J. N. Am. Benthol. Soc., 2006, 25(2):513–530

Comparison of the performance of different bioassessment methods: similar evaluations of biotic integrity from separate programs and procedures

David B. Herbst[1]

Sierra Nevada Aquatic Research Laboratory, University of California, Route 1, Box 198, Mammoth Lakes, California 93546 USA

Erik L. Silldorff[2]

Princeton Hydro, 1108 Old York Road, Suite 1, P.O. Box 720, Ringoes, New Jersey 08551 USA

Abstract. Regional bioassessment programs of states, various federal agencies, and other governmental and private groups often use different methods to collect and analyze stream invertebrate samples. This lack of uniformity has created concern and confusion over the comparability of disparate sources of data, but few studies have attempted to evaluate differences in performance between methods or to reconcile the results produced from different methods. We conducted concurrent sampling at 40 sites in the eastern Sierra Nevada of California using 3 bioassessment methods to obtain directly comparable data sets. The riffle-based methods (University of California Sierra Nevada Aquatic Research Laboratory [UC-SNARL, Lahontan Water Board], California Stream Bioassessment Protocol, and US Forest Service Region 5) differed at each stage from field sample collection to laboratory processing and data analysis. We used a performance-based methods system to compare precision, uniformity, discrimination, accuracy, and correlations among multimetric Index of Biotic Integrity (IBI) scores and multivariate **R**iver **I**n**v**ertebrate **P**rediction **a**nd **C**lassification **S**ystem (RIVPACS)-type observed/expected (O/E) ratios. Reference and test sites were identified using local and upstream-watershed disturbance criteria, and invertebrate community measures and models were then developed to discriminate between reference and test sites. The more-intensive UC-SNARL method showed slightly, but consistently, greater sensitivity for discriminating impairment than the other 2 methods. The UC-SNARL method produced greater differences between reference- and test-site means relative to lower reference-site standard deviations than the other 2 methods. However, assessment scores were highly correlated among methods and distinguished reference from test sites with similar accuracy among methods despite the slight differences in performance. Our results show that differing bioassessment methods can yield very similar, effective discrimination of impaired biological condition even though they have multiple differences in field and laboratory protocols (mesh size, replication, area sampled, taxonomic resolution, total counts). Moreover, this conclusion did not depend on the approach taken to data analysis because both multimetric IBIs and multivariate RIVPACS-type O/Es were in close agreement. Methodological uniformity is important when coordinating monitoring programs, but our results suggest that data from multiple sources could potentially be used interchangeably and for cross-validation of assessments of stream biological integrity.

Key words: bioassessment, impairment detection, methods comparison, metric precision, multimetric IBI, RIVPACS, Sierra Nevada, stream macroinvertebrates.

Surveys of the different stream bioassessment protocols used among federal, state and local programs show considerable variation in the procedures and tools used to collect and process samples (Gurtz and Muir 1994, Carter and Resh 2001). Comparisons of the data derived from collections taken with various types of sampling equipment, subsampling counts, and levels of taxonomic resolution have provided a basis for evaluating some of the field and laboratory methods in use (Resh and McElravy 1993, Resh and Jackson 1993, Barbour and Gerritsen 1996, Courtemanch 1996, Vinson and Hawkins 1996, Lenat and Resh 2001). The techniques used to analyze bioassess-

[1] E-mail addresses: herbst@lifesci.ucsb.edu
[2] esilldorff@princetonhydro.com

ment data also have been compared using the same sets of biological data (e.g., Fore and Karr 1996, Reynoldson et al. 1997). What has not been done for more than a few data sets (e.g., Houston et al. 2002, Cao et al. 2005) is a comparison of bioassessment results from concurrent or side-by-side sampling using methods that differ at several stages from field collections through laboratory processing and identification to the data analyses used to assess biological impairment. Such a comparison provides the most realistic context for evaluating the results produced from different monitoring programs. It also provides the information needed for calibration of methods to enable interagency cooperation and data sharing when developing biological criteria for water quality.

Organized bioassessment programs for monitoring water quality have been in operation in California since ~1993. Extensive data sets have been collected by several large agencies including the Aquatic Bioassessment Laboratory of the California Department of Fish and Game, the US Forest Service on National Forest lands, and the Lahontan Regional Water Quality Control Board in watersheds on the east slopes of the Sierra Nevada. These programs have used different field and laboratory protocols for sampling, processing, identifying, and analyzing data. Other programs with other methods also exist in California, but our study contrasts the 3 large programs listed above. These programs also were emphasized in a report reviewing the status of bioassessment in California (Barbour and Hill 2003).

Use of a performance-based method system (PBMS) has been suggested when evaluating the comparability of bioassessment methods (Diamond et al. 1996, Barbour et al. 1999). PBMS compares bioassessment results to a performance standard. If performance measures meet or exceed the standard, the method is considered acceptable for use in monitoring. Performance standards may be defined based on required data-quality objectives (DQOs) of a program or relative to a reference, or accepted, method. Methods are compared on the basis of performance characteristics that include precision, bias, discrimination power (ability to distinguish test from reference sites), and accuracy, particularly with respect to minimizing Type II error rate (i.e., misclassification of an impaired site as unimpaired). PBMS can identify differences between bioassessment methods and can inform decisions regarding the most appropriate method(s) for meeting defined DQOs.

The objectives of our study were to use different methods in the same set of sampling sites to: 1) evaluate differences in the ability of 3 common bioassessment methods used in California to meet PBMS criteria, 2) evaluate whether combined differences in field collection, laboratory processing, and data analysis affect the outcome of assessment of biological impairment, 3) provide explicit descriptions of the steps involved in multivariate **R**iver **I**nvertebrate **P**rediction **a**nd **C**lassification **S**ystem (RIVPACS)-type model and multimetric-model development, and 4) compare the costs and benefits of the 3 methods relative to their abilities to discriminate impairment.

Methods

Forty streams of various sizes (order, mean width, watershed area) were selected to represent least-impaired reference sites and a variety of impaired sites in a geographic region restricted to the eastern slopes of the Sierra Nevada (Great Basin watersheds between lat 37–40°N and long 118–120°W). The streams were sampled at the same sites and on the same dates using each of 3 methods: 1) University of California Sierra Nevada Aquatic Research Laboratory (UC-SNARL) Protocol for the Lahontan Water Quality Board, 2) California Stream Bioassessment Protocol (CSBP) for the California Department of Fish and Game, and 3) Utah State University Protocol for US Forest Service Region 5 (USFS.R5). Impaired (test) sites were selected from disturbed landscapes over a gradient of physical habitat degradation related mainly to livestock grazing and altered channel geomorphology (erosion and sediment pollution). Reference sites were selected based on initial screening for low upstream density of road crossings (a measure of watershed development), low local bank erosion, and minimal exposure to local and upstream pollution or landscape disturbance (Table 1). Sites were grouped into 24 reference and 16 test sites based on the criteria above for development of multimetric Indices of Biological Integrity (IBIs) and RIVPACS-type observed/expected (O/E) ratios (see below). Most reference sites (14 of 24) had low upstream density of road crossings (<0.2/km), low local bank erosion (<25% bank erosion), and no known pollution sources, but others (10 of 24) met only one criterion, with either density of upstream road crossing >0.2/km and local bank erosion <25% or density of upstream road crossing <0.2/km and local bank erosion >25% and no known pollution sources.

Sampling protocols

A 150-m-long study reach, located by GPS-UTM coordinates and elevation (near the lower end of each site), was identified in each stream (site), and all samples, regardless of method, were collected within these study reaches.

TABLE 1. Stream identification, size, and reference–test site classification. Sites are sorted by stream size and the density of upstream road-crossings (primary reference-site selection criterion). Large streams were >400 cm wide or had upstream length >5 km (16 reference streams). Small streams were <400 cm wide or had upstream channel length <5 km (8 reference streams). X = known local or upstream source of point- or nonpoint-source pollution present (usually grazing or altered channel structure), R = reference site (<0.2 road crossings/km or reach-scale bank erosion <25% with no pollution source), T = test site.

Stream names (codes)	Width (cm)	Upstream length of channel (km)	Upstream road crossings (/km)	% bank erosion	Pollution source present	Reference–test classification
Large streams						
Truck.forest (TF)	737	11.3	0.000	0.0		R
ECarson (EC)	1484	37.3	0.000	3.3		R
Silver (SV)	711	22.2	0.000	10.0		R
WWalker.Leavitt (W)	1253	24.6	0.000	40.0		R
Convict (CN)	415	16.6	0.043	0.0		R
Wolf (WO)	636	12.8	0.076	20.0		R
WWalker.Pickel (WP)	1464	27.8	0.102	33.3	X	T
Robinson.honey (RH)	817	23.1	0.112	26.7		R
Buckeye (B)	422	30.3	0.122	76.7	X	T
Sagehen (S)	382	6.0	0.123	3.3		R
Robinson.below (RB)	672	34.8	0.134	63.3	X	T
Lee (L)	951	12.6	0.145	10.0		R
Rush (R)	963	30.3	0.170	26.7	X	T
Deadman (D)	489	17.3	0.174	13.3		R
Owens.belowtun (OT)	1008	23.9	0.188	0.0	X	T
Owens.abovetun (OA)	644	23.3	0.189	0.0	X	T
Owens.spring (OS)	753	19.2	0.191	0.0		R
EWalker (EW)	919	24.6	0.221	90.0	X	T
Owens.417 (O4)	964	27.8	0.225	26.7	X	T
Owens.power (OP)	994	32.4	0.235	3.3	X	T
Truck.Celio (TC)	736	12.8	0.280	6.7		R
WCarson.blm (WC)	1255	33.4	0.312	6.7		R
Truck.park (TP)	921	13.7	0.315	7.0		R
Truck.Bart (TB)	885	21.8	0.327	33.0	X	T
Owens.bridge (OX)	1556	42.2	0.389	16.7	X	T
Owens.Benton (OB)	1132	44.2	0.395	33.3	X	T
Mammoth (M)	660	17.0	0.560	10.0		R
Cold (C)	523	6.9	0.565	20.0		R
Small streams						
Trib.Silver (T)	75	2.0	0.000	0.0		R
Forestdale (F)	318	2.0	0.000	3.3		R
Willow (WW)	307	10.4	0.000	3.3		R
Spratt (SP)	174	7.2	0.132	10.0		R
WCarson.faith (WF)	479	4.3	0.195	3.3		R
Kirman (K)	96	2.8	0.232	10.0	X	T
Cottonwood (CT)	153	8.0	0.269	0.0		R
Cowcamp (CW)	114	3.4	0.286	10.0		R
Slinkard (SL)	66	8.0	0.365	0.0		R
Bagley.meadow (BM)	133	2.0	0.629	10.0	X	T
Bagley.control (BC)	136	2.7	0.862	10.0	X	T
Poore (P)	207	4.4	0.890	3.3	X	T

Physicochemical variables.—Riffle and pool habitats were delineated (longitudinal distribution and length) and flagged for transect locations. The slope of the reach was measured with an autolevel and stadia rod, and sinuosity was estimated as the ratio of 150 m of reach length to the linear distance between the upper and lower ends of the reach. Bank and channel habitat were measured along 15 transect cross-sections spaced at 10-m intervals over the length of the reach. Water depth, substrate type, and current velocity were measured at 5 equidistant points along each transect. Stream width, bank structure (cover/substrate type and stability rating), riparian canopy cover, and bank angle were measured at each transect location. Bank

structure was rated as open, vegetated, or armored (rock or log), and as stable or eroded (evidence of bank erosion, collapse, or scour scars) between water level and bankfull channel level. Bank angles were scored as shallow, moderate, or undercut (<30°, 30–90°, and >90°, respectively), and riparian cover was measured from vegetation reflected on a grid in a concave mirror densiometer as the sum of grid points for measurements taken at each stream edge and at midstream facing up- and downstream. The type and amount of riparian vegetation along the reach was estimated by qualitative visual evaluation. Embeddedness of cobble-size substrate was estimated for 25 cobbles (encountered during transect surveys or supplemented with randomly selected cobbles) as the volume of the rock buried by silt or fine sand. Discharge was calculated from cross-sectional area and current velocity. A suite of basic water-chemistry and related variables including dissolved O_2, conductivity, pH, temperature, and turbidity were measured at each site.

UC-SNARL.—Five replicate samples of benthic macroinvertebrates were taken in riffles using a 30-cm-wide D-frame kick net with a 50-cm-long bag with 250-μm mesh. Each replicate was a composite from three 30.5 × 30.5-cm sample areas (0.093 m^2 each, 0.279 m^2 total) taken across the riffle transect (or in upstream series for small streams) over zones of varied depth, substrate, and current. Sample transects were selected using a random number table for locations corresponding to a delineated riffle segment. Each kick sample was taken using a mixture of feet and hands to dislodge and rub substrates for 30 s to 1 min so that both mobile and attached invertebrates were washed into the downstream net that was held against the bottom. These composited replicates were intended to represent varied microhabitat conditions and reduce variability among sample replicates. Samples were processed in the field by washing in buckets and removing large organic and rock debris followed by repeated elutriation of the sample to remove invertebrates from remnant sand and gravel debris. The remaining rock and gravel debris was inspected in a shallow white pan to remove any remaining organisms, including caddisflies with stone cases and snails or other mollusks with shells. Elutriated and inspected sample fractions were preserved in ethanol, and a small volume of rose Bengal stain was added to aid in laboratory processing. Invertebrate field samples were subsampled in the laboratory using a rotating drum splitter. Invertebrates were sorted under a stereomicroscope at 10× magnification, and minimum count of 250 organisms was removed from each replicate for identification (in practice ranging mostly from 400–500 individuals). Individuals (including midges and mites) were identified to the lowest practical taxonomic level (usually genus, species, or species group) depending on the availability of taxonomic keys. Oligochaetes and ostracods were not further identified. All sample sorting was done to achieve <5% error in removal, and quality-control verifications of every taxon identified in every sample were done by DBH. Unprocessed sample remnants also were searched (using a 3× magnification visor) for rare and large taxa not encountered in the processed sample, and single counts of those individuals were added to the total.

CSBP.—Samples were taken within the same study reaches at locations adjacent to the locations of the 1st, 3rd, and 5th UC-SNARL replicates. Three replicate CSBP samples were taken using a 30-cm-wide D-frame kick net fitted with a 50-cm-long 500-μm-mesh net. Each replicate was a composite from three 30.5 × 61-cm (width × length) sample areas (0.186 m^2 each, 0.558 m^2 total). Samples were processed in the field, preserved, and stained as described above for the UC-SNARL method. Laboratory subsampling was done by spreading the field sample over a large shallow white pan with a grid drawn on the bottom. All organisms were removed from grid sectors selected with a random number generator until a fixed count of 300 ind./sample was reached. Invertebrates were identified at the same level of taxonomic resolution as the UC-SNARL method except that midges were identified only to subfamily and all mites were left at Hydracarina. Quality-control checks of laboratory processing and identifications were done as for the UC-SNARL samples. A rare-and-large-taxa search was done as above.

USFS.R5.—Single composite samples were taken at eight 30.5 × 30.5-cm sample areas (0.093 m^2 each, 0.74 m^2 total) in the 4 longest riffles in the study reach (2 samples in each riffle, selected at random from a 9-point grid). When <4 riffles were available, sample locations were assigned in proportion to the length of each riffle. Samples were taken using a 30-cm-wide D-frame kick net fitted with a 50-cm-long 500-μm-mesh net. Samples were processed in the field, preserved, and stained as described above for the UC-SNARL method. Subsampling was done as described above for the CSBP method but to a fixed count of 500 organisms. Specimens were identified to the same level of taxonomic resolution as in the UC-SNARL method, including identification of midges and mites to genus and some species groups. Quality-control checks of laboratory processing and identifications were done as for the UC-SNARL samples, as were checks for rare and large taxa. The basic differences among methods are summarized in Table 2.

TABLE 2. Summary of differences in field and laboratory protocols between bioassessment methods. All methods were based on riffle-stratified habitat sampling for macroinvertebrates. UC-SNARL = University of California Sierra Nevada Aquatic Research Laboratory Protocol (Lahontan Water Quality Board), CSBP = California Stream Bioassessment Protocol (California Department of Fish and Game), USFS.R5 = Utah State University Protocol (US Forest Service Region 5).

	UC-SNARL	CSBP	USFS.R5
Net type, mesh	D-frame, 250 μm	D-frame, 500 μm	D-frame, 500 μm
Replication	5 composites of 3	3 composites of 3	1 composite of 8
Area sampled	1.39 m^2	1.67 m^2	0.74 m^2
Subsampling	Drum splitter	Grid tray	Grid tray
Enumeration	250–500 count	300 fixed count	500 fixed count
Taxonomic resolution	Genus/species for all taxa	Genus/species for all taxa except midges and mites to subfamily/family	Genus/species for all taxa

Analytical Methods

Data collected with the UC-SNARL and CSBP methods typically are analyzed using the multimetric calculations recommended by the USEPA (*multimetric modeling*, Barbour et al. 1999), whereas data collected with the USFS.R5 method usually are analyzed using a series of multivariate statistical methods first developed in Great Britain and referred to as RIVPACS-type models or multivariate predictive models (Moss et al. 1987). In our study, data sets from all 3 methods were analyzed using both the multimetric modeling and the multivariate RIVPACS-type modeling approaches so that field, laboratory, and analytical methods could be compared systematically.

Multimetric IBI model

Our calculation of a multimetic IBI model closely followed the recommendations and procedures outlined in the USEPA Rapid Bioassessment Protocol document (Barbour et al. 1999). Multimetric IBI models have not been developed and implemented for the eastern slopes of the Sierra Nevada, California, for any of the 3 methods we evaluated, so new multimetric IBI models were constructed during our study. Sixty-nine metrics were calculated for each sample across the 3 methods. The 69 metrics were created from 28 basic metrics by varying the calculation of a metric slightly. For example, taxa richness was standardized to different sampling levels using a rarefaction procedure; and dominance was calculated either as the most common taxon, the 3 most abundant taxa, or the number of taxa required to attain 50% of the total count.

Three criteria (power, consistency, and uniqueness) were used to identify a set of core metrics that could be more thoroughly evaluated for inclusion in a multimetric IBI. For the first 2 criteria (power and consistency), our evaluation was based on the overlap between the test and reference scores for each metric as a means of assessing the strength of the impairment signal relative to the background variability in that metric's scores.

Power.—Power was the most important consideration for including or excluding metrics for further consideration. Power was measured empirically as the overlap between test and reference scores, with overlap measured as the proportion of test (i.e., impaired) sites that exceeded various percentiles of the reference-site distribution of values for that metric. Overlap based on percentiles essentially evaluates the signal-to-noise ratio by considering the separation between the centers of the test- and reference-site distributions simultaneously with the spread of values around these centers. The sample size used for our study (24 reference streams) sometimes created discrete jumps between the values for adjacent percentiles. Therefore, the overlap between test- and reference-site distributions was evaluated broadly by considering multiple percentiles (range: successive elimination of the lowest 6 of the 24 reference streams in turn, or $\sim 4^{th}$–25^{th} percentiles) for each metric rather than choosing a single percentile for all comparisons. Metrics for which $<40\%$ of test-sites scored above the reference-site threshold (or below an upper threshold for reverse-scale metrics) were identified as potential candidates. Additional weight was given to metrics with least overlap and, thus, high power to discriminate between the reference and test distributions.

Consistency.—Consistency was defined as a systematic decrease in the proportional overlap between test- and reference-site distributions for increasing percentile thresholds of the reference-site distribution. Consistency primarily reflected the shapes of the test- and reference-site distributions and the behavior of the tails of these distributions. Therefore, this measure was used primarily to flag metrics with marked inconsistencies, particularly in the reference-site class. Rank-

ordered plots of metric scores also were used to evaluate the shapes of these distributions and, thus, the consistency of reference and test scores.

Uniqueness.—The uniqueness of a metric relative to other metrics was evaluated quantitatively with Pearson's correlation coefficients and conceptually by examining the possible dependencies among metrics. Like consistency, uniqueness was used to highlight metrics with numerous strong correlations with other metrics that had suitable power and consistency. Metrics with strong correlations (typically, $r > |0.8–0.9|$) and a conceptual relationship to other metrics were excluded from further consideration.

Screening of metrics using these 3 criteria yielded 22 candidate metrics that were considered more completely for inclusion in a final multimetric IBI for ≥ 1 of the 3 methods. Building specific multimetric IBIs for each method relied on 3 quantitative and qualitative measures of the individual metrics and the complement of metrics under consideration. These measures were power, uniqueness, and representation among different metric categories (Barbour et al. 1999). Power and uniqueness were measured as before, but at this stage of the selection process, both measures were given similar weights. Thus, uniqueness played a more important role at this stage in IBI creation than in the first stage, and the final set of metrics was selected to minimize or eliminate pairwise correlations where $r > |0.8|$. The 22 candidate metrics were assigned to 1 or 2 of 4 broad metric categories (richness measures, composition measures, tolerance measures, and functional/habit measures). Metrics for each candidate IBI index were selected to yield equal or nearly equal representation among the 4 categories of metrics. More richness measures met the selection criteria than metrics in other categories for each of the 3 methods considered in our study. Moreover, the richness measures often had the strongest discriminatory power. Thus, slightly more richness measures than metrics from the other 3 categories were included in our candidate multimetric IBIs.

Before constructing the candidate IBIs for each method, individual scores for the different metrics were converted to standardized scores on a continuous 0 to 10 scale so that metrics could be aggregated into a multimetric IBI. For each metric, any value greater than or equal to the median value of the reference-site distribution was scored as 10. The minimum value of the test-site distribution was scored as 0 because this value represented the worst empirical value attained in our study. Any metric score between the reference-site median and the test-site minimum values was scored by interpolating between these 2 numbers.

The candidate IBI multimetric score was calculated by summing the scaled metric scores and multiplying this sum by the quotient (10/[no. of metrics]) so that the final scores for all IBIs theoretically ranged from 0 to 100, with equal weight given to each metric in the calculation. Four performance characteristics were then quantified and examined: 1) power based on different percentiles of the reference-site IBI scores (as above), 2) the coefficient of variation (CV) for the reference-site scores as a measure of variability or noise, 3) the ratio of the reference-site mean to the test-site mean score as a measure of the impairment signal, and 4) the standardized difference between the reference-site and test-site means $((\bar{X}_{ref} - \bar{X}_{test})/\hat{\sigma}_{ref})$. In addition, the number of metrics falling within each of the 4 metric categories, the maximum r value among metrics within the IBI, and the number of correlations among metrics with $r > |0.707|$ ($R^2 > 0.50$) were determined. These 7 criteria were used to select a final optimal IBI with 6 to 8 metrics to use with data obtained by each of the 3 methods (Table 3). In addition, the CSBP and USFS.R5 data were analyzed using the 6-metric optimal IBI developed for the UC-SNARL method (standardized IBI) to standardize the analytical step and to focus on the effect of differences in field and laboratory techniques among methods. The UC-SNARL 6-metric IBI was used for this comparison because it was the only final IBI in which each of the component metrics performed sufficiently well for all 3 methods. Alternative candidate multimetric IBIs based on 5 to 15 metrics also were evaluated, and the results were comparable to the results we present, with no qualitative changes to our conclusions based on differences among IBIs with strong performance.

Multivariate RIVPACS-type model

The original RIVPACS models were developed by a team of researchers in the UK and have been used extensively, in different forms, in Australia, Canada, and the US (e.g., Moss et al. 1987, Reynoldson et al. 1995, Marchant et al. 1997, Hawkins et al. 2000, CEH 2003). Detailed steps for building these multivariate models have been outlined elsewhere (Moss et al. 1987, Moss 2000). Therefore, only our decisions on important details are presented. Building RIVPACS-type models can be described conceptually as a 5-step process, although this process has been defined with varying numbers of steps in the literature (e.g., Moss et al. 1987, Marchant et al. 1997, Ostermiller and Hawkins 2004). These 5 conceptual steps are:

1. Identify relatively homogeneous groups of reference sites based primarily or exclusively on the biological

TABLE 3. Metrics used for development of multimetric Indices of Biological Integrity (IBIs) for each method. Method abbreviations as in Table 2. EPT = Ephemeroptera, Trichoptera, and Plecoptera taxa.

Metrics selected for IBI development	Abbreviation	UC-SNARL	CSBP	USFS.R5
Richness (number of taxa/sample)	rich	X	X	X
% EPT of total abundance	perc.ept.abund		X	
Ephemeroptera richness	e.rich			X
Plecoptera richness	p.rich		X	X
Trichoptera richness	t.rich	X	X	
% EPT richness of total richness	perc.ept.rich	X		
Diptera richness	dip.rich		X	
% chironomid richness of total richness	perc.chiro.rich			X
Biotic index (modified Hilsenhoff)	bi	X	X	X
% of taxa intolerant of pollution (tolerance values of 0, 1, or 2)	intol.numb.perc		X	X
% of taxa tolerant of pollution (tolerance values of 7, 8, 9, or 10)	tol.numb.perc	X		
% shredder feeding guild	shredder	X	X	X

communities sampled at different sites (most frequently done using cluster analysis).
2. Develop decision rules for classifying sites into the groups identified in Step 1 based only on the physicochemical setting of the stream and its watershed (typically accomplished with discriminant analysis).
3. Use the decision rules established in Step 2 to assign the probability of sites belonging to each of the groups identified in Step 1 (typically obtained through the discriminant analysis routine or software).
4. Calculate the probability that each taxon in the data set will be collected at each site based on the physicochemical setting of a site, the reference-site biological data, and the models used in Steps 2 and 3 (this is the most novel step and involves a number of specific calculations; some details are provided below).
5. Calculate the Expected taxa richness (E) as the sum of the probabilities from Step 4 and the Observed taxa richness (O) from field sampling of a site, and use the ratio of these values (O/E) as the index or test statistic for each site (the taxa richness calculations are most often done for just the most common taxa; see below for how this threshold of *common* is defined).

O/E ratios usually center on 1.0 for reference sites but are <1.0 for sites that have been altered by anthropogenic stresses. This reduction in the ratio presumably occurs because taxa that would be expected at a site have been lost as a result of the anthropogenic impacts to that site, thus, reducing the numerator in the O/E ratio.

A number of analytical steps and decisions about specifically how to build the RIVPACS-type model underlie these 5 conceptual steps. The sensitivity of the final model output to these choices has been evaluated to a limited extent, but no consensus exists for the specific decisions that should be made at each step in the model construction (Moss et al. 1987, 1999, Ostermiller and Hawkins 2004). Thus, building a RIVPACS-type model is a somewhat subjective process, and researchers should document the decisions made during the model-building process.

Step 1.—A suite of cluster-analysis methods were used to identify the most consistent grouping structure of the reference sites during this initial step. The clustering methods used were: 1) Ward's clustering, 2) flexible-β Weighted Pair-Group Means with Arithmetic averaging (flexible-β WPGMA; flexible-β Unweighted Pair-Group Means with Arithmetic averaging [UPGMA] was unavailable), and 3) UPGMA using average linkage. Analyses were done using a specialized S-Plus clustering procedure written by D. L. Lorenz (Mounds View, Minnesota) and verified for selected clustering outputs using established analytical procedures in S-Plus (version 6.0, Insightful, Inc., Seattle, Washington) and SAS (release 6.12, SAS Institute, Cary, North Carolina). For all final analyses, Sorensen's similarity measure with presence/absence data was used to measure the similarity among samples. For data collected by each method, 3 or 4 clusters of reference sites were identified, with the total number of sites in each group ranging from 4 to 15.

Step 2.—A separate and distinct discriminant analysis model was constructed for each of the 3 methods. The groups of sites identified in the cluster analysis were differentiated using a subset of physical habitat variables at each site. Only abiotic variables that were unlikely to be affected by human disturbance were included as candidate variables for the discriminant analysis model. The candidate variables that met this criterion were: elevation, latitude, longitude, sampling

date, azimuth, distance to headwaters, watershed area, slope, depth, width, % of boulder outcrops, and 2 climatic statistics (annual precipitation, number of days with precipitation) obtained through Climate Source, Inc. (http://www.climatesource.com). The final discriminant model was selected through a series of manual variable-selection steps in which candidate models were run and evaluated based on both apparent and cross-validation error rates. For each of the 3 methods, the final model selected was a subjective choice among models with both low error rates for each group and a complement of predictor variables that were conceptually distinct but plausible drivers for differences among the invertebrate communities in these reference sites. The final models selected for the 3 methods each contained 3 environmental predictor variables: UC-SNARL – depth, sampling date, latitude; CSBP – width, sampling date, elevation; and USFS.R5 – depth, sampling date, azimuth.

Step 3.—A proportional prior was used in the above discriminant analysis models for the prediction of each stream's group membership. Thus, any new site had a larger probability of belonging to the group of sites with the largest number of members than the group of sites with the smallest number of members (consistent with the original British formulation; Moss et al. 1987).

Step 4.—The probability that a taxon would be present at a site was calculated as a weighted mean value. The observed proportion of sites in each group of reference streams in which that taxon was found (F_t of Ostermiller and Hawkins 2004) was multiplied by the probability that the site belonged to each stream group obtained from the discriminant analysis in Step 3 (P_g of Ostermiller and Hawkins 2004). The following pair of examples will clarify these calculations, which are at the core of RIVPACS-type modeling: 1) Suppose *Baetis* was found at 8 of 16 streams in reference group A, 5 of 5 streams in reference group B, and 4 of 4 streams in reference group C. For *Baetis* at stream X, which has probabilities of membership (based on environmental conditions; Step 3) in groups A, B, and C of 0.75, 0.15, and 0.10, respectively, the final probability that *Baetis* will be present at stream X is 0.625, i.e.:

$$\Pr(\textit{Baetis} \text{ at stream X}) = \frac{8}{16}(0.75) + \frac{5}{5}(0.15) + \frac{4}{4}(0.10) = 0.625$$

2) For *Baetis* at stream Z, which has probabilities of membership in groups A, B, and C of 0.05, 0.50, and 0.45, respectively, the final probability that *Baetis* will be present at stream Z is 0.975, i.e.:

$$\Pr(\textit{Baetis} \text{ at stream Z}) = \frac{8}{16}(0.05) + \frac{5}{5}(0.50) + \frac{4}{4}(0.45) = 0.975$$

Thus, for stream X, the low probabilities of being in groups B and C translate into a low probability of *Baetis* being present at the site. For stream Z, the high probabilities of being in groups B and C translates into a high probability of *Baetis* being present at the site.

Step 5.—The final calculation of O and E taxa richness used a probability threshold for including taxa in the calculations for each site (a P_t cutoff as described by Ostermiller and Hawkins 2004). Use of such thresholds frequently has led to improved model performance (Moss et al. 1987, Marchant et al. 1997, Ostermiller and Hawkins 2004). All invertebrate taxa with a probability of being present at a site <0.50 (i.e., $<50\%$ predicted probability) were removed from both the E and O richness calculations. These calculations are based on different subsets of taxa for each site because the probabilities of group membership and, thus, the probability of a taxon being expected at a site, are calculated separately for each site. Thus, our O and E taxa richness values are only for those common taxa that were collected at $>50\%$ of sites in one or more of the groups showing the greatest affinity to any given site. The O/E ratio was calculated as the biological condition index or test statistic for each site.

PBMS

A wide variety of metrics was screened for inclusion in IBI development depending on their abilities to separate test and reference sites and minimize background variability. Screening resulted in selection of 12 metrics that were used as a standard system for comparison based on the same set of indicators across all methods (Table 3). Four PBMS criteria (precision, consistency, discriminatory power, and accuracy), described in technical guidance documents (Diamond et al. 1996, Barbour et al. 1999, Barbour and Hill 2003), were used to evaluate and compare methods.

Precision and consistency.—The CVs for metrics at reference sites were used as a standardized measure of precision. The number of metrics that met predetermined DQOs (CV = 10–15%, 15–20%, or 20–25%) was determined for the 12 metrics used to develop the IBIs, the aggregate multimetric IBI score, and the O/E ratio of the RIVPACS-type model. The ratios of CVs of metrics at reference sites in different stream-size classes (small vs large streams) were used as a measure of consistency (equivalence in metric precision for different stream types or ecoregions). A ratio ~1.0 indicates high consistency.

Discriminatory power and sensitivity.—Discriminatory

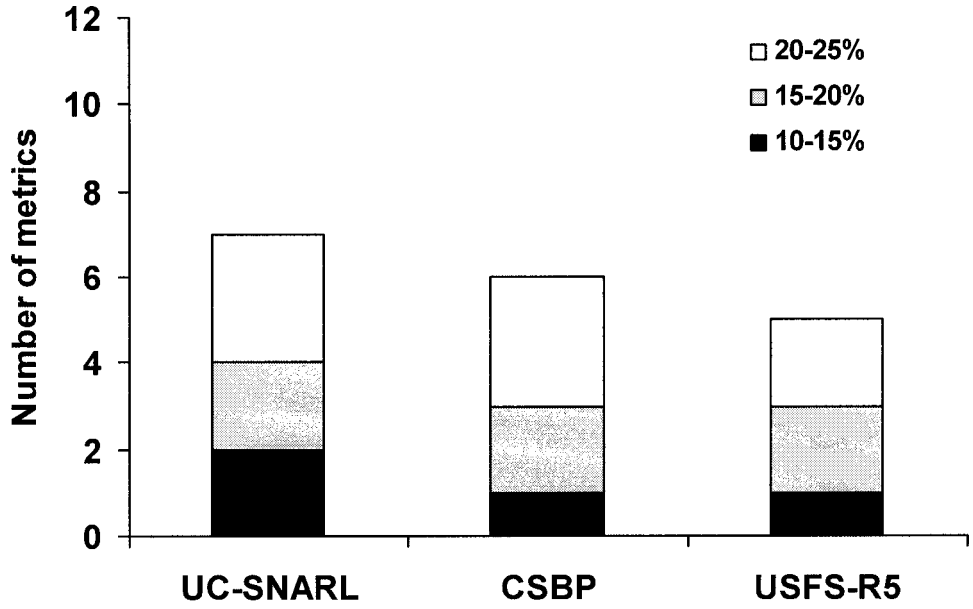

FIG. 1. Number (of 12) metrics that satisfied data-quality objectives (DQOs) at each level of variability (coefficient of variation [CV] values) when preparing the multimetric model for reference sites. Method abbreviations as in Table 2.

power and method sensitivity were estimated with 2 statistics: the ratio of the reference-site mean to the test-site mean and the difference between reference-site and test-site means standardized by the reference-site standard deviation ($(\bar{X}_{ref} - \bar{X}_{test})/\hat{\sigma}_{ref}$). The ratio of means identifies the *signal* of the average reference site relative to the average test site without taking into account the variability in site assessments (higher ratios indicate greater power). The standardized difference in means also puts the difference in means onto a standard scale but instead uses the reference-site variability to scale the difference in means (large values indicate high sensitivity). Thus, the 2 statistics give *signal* and *signal:noise ratio* estimates for the different assessment methods.

Accuracy.—We did not know a priori that test sites were impaired, but test sites were exposed to stress or disturbance and formed a class distinctly different from the undisturbed or least-exposed reference sites (Table 1). Thus, test sites could be used to compare the assessment methods presuming some level of impairment. The actual discriminatory power and method sensitivity for the set of 24 reference and 16 test sites was defined as the number of test sites that would be classified as unimpaired (misclassification of these test sites, or false positives, a Type II statistical error) based on different empirical impairment thresholds (misclassification of these reference sites, or false negatives, a Type I statistical error rate). The empirical thresholds used were the lower observed IBI scores or O/E ratios for reference sites (i.e., the lower percentiles of the reference distribution). Small Type II error rates were used as an indicator of accuracy.

Comparison of methods

Assessments were compared among methods using Lin's concordance correlations (Zar 1999). This statistic was used because it is designed to test whether the results of one method are reproducible by another, given paired observations with similar ranges, and is considered superior to other correlation measures for this purpose (Lin 1989). Pairwise correlations were calculated between the optimum IBI scores for each method, between standardized IBI scores for each method, and between O/E ratios for each method. To help visualize the correspondence among methods, optimal and standardized IBI scores for each method were plotted relative to the ranking of sites based on their UC-SNARL IBI scores, and O/E ratios for each method were plotted relative to the ranking of sites based on their UC-SNARL O/E ratios.

Discrimination of transitions in assessment scores that indicate loss of biological integrity is an important way to define the environmental thresholds at which impairment of structure and function occurs. Distinguishing gradations in biological structure and function is a key underpinning of the regulatory process of assigning streams to different categories of aquatic life use attainment (Jackson 2004). The clarity with which different methods permit identification of thresholds and intermediate subdivisions of impairment is

another feature then that should be considered when comparing model performance. Plots of ranked IBI scores and O/E ratios for each method were inspected visually for transitions in assessment scores/ratios and for intergradation of scores/ratios from reference and test sites.

Cost/benefit analysis

Evaluation of alternative assessment approaches requires that the performance characteristics of the methods be compared and that the cost:benefit ratios of the methods be considered. A balance must be achieved between the accuracy and utility of the assessment results and the expense in time and cumulative effort if monitoring efforts are to be sustained. An estimate of the relative cost of each method was obtained from field and laboratory observations of person-hours required to complete tasks of habitat surveys, sample collection, processing, sorting, identification, and counting. The data-analysis phase was accounted in this cost estimation qualitatively in terms of the level of expertise and number of steps required to obtain complete results.

Results

PBMS

Precision and consistency.—More of the 12 metrics used for IBI development had reference-site CVs below DQOs when calculated from the UC-SNARL reference-site data set than from the CSBP or USFS.R5 reference-site data sets (Fig. 1). Reference-site CVs for IBIs and O/E ratios were all below a DQO of 15% (Table 4). IBI scores and O/E ratios based on data obtained with the UC-SNARL method were ~⅓ less variable than IBIs and O/E ratios based on data obtained with the other 2 methods. CVs of metrics at reference sites usually differed between stream-size classes. However, this disparity in the precision of measurements of community attributes between different habitat types (called "bias" by Diamond et al. 1996) depended more on the metric being evaluated than on the methods being compared (Fig. 2). For example, richness tended to be more variable in small streams than in large streams.

Discriminatory power and sensitivity.—Relatively high test-site means for IBI and O/E values for the UC-SNARL method resulted in a reduced impairment signal (ratio of reference to test means) compared to the other methods (Table 4). Thus, the apparent discriminatory power was slightly lower on average for UC-SNARL method reference sites relative to impaired sites. However, the lower standard deviation for UC-SNARL (most intensive sampling methodology) led to higher standardized differences between reference- and test-site means. Thus, the UC-SNARL method had greater overall sensitivity than the other 2 methods when both signal and noise components were considered. This result suggests that the UC-SNARL method, with its reduced variance, might provide better ability to distinguish impaired sites from the reference condition than the other 2 methods.

Accuracy.—Overlap between reference- and test-site distributions of metrics was minimal with all methods using both multimetric and multivariate models (Table 4). For all but the minimum empirical threshold, 0 to 3 test sites would be misclassified as unimpaired across all methods (Table 5). Minor differences, which probably represent random variability, existed among the methods, but the CSBP method had a slightly stronger tendency to misclassify a larger number of test sites than the other 2 methods. In addition, the

TABLE 4. Precision, discriminatory power, and sensitivity for all methods based on optimized Index of Biological Integrity (IBI) scores from multimetric models and observed/expected (O/E) ratios from **R**iver **I**n**v**ertebrate **P**rediction **a**nd **C**lassification **S**ystem (RIVPACS)-type models. Method abbreviations as in Table 2. Numbers after method names indicate the number of metrics in the optimized metric set IBI for that method. Standardized difference calculated as $(\bar{X}_{ref} - \bar{X}_{test})/\hat{\sigma}_{ref}$.

	Optimized metric set IBI			RIVPACS model O/E		
	UC-SNARL-6	CSBP-8	USFS.R5–7	UC-SNARL	CSBP	USFS.R5
Reference sites						
Mean	89.35	86.36	85.74	0.999	1.018	1.032
Standard deviation	8.84	12.25	11.24	0.101	0.143	0.142
Coefficient of variation	0.099	0.142	0.131	0.101	0.140	0.138
Test sites						
Mean	49.98	42.66	45.06	0.606	0.557	0.541
Standard deviation	17.05	17.46	15.52	0.130	0.167	0.158
Reference mean/test mean	1.79	2.02	1.90	1.65	1.83	1.91
Standardized difference	4.46	3.57	3.62	3.89	3.23	3.45

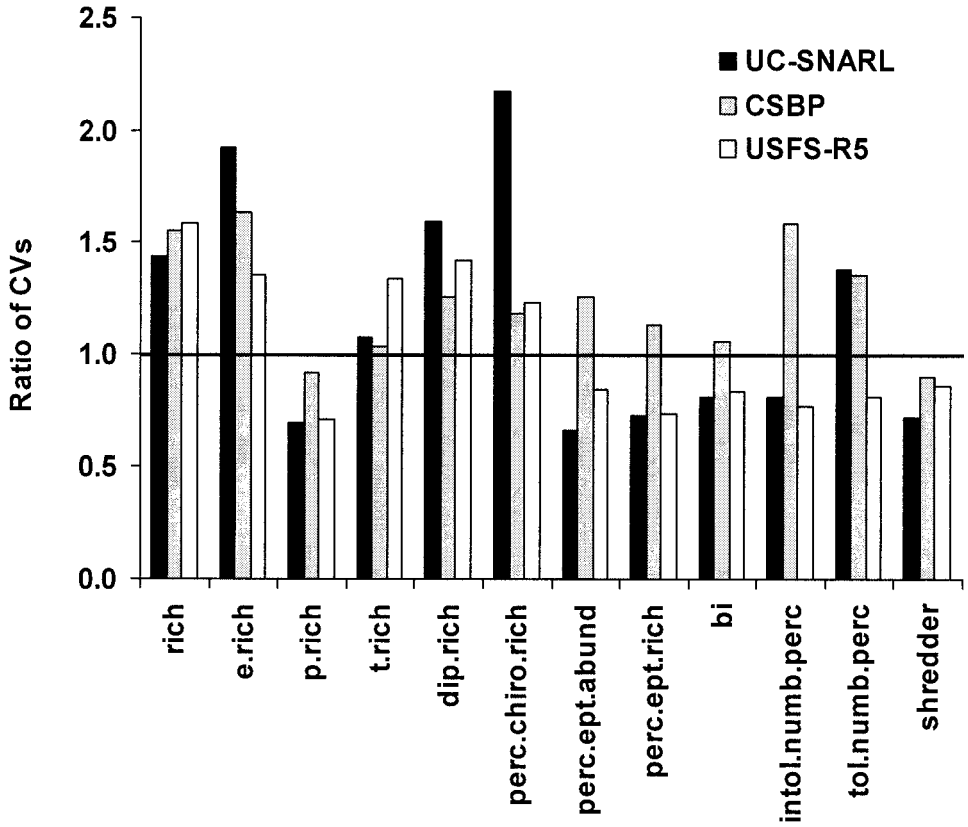

FIG. 2. Ratio of coefficients of variation (CVs) of small to large streams for individual metrics used to develop the multimetric model for reference sites (definitions of small and large streams as in Table 1). Deviation from a ratio ~1.0 (black horizontal line) indicates metric bias between stream size classes. Method abbreviations as in Table 2 and metric abbreviations as in Table 3.

multimetric models had more misclassifications at the lowest impairment thresholds than the multivariate models. Nevertheless, the differences among methods were minor, and all methods provided low misclassification rates for the 16 test sites (i.e., low Type II errors). For example, at a threshold Type I error rate with the lowest 4 of 24 reference sites excluded (~17th percentile), the Type II errors were reduced to 1 (SNARL and USFS.R5) or 0 (CSBP) misclassifications (0 to ~6%) of the 16 presumed-impaired test sites using the multimetric IBI.

Comparison of methods

Direct comparisons of the ratings of site quality (IBI score or O/E ratio) showed close correspondence between IBI scores or O/E ratios based on the 3 methods at most sites (Figs 3, 4). Pairwise correlations of optimum IBIs between methods were high ($r \geq 0.875$) for all comparisons (Fig. 3A) and were higher ($r \geq 0.916$) when CSBP and USFS.R5 data were standardized to the set of metrics used for UC-SNARL (Fig. 3B). Pairwise correlations of O/E ratios between

TABLE 5. Estimated number of misclassified test sites (Type II error) at specified thresholds of Type I error. Thresholds of Type I error were set at different percentiles of the reference-site distribution of optimized Index of Biological Integrity (IBI) scores or **R**iver **I**nvertebrate **P**rediction **a**nd **C**lassification **S**ystem (RIVPACS)-type model observed/expected (O/E) ratios. The number of test sites was 16 and the number of reference sites was 24. Method abbreviations as in Table 2. Numbers after method names as in Table 4.

Threshold Type I error	Percentile of reference-site distribution	IBI			O/E		
		UC-SNARL-6	CSBP-8	USFS.R5–7	UC-SNARL	CSBP	USFS.R5
Lowest reference-site score	4.2	2	3	4	2	2	2
2nd lowest reference-site score	8.3	1	3	1	2	2	0
3rd lowest reference-site score	12.8	1	3	1	0	2	0
4th lowest reference-site score	16.7	1	0	1	0	1	0

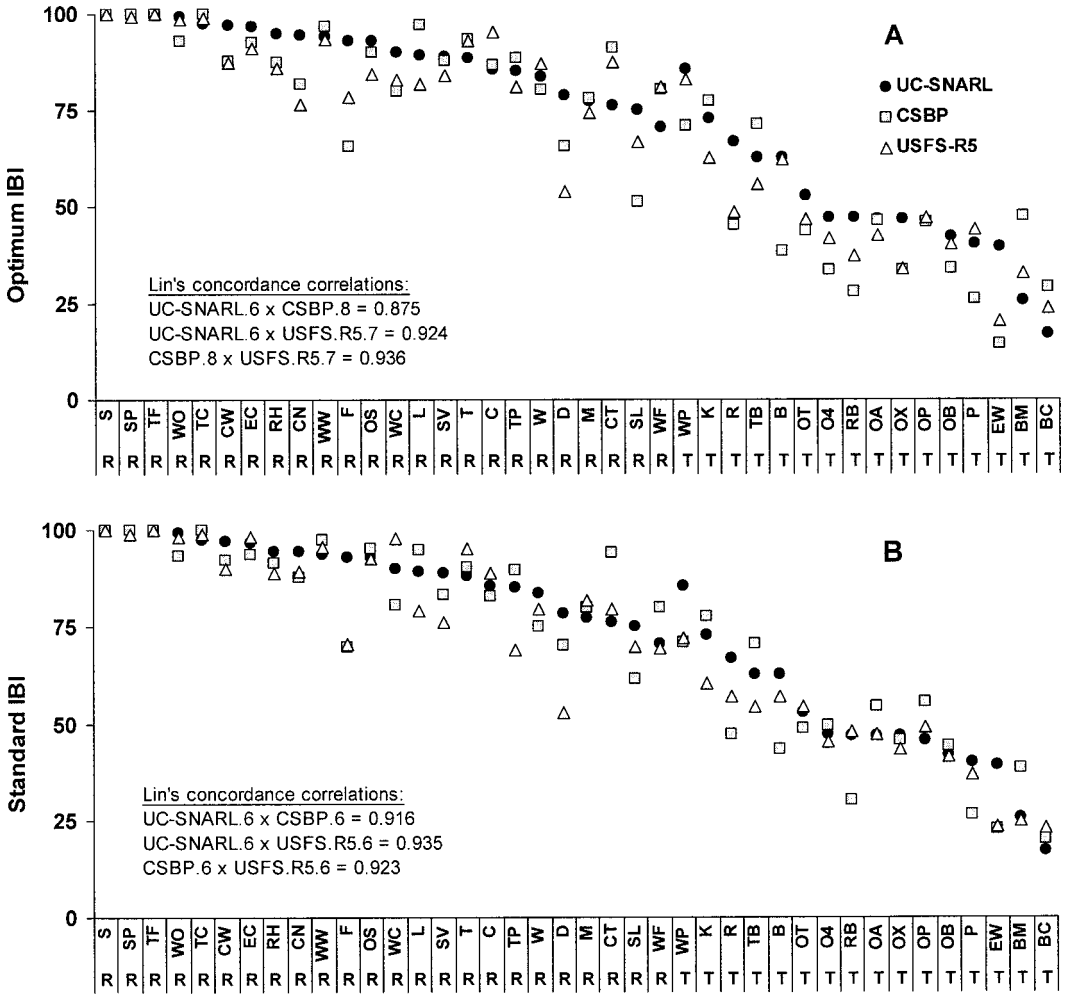

FIG. 3. Multimetric Index of Biological Integrity (IBI) scores based on a metric set optimized for each method (A) and IBI scores standardized to the metrics used in the UC-SNARL method multimetric IBI (B) for all sites. Sites are ordered by UC-SNARL scores. Stream codes as in Table 1 and method abbreviations as in Table 2. R = reference site, T = test site.

methods also were high ($r \geq 0.839$), but were lower than those observed for IBI scores (Fig. 4). The discrepancy among assessments (2 of 3) in placing the WWalker.Pickel (WP) test site in the reference range suggests that livestock grazing on this reach may have had only slight impact on the integrity of the benthic invertebrate community (Fig. 5). Kirman (K) and Slinkard (SL) Creeks also were placed by all methods in intermediate IBI and O/E ranges, indicating the reference site (SL) may have been overrated (it is under restoration), and the test site (K) was only moderately impaired (Figs 5, 6).

Inspection of ranked IBIs of all sites for each method shows some differences in the ordering of sites, but all methods display a break in the form of the distribution at an IBI ~75 (Fig. 5A, B, C). Most sites above the break were reference sites, and most below the break were test sites. These graphs also show where reference and test sites intergrade and the extent to which this intergradation affects detection of impaired condition (as in Table 5). Similar graphs for O/E ratios for each method also show an abrupt transition from reference to test site at an O/E ratio ~0.80 (Fig. 6A, B, C), but the transition is less well-defined than for IBI scores. Separate data (DBH, unpublished data) indicate that these thresholds correspond to combined habitat alterations over stressor gradients related to erosion (at >60% fines, sand, and gravel substrate composition), exposed banks and agricultural return flows (at conductivity >200 μS/cm), and bank-vegetation loss (at riparian cover <30%).

Cost/benefit analysis

The cost of field and laboratory efforts for each method was evaluated from records of the time and

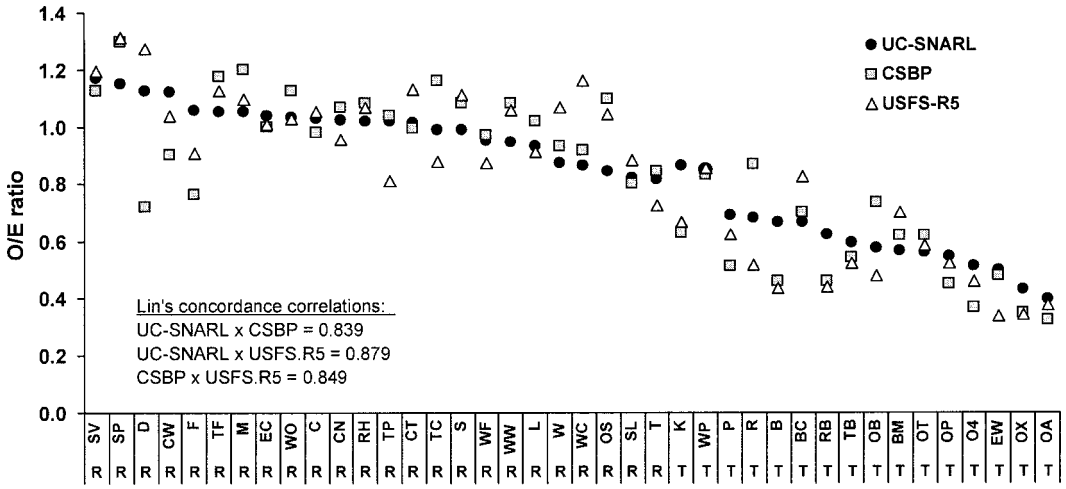

FIG. 4. **R**iver **I**nvertebrate **P**rediction **a**nd **C**lassification **S**ystem (RIVPACS)-type observed/expected (O/E) ratios for each method for all sites. Sites are ordered by UC-SNARL ratios. Stream codes as in Table 1 and method abbreviations as in Table 2. R = reference site, T = test site.

personnel necessary to complete all tasks related to sample collection, processing, sorting, counting, and identification, and including field habitat surveys. Field effort was nearly equal for all methods and made up a smaller fraction of the total effort than effort in the laboratory (Fig. 7). The number of replicates caused the UC-SNARL method to require 1.5 to 3× the laboratory effort of the CSBP and USFS.R5 methods, respectively. Data analysis efforts were more difficult to evaluate because expertise in statistical methods was more relevant than time requirements. Multivariate analysis involves a stepwise approach to model building that requires knowledge of a complex series of statistical operations, whereas multimetric data analysis uses only a simple combination of scaled metrics for IBI development. Therefore, RIVPACS-type modeling may require a greater initial investment of time or expense in development of analytical tools.

Discussion

Justifying uniform bioassessment methods

The use of differing methods to collect, process, identify, and analyze samples of stream macroinvertebrates for evaluations of water quality creates potential discrepancies in results and in the conclusions drawn. Our study directly addressed how the combined differences between methods affect the comparability of results and assessments, and used a PBMS to assess precision, discrimination, and accuracy. Three dissimilar methods showed only small differences in performance and had closely correlated assessment scores, whether derived from multimetric models or multivariate RIVPACS-type models. The consistent agreement across indicators produced by different bioassessment procedures suggests that output is often directly comparable, data sharing is possible, and specified alternative techniques can be applied confidently to the measurement of biological health in streams.

Conformity in bioassessment methods has been identified as an important step toward enabling data sharing among agencies. Use of uniform methods could permit assessments over broad geographic areas using data combined from different sources, decrease duplication of effort (cost savings), and minimize the potential for conflicting interpretation of results. A common foundation for evaluating water-quality status and trends would mean that reports of ambient conditions over broad regions could be unambiguously understood by the public without any need for adjustment of results.

An alternative view is that data sharing among programs that together could cover large geographic areas is not often useful or advisable. Stream communities in distant areas share less biogeographic affinity than communities in adjacent areas (especially in the western US) and may not have common species pools contributing to their assembly. Differences between streams in large geographical areas may have less to do with detecting impairment than with natural differences in faunal composition. Furthermore, duplication of effort by different management jurisdictions is probably infrequent, and agreement among results from different approaches may actually strengthen interpretation, making conclusions more reliable through cross-confirmation. In situations where sharing of data could demonstrably improve bioassessment efforts, a means of calibrating or converting

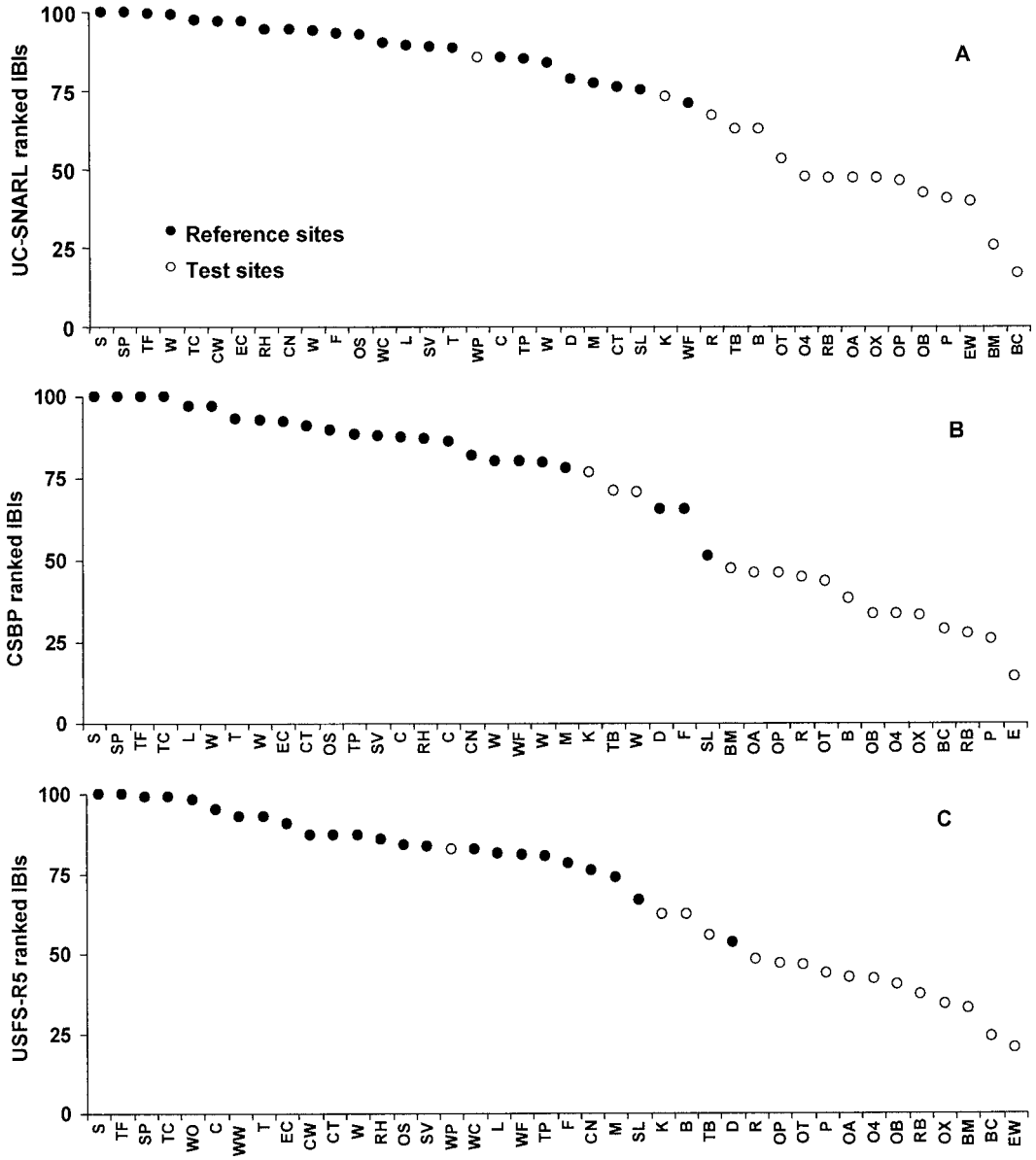

FIG. 5. Distribution of ranked optimized multimetric Index of Biological Integrity (IBI) scores for each site and method. A.—University of California Sierra Nevada Aquatic Research Laboratory Protocol (SNARL; Lahontan Water Quality Board). B.—California Stream Bioassessment Protocol (CSBP; California Department of Fish and Game). C.—Utah State University Protocol (US Forest Service Region 5; USFS.R5). Stream codes as in Table 1. Site order varies by method.

results to the lowest-common-denominator method used might be all that is necessary to facilitate the exchange. One also could argue that programs or monitoring projects with an established legacy of information through long-term data collection should maintain methods for the sake of internal consistency rather than undertake expensive resampling of existing study sites. Thus, as we evaluate the need for data sharing, we must consider what could be gained and what might be lost or ineffectively achieved, given differing monitoring objectives.

The value of independent assessments

PBMS contrasts showed broad agreement in test-site assessments and similar accuracy in distinguishing reference from test sites among the methods despite some differences in individual-metric and final-model precision that led to small differences in method sensitivity. IBI scores produced by the 3 methods were in agreement in distinguishing impairment (nonattainment) for 15 of 16 test sites exposed to disturbance from livestock grazing and channel alteration when the threshold Type I error was set at the $\sim 17^{th}$

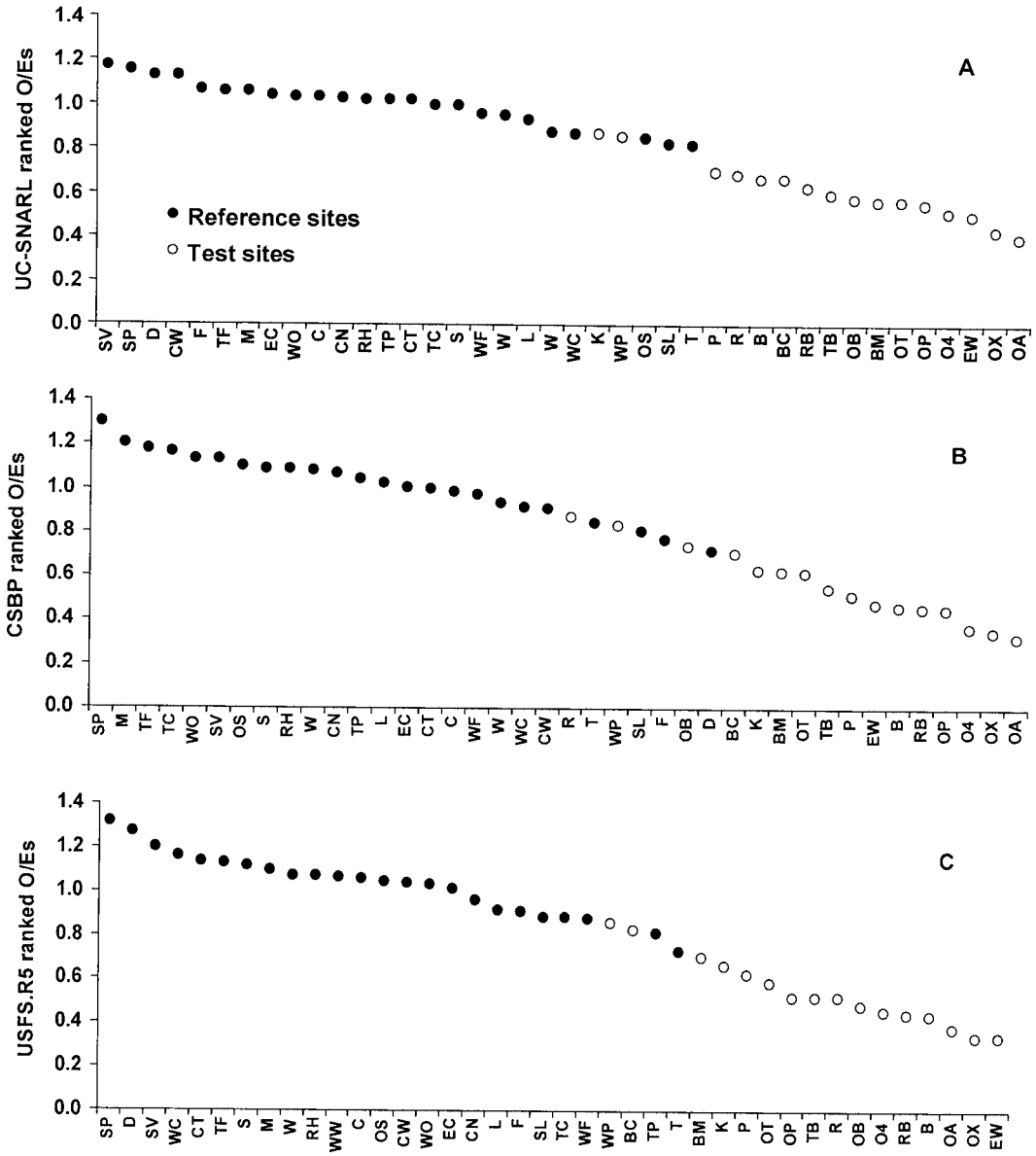

FIG. 6. Distribution of ranked **R**iver **I**nvertebrate **P**rediction **a**nd **C**lassification **S**ystem (RIVPACS)-type observed/expected (O/E) ratios for each site and method. A.—University of California Sierra Nevada Aquatic Research Laboratory Protocol (SNARL; Lahontan Water Quality Board). B.— California Stream Bioassessment Protocol (CSBP; California Department of Fish and Game). C.—Utah State University Protocol (US Forest Service Region 5; USFS.R5). Stream codes as in Table 1. Site order varies by method.

percentile of reference sites (corresponding approximately to an IBI score <76–78 and an O/E ratio <0.85–0.88). The single IBI assessment that was not in agreement was for a site (WP) where impact may have been minimal because livestock grazing effects were not evident in sediment deposition. Here IBI scores matched the reference condition for SNARL and USFS.R5 methods, and fell below the threshold for the CSBP method. The RIVPACS-type model assessments of this site showed O/E ratios just below reference attainment for all methods. O/E ratios from 3 methods disagreed for only one test site (Rush:R), where the CSBP produced a score indicating attainment and scores from the other methods fell below the threshold (Fig. 6).

Just as independent tests of results from clinical trials are important to ensuring public health safety, so may independent assessments provide confidence in judging whether stream biological integrity is intact. Repeated tests provide greater certainty when results agree, especially when differences in methods provide multiple lines of evidence that support the same

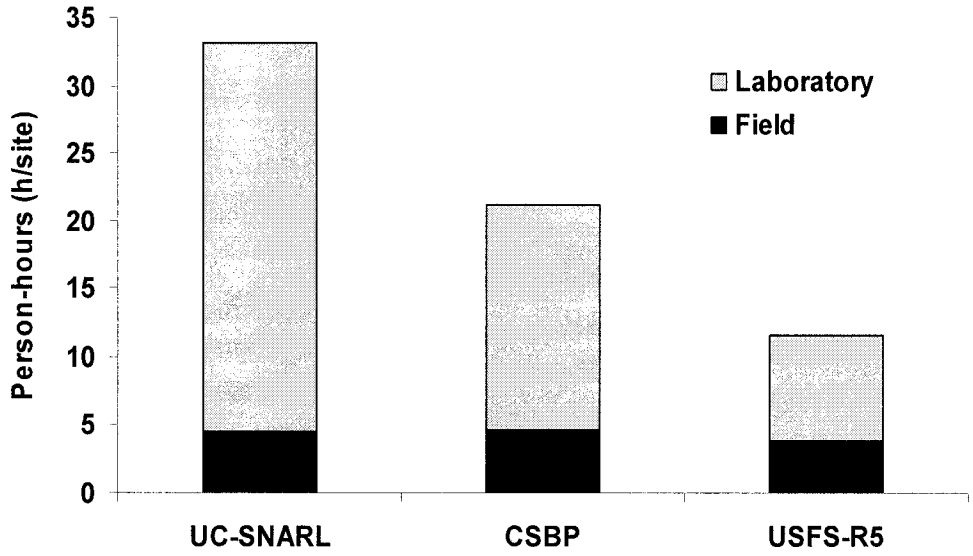

FIG. 7. Total person-hours of effort spent completing field and laboratory tasks for a single site or reach bioassessment survey (sample collection, habitat survey, sample processing, sorting, identifications, and counts) for each method. Method abbreviations as in Table 2.

conclusion. Differing test results give reason to question the assessment. This type of information is valuable for ensuring that errors are minimized beyond Type I and II statistical levels and that aquatic resources are protected or restored where problems are most clearly identified. The results of our study showed that a high degree of certainty for assessments of biological condition can be obtained through the collective consideration of multiple data sources. Integrated assessments are not simply redundant information, but where added certainty is required (where risks and costs are high), conclusions may be reinforced (or cast in doubt) if separate sources of data are considered.

Agreement among methods

One approach to determining the agreement or reproducibility of measurements between methods is Lin's concordance correlation (Zar 1999). Pairwise comparisons of results from different methods agreed closely for optimum-metric IBIs (Fig. 3A), were slightly improved for standard-metric IBIs (Fig. 3B), and were slightly reduced for O/E ratios between methods (Fig. 4). These contrasts suggest that between-method data sharing and integration may be simpler for IBI scores than for O/E ratios, and that between-method data sharing may be further improved simply by calibrating metric sets.

Spearman rank correlation can be used appropriately for comparing orders of site scores when comparing bioassessment results that are scaled differently (IBI vs O/E). Spearman correlation coefficients were lower for IBI scores vs O/E ratios ($r = 0.70$–0.86) than for IBI scores between methods ($r = 0.88$–0.98), but were similar to coefficients for O/E ratios between methods ($r = 0.79$–0.84). The best cross-analysis correlations were between CSBP IBI scores and O/E ratios, suggesting that the lower taxonomic resolution of the CSBP method may produce RIVPACS-type models that match the behavior of multimetric models. The CSBP method used only family or subfamily identification of mites and midges, reducing the emphasis on these common components of the benthic stream fauna that might appear in multivariate models. It is plausible that CSBP O/E ratios may more closely resemble the multimetric scores because the IBIs constructed in our study did not use metrics specific to mites and midges (with the exception of 1 metric [of 7] used for the USFS.R5 IBI; Table 3).

The IBI scores and O/E ratios yielded comparable assessments over all sites despite differences in their computation. However, multimetric and multivariate approaches to contrasting test and reference sites use procedures that are not consistent among data sets. Multimetric calculation of a single IBI involves selection, standardization, and summation of the metrics that produce the best separation of reference from test sites or the best correlations with stressor gradients. Thus, the number and type of metrics used to compute the IBI may vary from one data set or project to another (though some programs use a fixed suite of metrics, as in the Pacific Northwest; Karr 1998). Construction of a multivariate RIVPACS-type

model involves subjective decisions regarding similarity measures, clustering algorithms, discriminant model building, and probability of capture threshold. The predictor variables and their coefficients in the discriminant models change from one data set to another such that test sites are evaluated only in the context of a circumscribed group of reference sites. This lack of uniformity and other potential biases and limitations of both multimetric indices and RIVPACS-type models (reviewed by Suter 1993, Karr and Chu 2000, Norris and Hawkins 2000) notwithstanding, our results suggest that similar assessments of impairment can be obtained using either of these analytical tools, even for data sets derived using differing field and laboratory bioassessment methods.

Deciding among methods

The methods compared here had substantial differences in protocol, but they were nearly equivalent in accuracy of discriminating predefined reference from test sites. The complementary results obtained when using different field and laboratory methods and analytical tools argue that the outputs from all approaches were robust, data and impairment assessments were interchangeable, and these different lines of evidence provide mutual support rather than confusion in interpretations of the biological-integrity component of water quality. However, the costs with regard to laboratory time required to achieve results were considerably greater for the most intensive method (UC-SNARL) than for the other 2 methods, for only a small gain in potential sensitivity in discriminating impaired condition.

Direct comparisons of methods provides an important foundation for integrating and guiding bioassessment programs. Methods comparisons such as our study can provide guidance for choosing between alternate methods or combining data for biomonitoring programs. Options for ambient monitoring and biocriteria development include: 1) continue using existing methods if assessments are in agreement (high correlations of IBIs and O/Es suggest data may be shared directly if necessary), 2) adopt the most cost-effective method where results show equal outcomes in assessment conclusions (the lowest cost method), 3) default to the method with the best potential for data-sharing in biocriteria development (having the most comprehensive data set, provided it meets DQOs), 4) use the method with the most precision, sensitivity, and potential for distinguishing moderate levels of impairment, and detecting biological transitions at stressor thresholds that help in defining tiered aquatic life uses, 5) consider integrating results of different methods to increase assessment certainty, and 6) convert data from the most intensive method(s) to the lowest common denominator (e.g., use the same metrics, adjust taxonomic resolution, use fixed counts) to correct any systematic bias in data sets that must be combined.

Acknowledgements

We thank the State of California Water Resources Control Board, Surface Water Ambient Monitoring Program, and US Forest Service for supporting our work. In particular, Tom Suk of the Lahontan Regional Board and Joseph Furnish of Region 5 USFS encouraged the development and maturation of our project. We also thank Chuck Hawkins for providing valuable insight in his review of the initial manuscript, Dave Lorenz for developing and sharing the S-Plus clustering code, and Jerry Diamond and Michael Barbour for helping to develop a forum for the importance of method comparisons. Peter Ode, Jim Harrington, and Andy Rehn contributed to useful discussions on achieving data comparability. We thank Ryan King and Pamela Silver for helping us to craft the final product, and anonymous referees for comments leading to the improvement of this paper.

Literature Cited

BARBOUR, M. T., AND J. GERRITSEN. 1996. Subsampling of benthic samples: a defense of the fixed-count method. Journal of the North American Benthological Society 15: 386–391.

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 481-B-99–002. Office of Water, US Environmental Protection Agency, Washington, DC.

BARBOUR, M. T., AND C. HILL. 2003. The status and future of biological assessment for California streams. Division of Water Quality, California State Water Resources Control Board, Sacramento, California. (Available from: http://www.swrcb.ca.gov/swamp/reports.html)

CAO, Y., C. P. HAWKINS, AND A. D. STOREY. 2005. A method for measuring the comparability of different sampling methods used in biological surveys: implications for data integration and synthesis. Freshwater Biology 50: 1105–1115.

CARTER, J. L., AND V. H. RESH. 2001. After site selection and before data analysis: sampling, sorting, and laboratory procedures used in stream benthic macroinvertebrate monitoring programs by USA state agencies. Journal of the North American Benthological Society 20:658–676.

CEH (CENTER FOR ECOLOGY AND HYDROLOGY). 2003. The RIVPACS type approach to bioassessment of rivers. Natural Environment Research Council, Dorset, UK.

(Available from: http://www.dorset.ceh.ac.uk/River_Ecology/River_Communities/Rivpacs_2003/rivpacs_introduction.htm)

COURTEMANCH, D. L. 1996. Commentary on the subsampling procedures used for rapid bioassessments. Journal of the North American Benthological Society 15:381–385.

DIAMOND, J. M., M. T. BARBOUR, AND J. B. STRIBLING. 1996. Characterizing and comparing bioassessment methods and their results: a perspective. Journal of the North American Benthological Society 15:713–727.

FORE, L. S., AND J. R. KARR. 1996. Assessing invertebrate responses to human activities: evaluating alternative approaches. Journal of the North American Benthological Society 15:212–231.

GURTZ, M. E., AND T. A. MUIR (EDITORS). 1994. Report of the interagency biological methods workshop. Open File Report 94–490. US Geological Survey, Raleigh, North Carolina.

HAWKINS, C. P., R. H. NORRIS, J. N. HOGUE, AND J. W. FEMINELLA. 2000. Development and evaluation of predictive models for measuring the biological integrity of streams. Ecological Applications 10:1456–1477.

HOUSTON, L., M. T. BARBOUR, D. LENAT, AND D. PENROSE. 2002. A multi-agency comparison of aquatic macroinvertebrate-based stream bioassessment methodologies. Ecological Indicators 1:279–292.

JACKSON, S. 2004. Using biological assessments to refine designated aquatic life uses: EPA/State workgroup. National Biological Assessment and Criteria Workshop, Coeur d'Alene, Idaho. (Available from: http://www.epa.gov/waterscience/biocriteria/modules/)

KARR, J. R. 1998. Rivers as sentinels: using the biology of rivers to guide landscape management. Pages 502–528 *in* R. J. Naiman and R. E. Bilby (editors). River ecology and management: lessons from the Pacific Coastal Ecoregion. Springer, New York.

KARR, J. R., AND E. W. CHU. 2000. Sustaining living rivers. Hydrobiologia 422:1–14.

LENAT, D. R., AND V. H. RESH. 2001. Taxonomy and stream ecology—The benefits of genus- and species-level identifications. Journal of the North American Benthological Society 20:287–298.

LIN, L. I.-K. 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45:255–268.

MARCHANT, R., A. HIRST, R. H. NORRIS, R. BUTCHER, L. METZELING, AND D. TILLER. 1997. Classification and prediction of macroinvertebrate assemblages from running waters in Victoria, Australia. Journal of the North American Benthological Society 16:664–681.

MOSS, D. 2000. Evolution of statistical methods in RIVPACS. Pages 25–37 *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.

MOSS, D., M. T. FURSE, J. T. 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.

MOSS, D., J. F. WRIGHT, M. T. FURSE, AND R. T. CLARKE. 1999. A comparison of alternative techniques for prediction of the fauna of running-water sites in Great Britain. Freshwater Biology 41:167–181.

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

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

RESH, V. H., AND J. K. JACKSON. 1993. Rapid assessment approaches to biomonitoring using benthic macroinvertebrates. Pages 195–233 *in* D. M. Rosenberg and V. H. Resh (editors). Freshwater biomonitoring and benthic macroinvertebrates. Chapman and Hall, New York.

RESH, V. H., AND E. P. MCELRAVY. 1993. Contemporary quantitative approaches to biomonitoring using benthic macroinvertebrates. Pages 159–194 *in* D. M. Rosenberg and V. H. Resh (editors). Freshwater biomonitoring and benthic macroinvertebrates. Chapman and Hall, New York.

REYNOLDSON, T. B., R. C. BAILEY, K. E. DAY, AND R. H. NORRIS. 1995. Biological guidelines for freshwater sediment based on **BE**nthic **A**ssessment of **S**edimen**T** (the BEAST) using a multivariate approach for predicting biological state. Australian Journal of Ecology 20:198–219.

REYNOLDSON, T. B., R. H. NORRIS, V. H. RESH, K. E. DAY, AND D. M. ROSENBERG. 1997. The reference condition: a comparison of multimetric and multivariate approaches to assess water-quality impairment using benthic macroinvertebrates. Journal of the North American Benthological Society 16:833–852.

SUTER, G. W. 1993. A critique of ecosystem health concepts and indexes. Environmental Toxicology and Chemistry 12:1533–1539.

VINSON, M. R., AND C. P. HAWKINS. 1996. Effects of sampling area and subsampling procedures on comparisons of taxa richness among streams. Journal of the North American Benthological Society 15:392–399.

ZAR, J. H. 1999. Biostatistical analysis. 4th edition. Prentice Hall, Upper Saddle River, New Jersey.

Received: 30 March 2005
Accepted: 16 January 2006