS-I	5	0.43	TMEAN, TMAX, GRADIENT, ROCK-HARD
SENS-T	5	0.54	ELEV, FIRST-32, ROCK-HARD
SIMPSON	4	0.43	TMAX, ELEV, PRECIP
TOL-I	3	0.22	TMEAN, TMIN
WA-O_2	4	0.49	PRECIP, FIRST-32, ROCK-HARDNESS
WA-ORG-N	3	0.44	WETD, TMIN
WA-SAL	4	0.62	WETD, LAST-32, LOG-CATCH
WA-SAPRO	5	0.64	LATITUDE, LONGITUDE, PRECIP, TMIN
WA-TROPH	3	0.52	WETD, LAST-32, LOG-CATCH

occurs at the value of the predictor variable that most efficiently (as measured by the mean within-group standard deviation [SD]) partitions overall variation of the dependent variable into 2 groups. CART then partitions each of these 2 groups, if justified, into 2 smaller groups or nodes in the same manner, although the partitioning variable may differ. We built CART models with the R (version 2.2.1; R Development Core Team, http://www.r-project.org/) routine, tree, and then cross-validated the models to determine when to stop the splitting. Cross validation was based on subsampling in which 10% of samples were randomly withheld for validation. This process was repeated 1000 times. We then chose the number of nodes that yielded the lowest average cumulative errors. We further evaluated the likelihood of overfitting by randomly splitting the 69 reference sites into 2 groups, 49 for calibration and 20 for validation. If a model explains much less variance in the validation samples than the calibration samples when both are from the same population of stream sites, overfitting probably has occurred (Vaughan and Ormerod 2005). We applied this procedure to 2 metrics (% of taxa requiring high dissolved O_2 [HIGH-O2-T] and % of total individuals in those taxa [HIGH-O2-I]) that were selected to represent relatively complex (5 nodes) and simple (3 nodes) models, respectively.

After developing CART models (Table 2) with the 17 potential predictor variables (Table 1), we subtracted the predicted value of each metric from its observed value to obtain each sample's residual value, i.e., the variation remaining after accounting for natural factors. We refer to the residuals of metrics as the adjusted metrics, and we refer to MMIs developed based on adjusted metrics as adjusted MMIs.

Selecting metrics.—We selected metrics for use in unadjusted and adjusted MMIs using the procedure

described by Barbour et al. (1999) with some modifications as follows:

1. Testing the capability of a metric to separate reference sites from test sites. We used the nonparametric *U*-test to examine how well a metric separated the 69 RC samples from the 100 TC samples. Those metrics that failed this test ($p > 0.05$) were not considered further. We also excluded metrics that passed the *U*-test, but responded contrary to expectations. For example, taxa richness and diversity indices are generally expected to decrease at disturbed sites (e.g., Rapport et al. 1985, Karr and Chu 1998), but some of these metrics were higher at test than reference sites. It would be difficult to argue that such responses represented biological impairment, so we excluded them.
2. Assessing redundancy. A specific value of the Pearson correlation coefficient (e.g., $r = 0.8$) between metrics often is chosen as a threshold above which one or more metrics are dropped from consideration. This method works when identifying pairs of correlated metrics. However, when >2 metrics are highly correlated, it is difficult to decide which metric should be kept by considering pairwise correlations. Therefore, we used cluster analysis to identify groups of strongly correlated metrics. We measured metric similarities with the absolute value of Pearson's correlation coefficient (*r*), and we used Ward Linkage (Orloci 1967) as the clustering method.
3. Estimating discrimination efficiency. We measured discrimination efficiency (DE) as the % of TC sites that had a metric value below the lower 25th percentile of RC-site values if the metric was expected to decrease at test sites and as the % of TC sites with a metric value above the upper 75th percentile of RC-site values if the metric was expected to increase at test sites (Stribling et al. 2000).
4. Selecting metrics. We retained the metric that had the highest DE in each group of correlated metrics. Two sets of metrics were selected: one set consisted of adjusted metrics and the other consisted of unadjusted metrics.

Metric scoring.—We rescaled metric values in 2 ways using methods described by Blocksom (2003). In method A, TC- and RC-site values were kept separate. For metrics that decrease with disturbance, we set the 75th percentile of RC-site values as the maximum (max) and the 25th percentile of TC-site values as the minimum (min) values of the metric and rescaled the values at each site as 100 × (site value – min)/(max – min). For metrics that increase with disturbance, we set the 25th percentile of RC-site values as the minimum and the 75th percentile of TC-site values as the maximum values and rescaled the values at each site as 100 × (1 – [site value – min]/[max – min]). We treated rescaled values >100 or <0 as 100 or 0, respectively. In method B, TC- and RC-site values were combined. For metrics that decrease with disturbance, we set the 95th percentile of combined RC- and TC-site values as the maximum and the 5th percentile as the minimum values and rescaled the values at each site as described for method A. For metrics that increase with disturbance, we set the 5th percentile as the maximum and the 95th percentile as the minimum values and rescaled the values at each site as described in method A.

Developing a RIVPACS-type model

The procedure for developing RIVPACS-type models is described in detail elsewhere (Hawkins et al. 2000b, Wright et al. 2000). Therefore, we give only a brief description here.

Classification.—We used flexible-β Unweighted Pair-Group Method using Arithmetic Averages ($\beta = -0.5$) and the Bray–Curtis Index to cluster 69 RC samples. We transformed diatom counts as $\log(x + 1)$ to down-weight abundant taxa, and we excluded taxa recorded at <3 sites from the classification step. From this analysis, we identified 7 assemblage groups on which we based the modeling.

Predicting taxonomic composition.—We used the all-possible-subsets procedure of Van Sickle et al. (2006) to identify the variables in a discriminant function model that best predicted group membership and, thus, the taxonomic composition of RC sites. This procedure evaluated how every combination of predictor variables affected RIVPACS model accuracy and precision. We chose the most parsimonious model (fewest and most easily measured variables) that was both accurate (mean RC-site O/E value near 1) and precise (small SD for RC-site O/E).

Calculating O/E.—The procedures for estimating the number of taxa expected under reference conditions and calculating O/E (the proportion of expected taxa observed in a sample) are well documented (e.g., Hawkins et al. 2000b, Wright et al. 2000). The number of taxa expected at a site (E) under reference conditions is calculated as $E = \sum_{i=1}^{n} \sum_{j=1}^{m} M_i F_{ij}$, where M_i = the probability of a site belonging to reference group *i* as predicted from the discriminant function model ($0 \leq M_i \leq 1$), *n* = the number of groups for which $M_i > 0$, F_{ij} = the frequency of taxon *j* occurring in reference group *i*, and *m* = the number of taxa in group *i* with F_{ij} greater

than a specified value (0.5 in our study; Hawkins et al. 2000b). O is the number of taxa observed in a sample that meet the requirement for F_{ij}.

$\Sigma M_i F_j$ is the probability of capture (PC) of taxon j at a site given a standard sampling effort and method, and we calculated O/E with a threshold value of PC $\geq$ 0.5 (i.e., O/E_{50}), which often yields more precise assessments than PC $>$ 0 (Hawkins et al. 2000b, Ostermiller and Hawkins 2004). To evaluate the performance of the model, we compared the SD of RC-site O/E_{50} values derived from the model with the SDs estimated for a null model and for the best possible model, i.e., where the SD is related only to random sampling error (Van Sickle et al. 2005). The null model predicts that E will be the same everywhere, the sum of the proportional frequencies of occurrence of above-threshold taxa across all RC sites (i.e., it assumes no systematic variation exists in taxa occurrences among reference sites). The null and the best-possible models set the lower and upper boundaries of precision for a RIVPACS-type model.

Evaluating indicator performances

Indicator precision.—We estimated MMI precision as the coefficient of variation (CV) of index values observed at RC and RV sites. The lower the CV, the more precise an indicator is. We compared the CV of indicator values observed at RC and RV sites. If the CV was substantially higher at RV than at RC sites, the models used to account for natural variation might be overfitted. The CV is directly comparable to the SD used to evaluate RIVPACS precision because CV = SD when the mean is 1.

Indicator accuracy.—The true biological impairment at a site is typically unknown (Cao and Hawkins 2005). Therefore, evaluating the true accuracy or bias of any indicator is difficult. However, we can determine whether an indicator systematically over- or underestimates relative biological condition at sites occurring within different environmental settings. To do so, we regressed RC-site indicator values on the 17 environmental variables (Table 1).

Indicator responsiveness and comparability.—Different indicators might not lead to the same inference about biological impairment if they differ in either precision or responsiveness to stress. We evaluated similarity in indicator assessments in 2 ways. First, we calculated the proportion of TC and TV sites that would be inferred as impaired with each indicator at 3 different threshold values: the 5^{th}, 10^{th}, and 25^{th} percentiles of values observed at RC sites. Second, we determined the number of sites for which indicators differed in assessments. O/E values derived from a RIVPACS-type model and MMIs measure different attributes of biological condition; therefore, we did not expect assessments based on these 2 types of indicators always to be similar (Hawkins 2006).

Results

MMIs

Natural variability in metric values.—Most of the 43 candidate metrics were highly variable across the 69 RC sites (Appendix). CVs were between 1 and 3.13 for 14 metrics, and between 0.50 and 1 for 13 other metrics. The average CV was 0.9 among all candidate metrics. These results highlight the need to account for the natural variability of biotic metrics if we desire precise and accurate MMIs.

CART models.—We developed CART models for 32 of the 43 metrics (Table 2, see Fig. 2 for example). For the other 11 metrics, cross-validation indicated that none of the 17 predictor variables was consistently associated with variation in metric values. The number of nodes in the 32 CART models ranged between 2 and 8 (Table 2). Using CART residuals reduced the SD of metric values among calibration samples by an average of 46% (range = 15–86%). That is, a large amount of the observed variability in metrics across RC sites could be attributed to $\geq$1 environmental gradients. Furthermore, the response of a biotic metric to one variable often depended on other variables; i.e., interactions among variables were common. Elevation, annual precipitation, annual mean maximum air temperature, and longitude were the variables most frequently associated with variation in metric values (Fig. 3).

The cross-validation used to develop CART models appeared to be generally successful in minimizing overfitting. When we used 49 randomly drawn RC sites to develop reduced-sample CART models for the 2 evaluation metrics (% HIGH-O2-T and % HIGH-O2-I), the number of nodes and the predictive variables used were identical to those in the models based on all 69 RC sites (all-sample CART model). However, the pseudo-R^2 values were slightly lower for the reduced-sample models than for the all-sample model (HIGH-O2-T: 0.49 cf. 0.51, HIGH-O2-I: 0.69 cf. 0.70). The predicted value for each node also differed slightly between the all-sample and reduced-sample models. When we applied the reduced-sample CART models to the other 20 RC sites, pseudo-R^2 decreased from 0.49 to 0.37 (27%) for the HIGH-O2-T metric and from 0.70 to 0.46 (34%) for the HIGH-O2-I metric, indicating that some overfitting may have occurred in the all-sample models. Because our sample size was so small, we did not think these results were strong enough to warrant

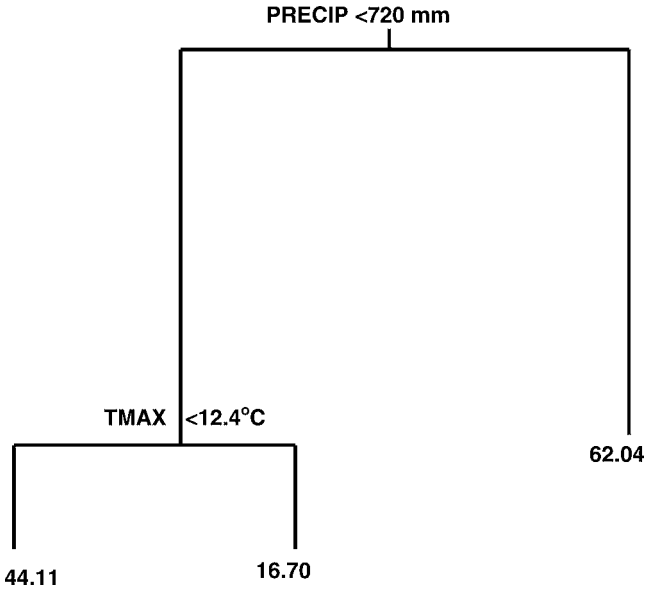

FIG. 2. Example of a Classification and Regression Tree (CART) model for the metric HIGH-O_2-I (% individuals requiring 100% O_2 saturation). CART models form prediction trees using binary splits to separate samples into bins that best partition the total variation among samples. In this example, the first split occurred at the value of the predictor variable (mean annual precipitation [PRECIP] = 720 mm) that most efficiently partitioned overall variation of the HIGH-O_2-I into 2 groups. If PRECIP at a site was <720 mm, the site was predicted into the group on the left, otherwise into the group on the right. The group on the left was further split at the value of the predictor variable (mean maximum monthly temperature [TMAX] <12.4°C) that partitioned the samples into 2 smaller groups in the same manner. A single metric value (at the bottom) is predicted for all samples within each final group.

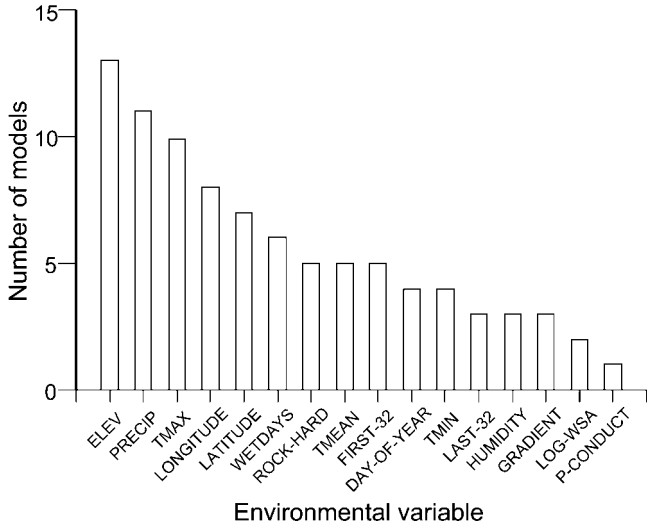

FIG. 3. Frequency histogram showing the number of Classification and Regression Tree (CART) models for which different environmental variables were selected as predictors. Full descriptions and explanations of abbreviations for predictors are shown in Table 1.

fitting models with fewer predictors to the data. However, once additional samples are available, these models should be re-evaluated for overfitting.

Metric discrimination between reference and test sites.—Values of 32 of the 43 unadjusted candidate metrics were significantly different between the RC sites and TC sites (*U*-test, $p < 0.05$; Table 3). Sample species richness, Simpson's Diversity Index, and Shannon's Diversity Index were higher at TC sites than at RC sites instead of being lower as expected. Dominance also was lower at TC sites than at RC sites, also contrary to expectation. We excluded these 4 metrics from further consideration because their response was difficult to interpret. ACE estimates of species richness were not significantly different between reference and test sites.

Of 32 adjusted metrics, 18 were significantly different between RC and TC sites. Two of the 18 were not significantly different between RC and TC sites in their unadjusted form (Table 3). The unadjusted responses of these 2 metrics to disturbance were apparently masked by the effects of strong natural environmental gradients. Ten of the 32 adjusted metrics were significantly different between RC and TC sites in their unadjusted but not their adjusted form, a result that implied that their apparent responses to human-caused stress emerged because stress gradients were strongly confounded with natural environmental gradients. Hence, these responses probably resulted from natural factors. Another 4 of the 32 adjusted metrics failed to separate RC from TC sites, as in their unadjusted forms. The CART adjustments also often reduced the magnitude of difference in metric values between RC and TC samples.

We excluded 3 of 18 significant adjusted metrics (species richness, Simpson's Diversity Index, and dominance) from further analysis for the same reasons as for the unadjusted metrics. We retained the remaining 15 adjusted metrics and 4 unadjusted metrics, REF-I, TEST-T, WA-POL, and TOL-T (see Table 2 for definitions) that were unresponsive to the natural gradients tested (i.e., no CART model was established) but discriminated between RC and TC sites, as candidates for further indicator development.

Metric redundancy and discrimination efficiency.—We classified the 28 candidate unadjusted metrics that distinguished RC sites from TC sites into 8 groups, each of which contained 1 to 6 metrics. The average within-group Pearson's r (for the groups with ≥ 2

TABLE 3. Probability (p) values associated with U-tests to determine the ability of metrics to discriminate between 69 reference-calibration (RC) and 100 test-calibration (TC) sites. Raw (unadjusted) metrics were adjusted to remove the effects of natural environmental gradients using Classification and Regression Tree (CART) models. See text for details. NA = natural gradients were not associated with metric values, U = unadjusted metric discriminated between RC and TC sites but adjusted metric did not, A = adjusted metric discriminated between RC and TC sites but unadjusted metric did not, B = both unadjusted and adjusted metrics discriminated between RC and TC sites, N = neither unadjusted nor adjusted metrics discriminated between RC and TC sites. See Appendix for full descriptions of metrics.

Metric	p-value: Unadjusted metric	p-value: Adjusted metric	Discriminate?
ALK-I	<0.05	>0.10	U
HIGH-O2-T	<0.01	>0.50	U
LOW-O2-I	<0.01	>0.05	U
MOB-I	<0.01	>0.10	U
NAVIC-I	<0.01	>0.05	U
NHETER-I	<0.01	>0.05	U
NHETER-T	<0.01	>0.10	U
OSAP-T	<0.01	>0.10	U
SENS-I	<0.01	>0.10	U
WA-TROPH	<0.01	>0.50	U
A/AN-I	>0.05	<0.05	A
CYMB-I	>0.10	<0.01	A
DOM[a]	<0.01	<0.05	B
HIGH-O2-I	<0.01	<0.05	B
LOW-O2-T	<0.01	<0.01	B
MOB-T	<0.01	<0.01	B
MPSAP-T	<0.01	<0.01	B
NAVIC-T	<0.01	<0.05	B
RICHNESS[a]	<0.01	<0.05	B
REF-I	<0.01	NA	B
REF-T	<0.01	<0.01	B
SENS-T	<0.01	<0.01	B
SHANNON[a]	<0.01	NA	B
SIMPSON[a]	<0.01	<0.01	B
TEST-I	<0.01	<0.01	B
TEST-T	<0.01	NA	B
TOL-T	<0.01	NA	B
WA-POL	<0.01	NA	B
TOL-I	<0.01	<0.01	B
WA-O2	<0.01	<0.01	B
WA-ORG-N	<0.01	<0.01	B
WA-SAL	<0.01	<0.05	B
WA-SAPRO	<0.01	<0.05	B
A/AN-T	>0.05	>0.50	N
ACE	>0.10	NA	N
ACID-I	>0.10	>0.10	N
ACID-T	>0.10	>0.50	N
ALK-T	>0.10	NA	N
CYMB-T	>0.10	NA	N
MPSAP-I	<0.05	NA	N
NF-I	>0.05	>0.05	N
NF-T	>0.10	NA	N
OSAP-I	>0.10	NA	N

[a] Metrics whose responses to disturbance were contrary to expectations

metrics) ranged from 0.49 to 0.86, indicating substantial redundancy. The DE of these metrics varied from 32 to 80%. We excluded ALK_I (see Table 2 for definition), the only metric in its group, from further consideration because of its low DE (37%). We then selected the metric with the highest DE in each of the other 7 groups for use in the MMI. The DE of the 7 selected metrics ranged between 48 and 75% for the 100 TC sites and 52 and 72% for the 25 TV sites. The mean Pearson's r among the selected metrics was 0.42 (range 0.21–0.68).

We also classified the 19 candidate adjusted metrics into 7 groups. Within-group correlations ranged between 0.38 and 0.71, indicating less redundancy among adjusted metrics than among unadjusted metrics. The DE of these metrics ranged between 33 and 76%. The DE of the 7 metrics selected for use in the MMI ranged between 43 and 76% for the 100 TC sites and 40 and 64% for the 25 TV sites. The mean Pearson's r among the 7 metrics was 0.29 (range 0.04–0.63). Because 3 of these 7 metrics were unadjusted, the adjusted indices actually were hybrids of adjusted and unadjusted metrics, but for convenience, we refer to the MMIs based on these metrics as adjusted MMIs.

The effects of the tree regressions on the scoring of metrics were substantial at many sites. For example, the difference between the values of the unadjusted and adjusted HIGH-O2-I was as large as 70 points on a 100-point scale at some sites, although the average difference was not as substantial (Fig. 4). CART-based adjustments generally reduced the variability of the metrics across the reference sites, but increased the variability among test sites.

Precision.—The CVs of the unadjusted MMIs based on scoring methods A (unadjusted MMI-A) and B (unadjusted MMI-B) for the 69 RC sites were high (0.27 and 0.21, respectively), indicating high natural variability. The CVs of unadjusted MMI-A and MMI-B for the 19 RV sites were even higher (0.31 and 0.24, respectively). In comparison, the adjusted MMIs were much more precise. The CVs of adjusted MMI-A and MMI-B for the RC sites were 0.14 and 0.13, respectively, whereas the CVs of adjusted MMI-A and MMI-B for the RV sites were slightly higher at 0.17 and 0.15, respectively. The differences in CVs between calibration and validation samples were smaller for the adjusted MMIs than for the unadjusted MMIs. This result suggests that, even if the models for individual metrics were somewhat overfit, any problems caused by overfitting were small relative to the gains in indicator performance achieved by modeling.

Bias.—Values of the 2 adjusted MMIs were not significantly correlated with any of the 17 natural environmental variables ($p > 0.05$), a result that

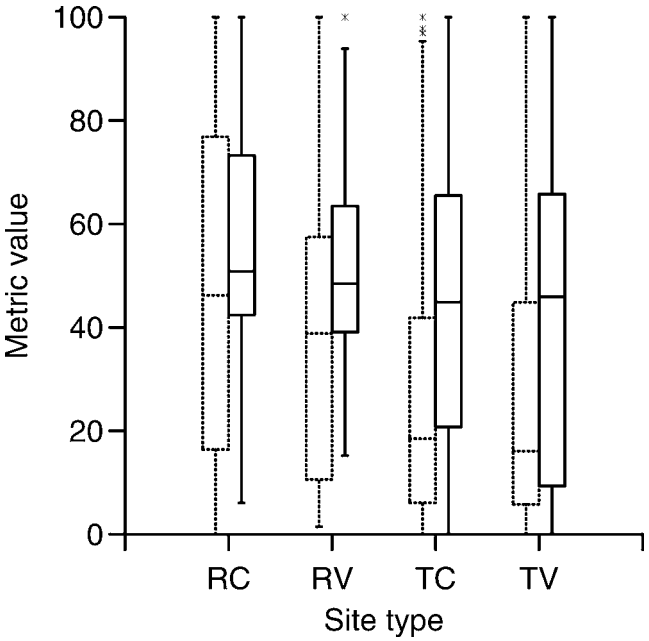

FIG. 4. Box-and-whisker plots of values of the metric HIGH-O_2-I (% individuals requiring 100% O_2 saturation) at reference (reference calibration [RC], reference validation [RV]) and test (test calibration [TC], test validation [TV]) sites. Dotted lines show the distributions of values for the unadjusted metric, and solid lines show values for the metric after adjusting them for the effect of natural environmental gradients using a Classification and Regression Tree (CART) model. Boxes encompass the interquartiles (25^{th}–75^{th} percentiles), small bars are means, stars are outliers, and range bars show the maximum and minimum values excluding outliers.

suggests that the adjusted MMIs should not systematically over- or underestimate condition under different environmental settings. In contrast, values of the 2 unadjusted MMIs increased significantly with annual precipitation and the number of wet days per year ($p < 0.05$) and tended to increase with latitude and decrease with maximum mean annual air temperature. These results suggest that the unadjusted MMIs probably would overestimate biological condition at sites in the wetter, northern regions of the state (Type I error), but underestimate biological condition at sites in the drier, southern part of the state (Type II error).

Responsiveness and comparability.—The apparent responsiveness to stress was similar among the 4 MMIs as measured by the % of TC and TV sites with values below the 5^{th}, 10^{th}, and 25^{th} percentiles of RC-site values (Table 4). However, the apparent comparability between the unadjusted and adjusted MMIs occurred mostly because precision and bias affected inferences regarding the condition (reference or not) of a site in opposite ways. For example, the unadjusted MMIs were far less precise than the adjusted MMIs, a difference that generally would cause fewer test site values to fall below a given percentile-defined impairment threshold. However, the unadjusted MMI values for many test sites (TC and TV) were so low that they were below the statistical threshold despite their low precision.

Improving precision by adjusting for natural gradients also reduced bias associated with the confounding effects of natural gradients on MMI values. Removing bias in these TC samples resulted in upward adjustment of many index values, although it did not substantially affect the % of TC sites rated in nonreference condition at a specific threshold. As a result, the distribution of unadjusted MMI values tended to be left-skewed (Fig. 5A, C) as compared with the distribution of adjusted MMI values (Fig. 5B, D). Furthermore, adjusted and adjusted MMI-A differed in their assessments of 16% of the test sites (TC and TV),

TABLE 4. Precision (coefficient of variation [CV]) and apparent responsiveness (% of test sites considered impaired) of 4 multimetric indices (MMI; based on unadjusted or adjusted metrics and scaling method A or B) and an observed/expected (O/E) measure of taxonomic completeness with probability capture = 0.50 (O/E_{50}). Unadjusted metrics were adjusted to remove the effects of natural environmental gradients using Classification and Regression Tree (CART) models. Apparent responsiveness was measured for 3 statistical threshold values that were derived from the distribution of reference-site values. RC = reference sites used for model calibration, RV= reference sites used for model validation, TC = test sites used for model calibration, TV = test sites used for model validation. RV samples were not used to validate the RIVPACS-type model.

			Apparent responsiveness								
	CV		5^{th} percentile			10^{th} percentile			25^{th} percentile		
Index	RC	RV	TC	TV	RV	TC	TV	RV	TC	TV	RV
Unadjusted MMI-A	0.27	0.31	54	60	16	60	68	21	78	80	26
Adjusted MMI-A	0.14	0.17	59	60	11	62	68	16	75	72	37
Unadjusted MMI-B	0.21	0.24	53	56	11	59	68	16	81	80	26
Adjusted MMI-B	0.13	0.15	59	64	5	62	64	11	71	68	21
O/E_{50}	0.17		12	12		19	20		38	48	

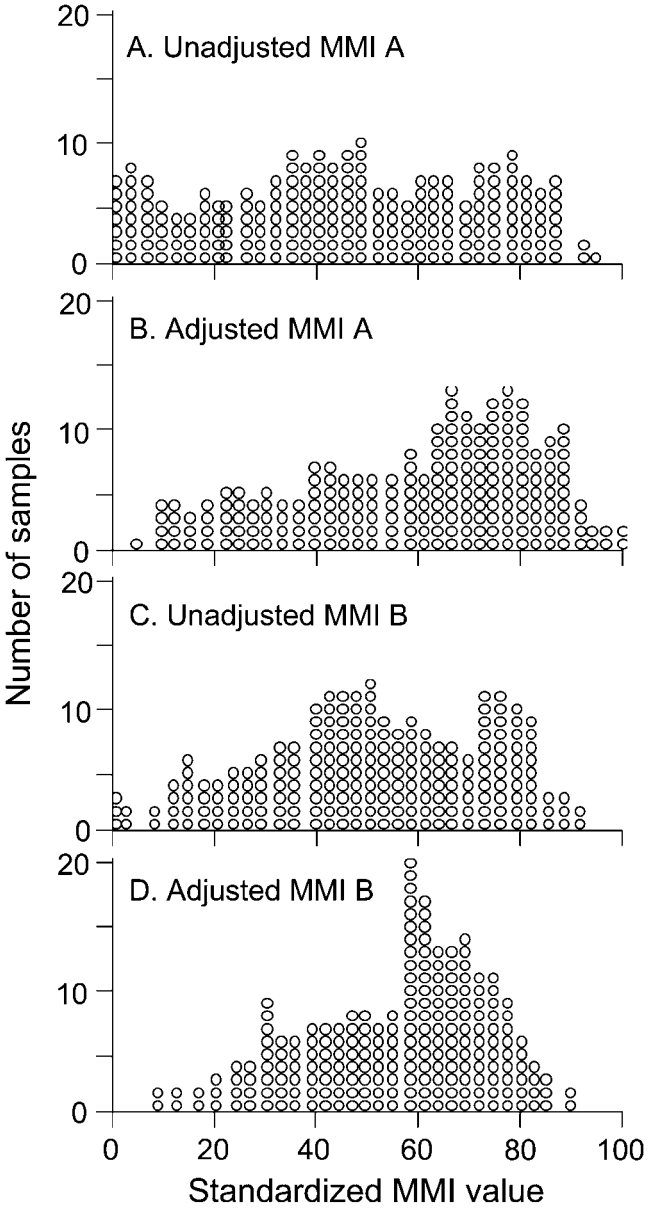

FIG. 5. Frequency distributions of values for the 4 multimetric indices (MMI) at 125 test sites. MMIs were calculated using 2 scaling methods (A and B), before and after adjusting metric values for the effect of natural environmental gradients using a Classification and Regression Tree (CART) model. A.—Unadjusted values calculated using scaling method A. B.—Adjusted values calculated using scaling method A. C.—Unadjusted values using calculated scaling method B. D.—Adjusted values calculated using scaling method B.

and adjusted and unadjusted MMI-B differed in their assessments of 17% of the test sites. Because the test sites were selected for their conspicuous and relatively severe environmental alterations, we expected most indicators, including imprecise and biased ones, to lead to similar inferences at these sites. However, we would expect adjusted and unadjusted MMIs to disagree more often when applied to less severely stressed sites. The scores of the adjusted MMI-A at TC sites (Fig. 5B) were noticeably more variable than the scores of the adjusted MMI-B at TC sites (Fig. 5D), implying that method A may be more capable of differentiating levels of biological impairments than method B.

RIVAPCS-type model

Classifications of reference sites.—The 69 RC sites were classified into 7 reasonably distinct groups (classification strength = 0.31). The number of sites in a group ranged between 5 and 15. Log-scaled catchment size, predicted alkalinity, and mean maximum monthly temperature were the best predictors of taxa occurrences.

O/E precision.—The CV of O/E_{50} for RC samples was 0.17, slightly higher than the CV of the 2 adjusted MMIs (0.14–0.15), but lower than the CVs of the 2 unadjusted MMIs (0.21–0.27). Model-based O/E precision was intermediate between those of the null model (CV = 0.23) and the best possible model (CV = 0.11), indicating this model improved performance by ~50%.

O/E responsiveness.—O/E_{50} rated 38 and 48% of the TC and TV sites, respectively, in nonreference condition when the 25th percentile of RC-site values was used as the threshold, 19 and 20% of the TC and TV sites, respectively, in nonreference condition when the 10th percentile was used, and 12 and 12% of the TC and TV sites, respectively, in nonreference condition when the 5th percentile was used (Table 4).

Discussion

Partitioning the effects of natural environmental gradients on indicators

The primary challenge we faced in developing diatom-based indicators for Idaho was adjusting for the extreme range in naturally occurring environmental conditions within the state, a situation typical of many other western states. Adjusting for these natural gradients was the key to both precise and accurate indicators. Our study also provided insight into several other issues that are relevant to the development, application, and interpretation of biological indicators.

Natural variability of most biotic metrics is often substantial, as we found in our study (Appendix). Finding effective ways to reduce the effect of naturally varying factors on ecological indicators has been a

focus of bioassessment for decades (e.g., Omernik 1987, Hawkins et al. 2000a). Classification by ecoregions has been the most widely used approach for adjusting for the natural variability in biotic metrics. However, in our analysis, the classification strength associated with ecoregion and bioregion was extremely low and was substantially lower than that achieved by the biotic classification used in the RIVPACS model. In this respect, our results are consistent with those of Pan et al. (1999, 2000) and Mazor et al. (2006), although others have found that ecoregion classifications can account for meaningful amounts of the natural variability in stream assemblages (e.g., Heino et al. 2002, Soininen et al. 2004). Given these inconsistent results, we think that investigators who develop indicators should carefully scrutinize the assumptions that often are made implicitly when ecoregion classifications or other geographically dependent regionalizations are used.

Use of geographic stratifications, such as ecoregion and stream order, implicitly assumes that the classification partitions the major environmental gradients that affect biological assemblages and that the variation between assemblage groups is substantially greater than within-group variation. We think this assumption is generally unrealistic, although it might be valid in certain cases. The environmental gradients that most strongly affect biological assemblages probably will vary with locality, spatial scope of the area of interest, and the type of biological assemblage examined. Any discrete regionalization scheme almost certainly will be messy in that not all biologically important factors covary in the same way. For example, sites within the Central-Southern Mountain bioregion in Idaho share several general environmental attributes, but annual precipitation varies markedly within the region and is a factor that was strongly associated with among-site variation in many metrics in our study. In regions dominated by complex environmental gradients, we think it is conceptually more logical and technically more practical to account separately for the effects of different environmental gradients on individual metrics through modeling than to account for the effects of multiple gradients on all metrics simultaneously through regionalizations (e.g., ecoregion) or stratifications (e.g., stream order).

CART offered a conceptually simple yet powerful way to model the complex relationships between natural environmental gradients and biological assemblages. Its advantages as compared to other modeling techniques include: 1) its ability to use a wide range of data types including numerical, categorical, and ranked data (De'ath and Fabricius 2000), 2) its ability to replace missing data with a closely correlated variable (Karels et al. 2004), 3) its freedom from assumption about the relationship between the response variable and the predictive variables (Rejwan et al. 1999), 4) its automatic identification of interactions among variables (Rejwan et al. 1999), and 5) model output that is easy to interpret (De'ath and Fabricius 2000). These attributes are particularly important for assemblage-based bioassessments, where data types are diverse, missing data are frequent, the relationships between metrics or indicators and environmental gradients are often nonlinear or unknown, and interactions among different environmental gradients are common.

With CART models, we were able to create comparable, site-specific expectations for metric values by adjusting for unique combinations of environmental factors. This procedure is conceptually similar to what RIVPACS-type models do—i.e., make site-specific predictions of the biota expected under reference conditions. Such site-specific adjustments improve precision by reducing the variability in both individual and combined (MMI) metric values across reference sites (Fig. 4, Table 3) and yield more accurate predictions of the biological expectations for test sites.

Confounding of natural gradients with human-caused stressor gradients is especially problematic for metric selection and interpretation. For example, agriculture, a land use sometimes associated with substantial stress to stream ecosystems, often decreases with increasing elevation. Thus, a metric that is strongly correlated with land use actually may be responding to the natural elevation gradient. The change in discrimination between reference sites and test sites for many metrics after adjustment for natural gradients (Table 3) strongly supports the argument that such metrics probably are not actually responsive to disturbances. Modeling should help to identify truly responsive metrics. Furthermore, responses of metrics to human disturbance gradients can be obscured by opposing responses to natural gradients. Modeling should help identify the metrics that provide unique biological signals to disturbance.

The use of models also made it possible to develop an MMI based on the same metrics and scaling throughout the state, thereby avoiding one of the drawbacks (uncertain consistency in what they measure) associated with more traditional MMIs. Metric responses to disturbance should be more consistent across regions after adjustment for the effects of natural environmental gradients. Thus, we probably can avoid the need to use different metrics, which will almost certainly have been selected and calibrated on different criteria, for assessing general disturbances in

different regions. The use of the same or similar metrics within states and across regions certainly would improve the comparability of bioassessments, although development of stressor-specific (diagnostic) indicators might require different metrics if regions differ in their dominant stressors. The development of an MMI requires that each classification stratum have a certain minimum number of reference sites. Modeling relaxes this major constraint as long as the reference sites used in modeling adequately characterize the range of natural environmental gradients in the region of interest.

The use of CART or other models to develop MMIs has distinct advantages. However, potential pitfalls also exist. A major concern is model overfitting (McKenzie et al. 2000). The CART routine that we used provided cross-validation procedures based on random withholding of a specified proportion of samples (10% in our case). In our study, cross-validation almost always resulted in a much smaller tree than did the nonvalidated model. However, robust cross-validation requires sufficiently large sample sizes that the randomly selected subsets of data do not strongly overlap. When sample sizes are large enough, the data set should be split into independent calibration and validation data sets. We were able to make only limited use of cross-validation because we had a relatively small number of reference sites. Another possible way to avoid overfitting is to use the recently developed Random–Forest regression techniques (Breiman 2001), a bootstrap-like version of CART that presumably is resistant to overfitting. We are in the process of exploring this tool. Even if CART models for individual metrics are not significantly overfit, the MMI might still be overfit if slight overfitting in individual metrics is cumulative. In our analysis, the overfitting of the 2 adjusted MMIs appeared to be modest because the CV of the adjusted MMI values was only slightly higher for the RV sites than for the RC sites.

None of the CART models that we developed explained all of the variability in RC-site metric values. Indeed, only small amounts of the total variance were explained for some metrics. CART models might not always account for high proportions of variation for several possible reasons. First, sampling error (the variability among replicate samples) might have accounted for much of the remaining variation, but we do not have estimates of sampling error for these data. Second, samples were collected during different years, and we did not address the effects of annual variation in climate and discharge on metric values (although there was no year-to-year signal apparent from the ordination results). Third, we certainly did not explore an exhaustive list of factors known to influence diatom assemblages, and some of the variation we observed may have been associated with unmeasured environmental factors, such as nutrient and flow regimes. The use of direct measures of such variables for bioassessment is problematic because their values are affected by human disturbance. Indentification of surrogates for these types of factors might result in more robust models in the future. Fourth, the number of species that could be assigned tolerance values varied considerably among sites because the responses of many species to disturbances are still largely unknown. Thus, some assemblages were less well characterized in terms of their environmental tolerances than other assemblages, and uneven characterization of tolerance could have produced a biased estimate of the true, overall tolerance profile. This source of error has rarely been addressed in the literature and requires attention. In general, we need more refined information regarding the tolerance of diatom species to different stressors. Alternatively, we could use indices, such as O/E, that are based solely on composition. However, O/E may be especially sensitive to inconsistent taxonomic assignments, a problem that we encountered and that has not been well documented in the diatom literature.

Expected responses of metrics

The responses of many biotic metrics to general human-disturbance gradients are well documented (e.g., Rapport et al. 1985, Karr and Chu 1998, Barbour et al. 1999). However, several metrics, including species richness, Shannon's Index, and Simpson's Index, increased at disturbed sites rather than decreasing as generally expected. Similar observations were reported by others (e.g., Chessman et al. 1999, Stevenson and Pan 1999, Naymik et al. 2005, Wang et al. 2005). Different explanations, including those based on Intermediate Disturbance Theory (e.g., Townsend et al. 1997, Yamamoto and Hatta 2004), nutrient-supply hypotheses (Stevenson and Pan 1999, Naymik et al. 2005), and sampling artifact (Mackey and Currie 2000, Cao and Hawkins 2005), have been suggested for these responses. Chessman et al. (1999) also suggested that the presence of unexpected diatom species at a site might be a more reliable indicator of disturbances than the absence of an expected species. Nevertheless, the value of a metric as a biological indicator will be restricted if its response is difficult to generalize.

Metric redundancy

Most advocates of MMIs agree that redundancy among metrics should be minimized when metrics are

selected for inclusion in a MMI (Barbour et al. 1999). However, the literature provides limited insight into the problems of defining redundancy and determining an acceptable level of redundancy. Redundancy can be defined in either statistical or biological terms (Karr et al. 1986, Lewis et al. 2001). If $\geq$2 metrics respond to a disturbance gradient in highly similar ways, these metrics provide a similar statistical signal. If identifying nonredundant statistical signals is the main objective in metric selection, only 1 of these correlated metrics should be used in a MMI. However, 2 very different biological attributes might show a similar statistical response across a stressor gradient. In this case, we might want to ignore the high correlation and use both metrics because they represent independent biological signals (Karr and Chu 1998). Selecting a set of metrics that maximizes biological signal and minimizes statistical redundancy is not easy, especially if we want the same metrics to provide broad coverage of the overall biological structure and function present across sites. Previous studies have used rather rough rules of thumb based on the values of simple correlation coefficients (r) to identify redundant metrics. For example, Fore and Grafe (2003) defined metrics with an r value $>$0.75 as redundant, and Paul et al. (2005) used a threshold r value of 0.8. We suggest that using such thresholds might not be adequate when the correlation structure among candidate metrics is complex, as we showed in our data set. Cluster analysis and ordination are capable of revealing complex data structure and appear to be promising ways to identify groups of statistically covarying metrics. Only expert knowledge of biology and ecology will enable decisions regarding the biological independence of metrics.

Metric scoring

A variety of methods exists for rescaling metric values (e.g., Blocksom 2003, Wang et al. 2005). However, the biological implications of these methods have rarely been examined. Blocksom (2003) showed (and we observed) that the method used to scale metrics does matter and could affect our interpretation of the degree of impairment that exists at a site. Blocksom (2003) recommended use of method B, which scales values between the 5th percentile and 95th percentile of all site values, because an MMI scaled in this way was more strongly correlated with a principal components analysis axis that represented major disturbance and was more precise than MMIs scaled in other ways. However, this method tended to bunch MMI values in our analysis (Fig. 5C, D). In comparison, method A, which scales values between the 75th percentile of reference-site scores and the 25th percentile of test-site scores, tended to spread values out from the center of the distribution of values (Fig. 5A, B). Moreover, we suspect that method B might be more sensitive than method A to the relative proportions of reference and test sites used in the scaling because it includes test sites in setting both the maximum and minimum values of a metric. From our perspective, method A appears to have slight advantages as compared to method B.

Methods A and B both scale values on the basis of observations at reference and test sites, whereas O/E is scaled only on the basis of observations at reference sites. Therefore, we might expect O/E to appear to be less sensitive to stress in a given region, all other things being equal. The results of our study are consistent with that expectation. However, given that the 2 indicators measure different attributes of biological assemblages, other factors could easily affect the responsiveness of these 2 types of indicators to stress. These factors include the overall environmental setting at a site and the specific stressors present in a region.

RIVPACS-type models for diatom assemblages

RIVPACS-type models have been widely used for macroinvertebrate-based bioassessments (e.g., Barbour et al. 1999, Wright et al. 2000, Hawkins 2006), but their application to diatom assemblages has been limited. Chessman et al. (1999) developed a model based on genus-level data and concluded that its performance was compromised by high temporal variability in diatom assemblages and inability to use water chemistry as a predictor. The 1st constraint might be overcome by including the sampling date as a predictor, as we did in the present study. The 2nd problem can be addressed by using surrogates for at least some water-chemistry variables (i.e., geology) or by making predictions based on watershed geology as we did here. Despite the challenges inherent in precise modeling of diatom assemblages, our RIVPACS-type model was as precise (CV = 0.17) as many macroinvertebrate and fish models (e.g., Hawkins 2006). It also was similar in precision to a diatom model developed for Appalachian streams (SD = 0.16) by D. M. Carlisle, CPH, M. R. Meador, M. Potapova, and J. Falcone (unpublished data). Issues of precision notwithstanding, diatom assemblages do not appear to lose taxa as readily as invertebrate assemblages in response to general stress. Therefore, O/E might be less useful as an indicator of diatom-assemblage condition than the more-responsive MMIs we developed. However, the use and interpretation of any indicator should be

driven primarily by consideration of the biological attribute it measures. If biodiversity is an important aspect of overall biological integrity, then taxon loss is a useful measure. Assessment of the relative degree of taxon loss across algal, invertebrate, and fish assemblages in response to stress enables us to describe more completely the overall biological condition of a given ecosystem.

Summary comments regarding the overall performances of indicators

The need to adjust expectations for natural background conditions seems inescapable and is generally recognized as central in bioassessment (Moss et al. 1987, Karr and Chu 1998). Derivation of site-specific expectations by modeling natural gradients is a traditional strength of RIVPACS-type assessments. Our success in developing a single, reasonably precise model for diatoms that is applicable to wadeable streams in Idaho demonstrates that O/E-type indicators are useful in assessing the condition of algal assemblages, just as they are for invertebrates and fish. Historically, adjustments of expectations for MMIs were made by identifying regions or water-body types within which similar biota were expected. This approach may have an advantage in appearing to be technically simpler to implement than modeling, but the evidence is now compelling that it is not always effective. Our results and those of Pont et al. (2006) clearly show that MMIs can benefit significantly from directly modeling the response of individual metrics to natural gradients. Regardless of the type of indicator, we think that the benefits of modeling (improved accuracy and precision) far outweigh whatever costs might be associated with construction and application of the models.

Acknowledgements

We thank Stephen Porter for providing the table of diatom tolerance values. Jan Stevenson, Don Charles, Leska Fore, and Mike Munn gave advice and commented at different stages of the project. John Van Sickle, Lester Yuan, and Jeff Ostermiller offered useful suggestions regarding the use of CART. This manuscript was further improved by the comments of 2 anonymous referees and Bruce Chessman. The study was funded primarily by a contract with the Idaho Department of Environmental Quality (IDEQ). Cyndi Grafe, previously with the IDEQ, managed early phases of this project. Completion of the work was facilitated by funds from EPA Science To Achieve Results (STAR) grants R-82863701 and R-82863701.

Literature Cited

BAHLS, L. L. 1993. Periphyton bioassessment methods for Montana streams. Department of Health and Environmental Sciences, Water Quality Bureau, Helena, Montana. (Available from: Department of Health and Environmental Sciences, Water Quality Bureau, 1400 Broadway, Helena, Montana 59620 USA.)

BAKER, E. A., K. E. WEHRLY, P. W. SEELBACH, L. WANG, M. WILEY, AND T. SIMON. 2005. A multimetric assessment of stream condition in the Northern Lakes and Forests Ecoregion using spatially explicit statistical modeling and regional normalization. Transactions of the American Fisheries Society 134:697–710.

BARBOUR, M. T., J. GERRITSEN, B. D. SNYDER, AND J. B. STRIBLING. 1999. Rapid bioassessment protocols for use in streams and wadeable rivers: periphyton, benthic macroinvertebrates, and fish. 2nd edition. EPA 841-B-99-002. Office of Water, US Environmental Protection Agency, Washington, DC.

BLOCKSOM, K. A. 2003. A performance comparison of metric scoring methods for a multimetric index for Mid-Atlantic Highlands streams. Environmental Management 31:670–682.

BOURG, N. A., W. J. MCSHEA, AND D. E. GILL. 2005. Putting a CART before the search: successful habitat prediction for a rare forest herb. Ecology 86:2793–2804.

BREIMAN, L. 2001. Random forests. Machine Learning 45:5–32.

BREIMAN, L., J. H. FRIEDMAN, R. A. OLSHEN, AND C. G. STONE. 1984. Classification and regression trees. Wadsworth Statistics/Probability series. Chapman and Hall, New York.

CAO, Y., AND C. P. HAWKINS. 2005. Simulating biological impairment to evaluate the accuracy of ecological indicators. Journal of Applied Ecology 42:954–965.

CHAO, A., AND S.-M. LEE. 1992. Estimating the number of classes via sample coverage. Journal of the American Statistical Association 87:210–217.

CHARLES, D., C. KNOWLES, AND R. S. DAVIS. 2002. Protocols for the analysis of algal samples collected as part of the U. S. Geological Survey National Water-Quality Assessment Program. Report No. 02-06. Patrick Center for Environmental Research, Academy of Natural Sciences, Philadelphia, Pennsylvania. (Available from: Patrick Center for Environmental Research, Academy of Natural Sciences, 1900 Benjamin Franklin Parkway, Philadelphia, Pennsylvania 19103-1195 USA.)

CHESSMAN, B., I. GROWS, J. CURREY, AND N. PLUNKETT-COLE. 1999. Predicting diatom communities at the genus level for the rapid biological assessment of rivers. Freshwater Biology 41:317–331.

COLES, J. F., T. F. CUFFNEY, G. MCMAHON, AND K. M. BEAULIEU. 2004. The effects of urbanization on the biological, physical, and chemical characteristics of coastal New England streams. U.S. Geological Survey Professional Paper 1696. US Geological Survey, Denver, Colorado. (Available from: Information Services, US Geological Survey, Box 25268, Denver Federal Center, Denver, Colorado 80225 USA.)

DALY, C., AND G. TAYLOR. 2000. United States average annual precipitation, maximum temperature, mean temperature, minimum temperature, relative humidity, mean date of first 32F temperature in autumn, mean date of last 32F temperature in spring, days with measurable precipitation, 1961–90. Spatial Climate Analysis Service, Oregon State University, Corvallis, Oregon. (Available from: http://www.ocs.oregonstate.edu/prism/index.phtml)

DE'ATH, G., AND K. E. FABRICIUS. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81:3178–3192.

DIXIT, S. S., AND J. P. SMOL. 1994. Diatoms as indicators in the Environmental Monitoring and Assessment Program-Surface Water (EMAP-SW). Environmental Monitoring and Assessment 31:275–306.

DUFRENE, M., AND P. LEGENDRE. 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecological Monographs 67:345–366.

FORE, L. S., AND C. GRAFE. 2003. Using diatoms to assess the biological condition of large rivers in Idaho (U.S.A.). Freshwater Biology 47:2015–2037.

GRIFFITH, M. B., B. H. HILL, F. H. MCCORMICK, P. R. KAUFMANN, A. T. HERLIHY, AND A. R. SELLE. 2005. Comparative application of indices of biotic integrity based on periphyton, macroinvertebrate, and fish to southern Rocky Mountain streams. Ecological Indicators 5:117–136.

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., 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 classifications 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.

HEINO, J., T. MUOTKA, R. PAAVOLA, H. HAMALAINEN, AND E. KOSKENNIEMI. 2002. Correspondence between regional delineations and spatial patterns in macroinvertebrate assemblages of boreal headwater streams. Journal of the North American Benthological Society 21:397–413.

HERLIHY, A. T., R. M. HUGHES, AND J. C. SIFNEOS. 2006. National clusters of fish species assemblages in the conterminous United States and their relationship to existing landscape classification schemes. Pages 87–112 *in* R. M. Hughes, L. Wang, and P. W. Seelbach (editors). Influences of landscapes on stream habitats and biological assemblages. American Fisheries Society Symposium 48. American Fisheries Society, Bethesda, Maryland.

HILL, B., A. T. HERLIHY, P. R. KAUFFMANN, S. J. DECELLES, AND M. A. VANDER BORGH. 2003. Assessment of streams of the eastern United States using a periphyton index of biotic integrity. Ecological Indicators 2:325–338.

HILL, B. H., A. T. HERLIHY, P. R. KAUFMANN, R. J. STEVENSON, F. H. MCCORMICK, AND C. B. JOHNSON. 2000. Use of periphyton assemblage data as an index of biotic integrity. Journal of the North American Benthological Society 19:50–67.

ID DEQ (IDAHO DEPARTMENT OF ENVIRONMENTAL QUALITY). 2004a. Selection of reference condition for small streams in Idaho: a systematic approach. Surface Water Quality Program, Department of Environmental Quality, Boise, Idaho. (Available from: Surface Water Quality Program, Department of Environmental Quality, Boise, Idaho 83706 USA.)

ID DEQ (IDAHO DEPARTMENT OF ENVIRONMENTAL QUALITY). 2004b. Beneficial use reconnaissance program field manual for streams. Surface Water Quality Program, Department of Environmental Quality, Boise, Idaho. (Available from: Surface Water Quality Program, Department of Environmental Quality, Boise, Idaho 83706 USA.)

JESSUP, B., AND J. GERRITSEN. 2000. Development of a multimetric index for biological assessment of Idaho streams using benthic macroinvertebrates. Prepared for the Idaho Department of Environmental Quality. Tetra-Tech, Owings Mills, Maryland. (Available from: Water Quality Division, Department of Environmental Quality State Office, 1410 N. Hilton, Boise, Idaho 83706 USA.)

KARELS, T. J., A. A. BRYANT, AND D. S. HIK. 2004. Comparison of discriminant function and classification tree analyses for age classification of marmots. Oikos 105:575–587.

KARR, J. R., AND E. W. CHU. 1998. Restoring life in running waters: better biological monitoring. Island Press, Washington, DC.

KARR, J. R., K. D. FAICH, P. L. ANGERMEIER, P. R. YANT, AND I. J. SCHLOSSER. 1986. Assessing biological integrity in running waters: a method and its rationale. Publication 5. Illinois Natural History Survey, Champaign, Illinois. (Available from: Illinois Natural History Survey, 1816 South Oak Street, Champaign, Illinois 61820 USA.)

KY DOW (KENTUCKY DIVISION OF WATER). 1993. Methods for assessing biological integrity of surface waters. Kentucky Department of Environmental Protection, Frankfort, Kentucky. (Available from: Kentucky Department of Environmental Protection, 14 Reilly Road, Frankfort, Kentucky 40601 USA.)

LELAND, H. V., AND S. D. PORTER. 2000. Distribution of benthic algae in the upper Illinois River Basin in relation to geology and land use. Freshwater Biology 44:279–301.

LEWIS, P. A., D. KLEMM, AND W. THOENY. 2001. Perspectives on use of a multimetric lake bioassessment integrity index using benthic macroinvertebrates. Northeastern Naturalist 8:233–246.

MACKEY, R. L., AND D. J. CURRIE. 2001. The diversity-disturbance relationship: is it generally strong and peaked? Ecology 82:3479–3492.

MAZOR, R. D., T. B. REYNOLDSON, D. M. ROSENBERG, AND V. H. RESH. 2006. Effects of biotic assemblage, classification, and assessment method on bioassessment performance. Canadian Journal of Fisheries and Aquatic Sciences 63: 394–411.

McKENZIE, D., D. L. PETERSON, AND J. K. AGEE. 2000. Fire frequency in the interior Columbia River basin: building regional models from fire history data. Ecological Applications 10:1497–1516.

MOSS, D., M. T. FURSE, J. F. WRIGHT, AND P. D. ARMITAGE. 1987. The prediction of the macroinvertebrate fauna of unpolluted running-water sites in Great Britain using environmental data. Freshwater Biology 17:41–52.

NAYMIK, J., Y.-D. PAN, AND J. FORD. 2005. Diatom assemblages as indicators of timber harvest effects in coastal Oregon streams. Journal of the North American Benthological Society 24:569–584.

NEWALL, P., N. BATE, AND L. METZELING. 2006. A comparison of diatom and macroinvertebrate classification of sites in the Kiewa River system, Australia. Hydrobiologia 572: 131–149.

NORRIS, R. H., AND C. P. HAWKINS. 2000. Monitoring river health. Hydrobiologia 435:5–17.

O'CONNOR, R. J., T. E. WALLS, AND R. M. HUGHES. 2000. Using multiple taxonomic groups to index the ecological condition of lakes. Environmental Monitoring and Assessment 61:207–229.

OMERNIK, J. M. 1987. Ecoregions of the conterminous United States. Annals of the Association of American Geographers 77:118–125.

ORLOCI, L. 1967. An agglomerative method for classification of plant communities. Journal of Ecology 55:193–206.

OSTERMILLER, J. D., AND C. P. HAWKINS. 2004. Effects of sampling error on bioassessments of stream ecosystems: application to RIVPACS-type models. Journal of the North American Benthological Society 23:363–382.

PAN, Y., AND R. J. STEVENSON. 1996. Gradient analysis of diatom assemblages in western Kentucky wetlands. Journal of Phycology 32:222–232.

PAN, Y., R. J. STEVENSON, B. H. HILL, AND A. T. HERLIHY. 2000. Ecoregions and benthic diatom assemblages in Mid-Atlantic Highlands streams, USA. Journal of the North American Benthological Society 19:518–540.

PAN, Y., R. J. STEVENSON, B. H. HILL, A. T. HERLIHY, AND G. B. COLLINS. 1996. Using diatoms as indicators of ecological conditions in lotic systems: a regional assessment. Journal of the North American Benthological Society 15:481–495.

PAN, Y., R. J. STEVENSON, B. H. HILL, P. R. KAUFMANN, AND A. T. HERLIHY. 1999. Spatial patterns and ecological determinants of benthic algal assemblages in Mid-Atlantic streams, USA. Journal of Phycology 35:460–468.

PAUL, M. J., J. GERRITSEN, C. P. HAWKINS, AND E. LEPPO. 2005. Development of biological assessment tools for Colorado. Prepared for Colorado Department of Public Health and Environment. (Available from: Water Quality Division, Colorado Department of Environmental Health, Denver, Colorado 80246-1530 USA.)

PETERSON, D. A., AND S. D. PORTER. 2002. Biological and chemical indicators of eutrophication in the Yellowstone River and major tributaries during August 2000. Proceedings of the 2002 National Monitoring Conference. National Water Quality Monitoring Council, Madison, Wisconsin. (Available from: http://wy.water.usgs.gov/YELL/nwqmc/nwqmc.pdf)

PONADER, K. C., D. CHARLES, AND T. J. BELTON. 2007. Diatom-based TP and TN inference models and indices for monitoring nutrient enrichment of New Jersey streams. Ecological Indicators 7:79–93.

PONT, D., B. HUGUENY, U. BEIER, D. GOFFAUX, A. MELCHER, R. NOBLE, C. ROGERS, N. ROSET, AND S. SCHMUTZ. 2006. Assessing river biotic condition at a continental scale: a European approach using functional metrics and fish assemblages. Journal of Applied Ecology 43:70–80.

POTAPOVA, M., AND D. CHARLES. 2002. Benthic diatoms in USA rivers: distributions along spatial and environmental gradients. Journal of Biogeography 29:167–187.

POTAPOVA, M., AND D. CHARLES. 2007. Diatom metrics for monitoring eutrophication in rivers of the United States. Ecological Indicators 7:48–70.

RAPPORT, D. J., H. A. REGIER, AND T. C. HUTCHINSON. 1985. Ecosystem behavior under stress. American Naturalist 125:617–640.

REJWAN, C., N. COLLINS, L. L. BRUNNER, B. J. SHUTER, AND M. S. RIDGWAY. 1999. Tree regression analysis on the nesting habitat of smallmouth bass. Ecology 80:341–348.

SABLE, K. A., AND E. WOHL. 2006. The relationship of lithology and watershed characteristics to fine sediment deposition in streams of the Oregon coast range. Environmental Management 37:659–670.

SOININEN, J., AND K. KONONEN. 2004. Comparative study of monitoring South-Finnish rivers and streams using macroinvertebrate and benthic diatom community structure. Aquatic Ecology 38:63–75.

SOININEN, J., R. PAAVOLA, AND T. MUOTKA. 2004. Benthic diatom communities in boreal streams: community structure in relation to environmental and spatial gradients. Ecography 27:330–342.

SONNEMAN, J. A., C. J. WALSH, P. F. BREEN, AND A. K. SHARPE. 2001. Effects of urbanization on streams of the Melbourne region, Victoria, Australia. II. Benthic diatom communities. Freshwater Biology 46:553–565.

STEVENSON, R. J., AND Y. D. PAN. 1999. Assessing environmental conditions in rivers and streams with diatoms. Pages 11–40 in E. F. Stoermer and J. P. Smol (editors). The diatoms: applications for environmental and earth sciences. Cambridge University Press, Cambridge, UK.

STODDARD, J. L., D. P. LARSEN, C. P. HAWKINS, R. K. JOHNSON, AND R. H. NORRIS. 2006. Setting expectations for the ecological condition of running waters: the concept of reference conditions. Ecological Applications 16:1267–1276.

STRIBLING, J. B., B. K. JESSUP, AND J. GERRITSEN. 2000. Development of biological and physical habitat criteria for Wyoming streams and their use in the TMDL process. Prepared for US Environmental Protection Agency. Tetra-Tech, Owings Mills, Maryland. (Available from: US Environmental Protection Agency Region 8, 1595 Wynkoop Street, Denver, Colorado 80202-1129 USA.)

TOOTH, S., T. S. MCCARTHY, D. BRANDT, P. J. HANCOX, AND R. MORRIS. 2002. Geological controls on the formation of alluvial meanders and floodplain wetlands: the example

of the Klip River, eastern Free State, South Africa. Earth Surface Processes and Landforms 27:797–815.

TOWNSEND, C. R., M. R. SCARSBROOK, AND S. DOLEDEC. 1997. The intermediate disturbance hypothesis, refugia, and biodiversity in streams. Limnology and Oceanography 42: 938–949.

VAN SICKLE, J. 1997. Using mean similarity dendrograms to evaluate classification. Journal of Agricultural, Biological, and Environmental Statistics 2:370–388.

VAN SICKLE, J., C. P. HAWKINS, D. P. LARSEN, AND A. T. HERLIHY. 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. D. HUFF, AND C. P. HAWKINS. 2006. Selecting discriminant function models for predicting the expected richness of aquatic macroinvertebrates. Freshwater Biology 61:359–372.

VAUGHAN, I. P., AND S. J. ORMEROD. 2005. The continuing challenges of testing species distribution models. Journal of Applied Ecology 42:720–730.

WANG, Y.-K., R. J. STEVENSON, AND L. METZMEIER. 2005. Development and evaluation of a diatom-based index of biotic integrity for the Interior Plateau Ecoregion, USA. Journal of the North American Benthological Society 24:990–1008.

WRIGHT, J. F., D. W. SUTCLIFFE, AND M. T. FURSE. 2000. Assessing the biological quality of fresh waters: RIVPACS and other techniques. Freshwater Biological Association, Ambleside, Cumbria, UK.

YAMAMOTO, T., AND G. HATTA. 2004. Pulsed nutrient supply as a factor inducing phytoplankton diversity. Ecological Modelling 171:247–270.

Received: 25 October 2006
Accepted: 27 March 2007

APPENDIX. Candidate metrics, their expected responses to human disturbances, and their coefficients of variation (CV) among 69 calibration reference sites. TV = tolerance value.

Metric	Units	Definition	Expected response	CV
A/AN-I	%	*Achnanthes*/(*Achnanthes*+*Navicula*) in individuals	Down	2.15
A/AN-T	%	*Achnanthes*/(*Achnanthes*+*Navicula*) in taxa numbers	Down	1.92
ACE		ACE estimate of taxa richness	Down	0.38
ACID-I	%	Individuals of taxa preferring pH <7	Variable	2.49
ACID-T	%	Number of taxa preferring pH <7	Variable	1.19
ALK-I	%	Individuals of taxa preferring pH >7	Variable	2.64
ALK-T	%	Number of taxa preferring pH >7	Variable	0.77
CYMB-I	%	*Cymbella* individuals	Down	1.98
CYMB-T	%	Number of *Cymbella* taxa	Down	0.77
DOM	%	Individuals in the most dominant taxon	Up	0.48
HIGH-O_2-I	%	Individuals requiring 100% O_2 saturation	Down	0.65
HIGH-O_2-T	%	Number of taxa requiring 100% O_2 saturation	Down	0.33
LOW-O_2-I	%	Individuals tolerant of <30% O_2 saturation	Up	1.65
LOW-O_2-T	%	Number of taxa tolerant of <30% O_2 saturation	Up	0.68
MOB-I	%	Individuals of mobile taxa	Up	0.87
MOB-T	%	Number of mobile taxa	Up	0.38
MPSAP-I	%	Individuals of meso- and polysaprobic taxa	Up	1.14
MPSAP-T	%	Number of meso- and polysaprobic taxa	Up	0.40
NHETER-I	%	Individuals of N heterotrophic taxa	Up	1.74
NHETER-T	%	Number of N heterotrophic taxa	Up	0.55
NAVIC-I	%	Individuals of *Navicula*	Up	1.30
NAVIC-T	%	Number of *Navicula* taxa	Up	0.51
NF-I	%	Individuals of N-fixing taxa	Down	3.13
NF-T	%	Number of N-fixing taxa	Down	1.00
OSAP-I	%	Individuals of oligosaprobic taxa	Down	1.17
OSAP-T	%	Number of oligosaprobic taxa	Down	0.56
RICHNESS		Number of taxa in sample	Down	0.35
REF-I	%	Individuals of taxa indicative of reference sites	Down	0.59
REF-T	%	Number of taxa indicative of reference sites	Down	0.42
SENS-I	%	Individuals of sensitive taxa (TV = 3)	Down	0.23
SENS-T	%	Number of sensitive taxa (TV = 3)	Down	0.15
SHANNON		H′ (Shannon diversity index)	Down	0.27
SIMPSON		1/D (Simpson diversity index)	Down	0.63
TEST-I	%	Individuals of taxa indicative of test sites	Up	1.17
TEST-T	%	Number of taxa indicative of test sites	Up	0.74
TOL-I	%	Individuals of the most tolerant taxa (TV = 1)	Up	1.63
TOL-T	%	Number of most tolerant taxa (TV = 1)	Up	0.68
WA-ORG-N		Weighted average of organic N use index	Up	0.17
WA-O_2		Weighted average of O_2 TVs	Up	0.19
WA-POL		Weighted average of pollution TVs	Down	0.08
WA-SAL		Weighted average of salinity TVs	Up	0.12
WA-SAPRO		Weighted average of saprobic values	Up	0.14
WA-TROPH		Weighted average of trophic index values	Up	0.12