Performance of Different Bioassessment Methods From California: Side by Side Comparisons of Field, Laboratory and Analysis Procedures for Streams of the Eastern Sierra Nevada

David B. Herbst and Erik L. Silldorff

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

Supported by the Lahontan Regional Water Quality Control Board

Contract 9-191-160-0

With additional support from the US Forest Service, Region 5

November 26, 2004

Correspondence: herbst@lifesci.ucsb.edu

(760) 935-4536

FINAL REPORT

ABSTRACT

Bioassessment programs within California, in other states, and among federal agencies have used different methods to collect and analyze stream invertebrate samples. While this has created concern and confusion over the comparability of these many disparate sources of data, few studies have attempted to evaluate differences in performance of these methods and reconcile the results produced from varied approaches. To obtain directly comparable data sets we conducted concurrent sampling at 40 sites using three bioassessment methodologies that differed at each stage, from field sample collection to laboratory processing and data analysis (California Stream Bioassessment Protocol, Region 5 US Forest Service, and UC-SNARL methods). We used a performance-based methods system to compare precision, bias, discrimination, accuracy, and correlations among multimetric and predictive model output assessment scores. Reference and test sites were first identified using local and upstream watershed disturbance criteria, and invertebrate community measures and models then developed to discriminate between these site classes. Differences in performance between methods were small, and the assessment scores were highly correlated and accurately distinguished test and reference sites. An examination of the association of impaired biological integrity with environmental stress gradients showed that the method using most replication and sample counts provided the clearest resolution of stressor effect thresholds and intermediate levels of impairment. Despite slight differences in performance and stress detection, these results demonstrate that even substantially different methods of bioassessment yield very similar, effective discrimination of impaired biological condition. This conclusion did not depend either on the data analysis approach used since both multimetric IBIs and multivariate RIVPACS predictive models were in close agreement. Furthermore, simple data conversion procedures can be used to calibrate the data from more intensive sampling protocols to a common 500 fixed-count that provides a uniform and consistent data form for analysis and biocriteria development. Existing data may thereby be integrated with future data collected using a unified standard approach.

INTRODUCTION

Surveys of the different protocols employed among federal, state and local programs for stream bioassessment have revealed 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 (Barbour and Gerritsen 1996, Courtemanch 1996, Lenat and Resh 2001, Resh and McElravy 1993, Resh and Jackson 1993, Vinson and Hawkins 1996). Comparisons of how bioassessment data are analyzed have also been presented, as contrasts of different analytical approaches based on 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. 2004) is to compare bioasessment results for coordinated side-by-side contrasts of methods that differ at each stage of the process, from field collection through laboratory processing and identification, to the analytical assessment of biological impairment. This provides both the most realistic context for evaluation of the results produced from different monitoring programs, and the information needed for the calibration and interconversion of methods for promotion of interagency cooperation in developing biological criteria for water quality.

Organized bioassessment programs for the monitoring of water quality have been in operation in California since about 1993, the date of the first meeting of the California Aquatic Bioassessment Workgroup. Several large programs involving extensive data sets have developed since that time including the work of the Aquatic Bioassessment Laboratory of the Department of Fish and Game, the US Forest Service on National Forest Lands, and the Lahontan Regional Water Quality Control Board in watersheds east of the Sierra Nevada (conducted by the University of California Sierra Nevada Aquatic Research Laboratory). Each of these programs has used differing field and lab protocols for sampling, processing, identifying, and analyzing data. Additional programs with still other methodologies exist in California, but this study contrasts these three larger programs that were also emphasized in a recent report on the status of bioassessment in California (Barbour and Hill 2003).

The use of a performance-based method system (PBMS) has been suggested to evaluate the comparability of bioassessment methods (see Barbour et al. 1999). The PBMS compares bioassessment results to a performance standard, and if performance measures meet or exceed the standard, the method is considered acceptable for use in monitoring. The performance standards may be defined based on required data quality objectives of a program, or relative to a reference, or accepted, methodology. Methods are compared based on performance characteristics that include precision, bias, discrimination power (of test from reference), and accuracy – especially in minimizing type II error rate (i.e. the frequency of not identifying a known impaired site to be impaired). Using the PBMS will allow differences between bioassessment methods to be resolved and decisions made on what method(s) are most appropriate for defined data quality objectives.

The objectives of this study were to (1) compare common bioassessment methods used in California to evaluate differences in meeting PBMS criteria, (2) determine whether any bias in representing community structure exists between methods and if so, how data sets might be calibrated, (3) evaluate whether differences in field collection, laboratory processing, and data analysis affect the outcome of assessment of biological impairment, (4) provide explicit descriptions of the steps involved in predictive model development and IBI construction, and (5) examine the correspondence between IBI scores produced by different methods and environmental stressor gradients. Relative costs and benefits related to the discrimination of moderate impairment levels were also considered.

METHODS

Forty streams of varied size (order, mean width, watershed area) were selected to represent both least-impaired reference sites and a variety of impaired sites in a geographic region restricted to the east slopes of the Sierra Nevada (Great Basin drainages between about 37º and 40º north latitude, and 118º and 120º west longitude). The same 40 streams were sampled at the same sites and on the same dates using each of the 3 primary methods, although the number of total samples taken by each method differed. The impaired (test) sites were selected from disturbed landscapes, with clear

physical habitat degradation related mainly to livestock grazing and altered channel geomorphology (erosion and sediment pollution). Reference streams were selected based on an initial screening for low upstream density of road crossings (a measure of watershed development), or low local bank erosion and minimal exposure to known local and upstream landscape disturbance (refer to Table 1).

For the development of IBIs appropriate to habitat type, streams were grouped into large and small classes based on channel size (the wetted perimeter stream width) and upstream watershed area. Small streams were those less than about 3 meters wide with watershed areas mostly less than 20 km^2. Large streams were greater than 3 meters width with watershed areas mostly greater than 20 km^2. Thirty streams were grouped in the large size class (19 reference and 11 test), and twelve streams in the small size class (8 reference and 4 test). Two streams were borderline and were thus used in both stream classes.

Sampling Protocols

Physical Habitat and Water Chemistry

Each site was defined as a 150-meter length study reach, located by GPS-UTM coordinates and elevation (near lower end of each site). The longitudinal distribution and length of riffle and pool habitats were first delineated, and flagged for marking transect locations. Slope of the reach was measured with an autolevel and stadia rod, and sinuosity estimated as the ratio of 150 meter to the linear distance between the upper and lower ends of the reach. Bank and channel habitat were measured over the length of each reach along 15 transect cross-sections spaced at 10 meter intervals. Water depth, substrate type and current velocity were measured at five equidistant points on each transect along with stream width, bank structure (cover/substrate type and stability rating), riparian canopy cover, and bank angle. Bank structure between water level and bankfull channel level was rated as open, vegetated, or armored (rock or log), and as stable or eroded (evidence of bank erosion, collapse or scour scars). 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 (sum of grid points for measurements taken at each stream edge and at mid-

stream facing up- and downstream). The type and amount of riparian vegetation along the reach was also estimated by qualitative visual evaluation. The embeddedness of cobble size substrate was estimated as the volume of the rock buried by silt or fine sand for 25 cobbles (encountered during transect surveys or supplemented with random selected cobbles). Discharge was calculated from each transect as the sum of one-fifth the width times depth and current velocity at each of the five transect points, and averaged. A suite of basic water chemistry and related parameters were also measured at each site: dissolved oxygen, conductivity, pH, temperature, and turbidity. Documentation also included photographs taken at mid-stream looking upstream at 0, 50, and 100 meters, and downstream at 150 meters.

UC-SNARL Protocol – Lahontan Water Quality Board

Benthic macroinvertebrate sampling consisted of 5 replicate samples taken in riffle zones using a 30-cm wide D-frame kick-net, having a 50 cm length bag with 250 µm mesh. Each replicate was composed of a composite of three 30x30 cm sample areas (0.09 m^2 each, 0.27 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 seconds to one minute so that both mobile and attached invertebrates were washed off and into the downstream net being held against the bottom. This composite of microhabitats was intended to represent varied microhabitat conditions and reduce the variability among replicate samples. Samples were processed in the field by washing and removing large organic and rock debris in sample buckets 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 shelled snails or other molluscs. Elutriated and inspected sample fractions were then preserved in ethanol, and a small volume of rose bengal stain was added to aid in lab processing. Invertebrate field samples were subsampled in the laboratory using a rotating drum splitter, sorted under a stereo microscope at 10X magnification, and identified to the

lowest practical taxonomic level (usually genus; species or species groups when possible based on the availability of taxonomic keys including midges and mites; oligochaetes and ostracods were not further identified). A minimum count of 250 organisms was removed from each replicate for identification (in practice averaging about 500 individuals). All sample sorting was conducted to achieve <5% error in removal, and quality control verifications of every taxon identified in every sample was performed by the lead author. Unprocessed sample remnants were also searched (using a 3X magnification visor) for rare and large taxa not encountered in the processed sample, and single counts of those found were added to the total.

California Stream Bioassessment Protocol (CSBP) - Department of Fish and Game

Samples collected using this method were taken within the same study reach at locations adjacent to the first, third and fifth SNARL sample replicates. Three replicate CSBP samples were taken using a 30-cm wide D-frame kick-net fitted with a 500 µm mesh net 50-cm in length. Each replicate consisted of a composite sample taken from 3 locations as in the SNARL method except that the collection areas were each 30-cm wide by 60-cm long (0.18 m^2 each, 0.54 m^2 total). Samples were field-processed, preserved and stained as described above for the SNARL method. Laboratory subsampling was performed by spreading the field sample over a large shallow white pan with a grid drawn on the bottom, and random-numbered grid sectors were selected and all organisms removed until reaching a fixed-count of 300 individuals. The benthic invertebrates were identified at the same level of taxonomic resolution as the SNARL method except that midges were identified only to subfamily and all mites were left at Hydracarina. Quality control checks of lab processing and identifications were performed as for the SNARL samples. A rare and large taxa search was also performed as above.

Utah State University Protocol – Region 5 US Forest Service (R5.USFS.USU)

Samples were obtained as a single composite taken at 8 locations, each 30x30 cm in area (0.09 m^2 each, 0.72 m^2 total), in the 4 longest riffle units in the study reach (2 in each riffle unit selected at random from a 9-point grid). When fewer riffle units were available, locations were assigned in proportion to the length of each unit. Samples were

taken using a 30-cm wide D-frame kick-net fitted with a 500 µm mesh net 50-cm in length. Samples were field-processed, preserved and stained as above. Subsampling was performed as in the CSBP method but to a fixed count of 500 organisms. Identifications were made at the same level of taxonomic resolution as in the SNARL method, including midges and mites to genus and some species groups. Quality control checks of lab processing and identifications were performed as for the SNARL samples, as were checks for rare and large taxa. Methods differences are summarized in Table 2.

Analytical Methods

Data collected with the UC-SNARL and CSBP methods are typically analyzed using the multi-metric calculations recommended by the U.S. Environmental Protection Agency (Barbour et al. 1999) while data collected with the R5.USFS.USU method are typically analyzed using a series of multivariate statistical models first developed in Great Britain and referred to as the River Invertebrate Prediction and Classification System (RIVPACS). In order to evaluate the field/lab and analytical methods in a controlled manner, we analyzed all 3 sets of methods using both the multi-metric models and the RIVPACS predictive models.

Data Preparation

For the CSBP and R5.USFS.USU methodologies, large and/or rare organisms were sorted and distinguished from the organisms obtained in the subsampled section of the sample. To avoid overestimating the abundance of these large and/or rare invertebrates, we first adjusted the final counts per sample by the fraction of the sample which was subsampled. Following this adjustment, the large and/or rare invertebrates were added to the sample data.

Multi-Metric Calculations

Our calculation of a multimetic index (referred to herein as an Index of Biological Integrity or IBI) closely follows the recommendations and procedures outlined in the U.S. EPA Rapid Bioassessment Protocol document (Barbour et al. 1999).

Metric Choice

We calculated 69 metrics for each sample across the 3 methodologies. These 69 metrics were created from 28 metric classes by slightly varying the calculation for a metric (e.g., taxa richness standardized to different sampling levels using a rarefaction procedure, dominance by variable numbers of the most common taxa). The full suite of metrics evaluated is listed in the Appendix. We selected 15 of these metrics for inclusion in a composite multimetric IBI. Calculation of these 15 metrics is further detailed in the Appendix.

Evaluation of Candidate Metrics

For all multimetric analyses, the 40 sampling sites were divided into 2 classes of streams, "large" and "small". Two of the 40 streams fell into an intermediate class between the large and small streams. These 2 sites (both were reference / unimpaired sites) were included in both the large and small classes of streams in order to increase the sample size for the reference distribution and to handle the ambiguity of assigning these streams to one of the size classes when they had characteristics of both.

Each individual metric was then evaluated for its discriminatory power by examining the proportion of "test" (i.e., impaired) streams in each stream class that exceeded (or fell below for reverse scale metrics) various quantiles of the reference distribution. The use of overlap based on quantiles essentially evaluates the signal-to-noise ratio by looking, simultaneously, at the separation between the centers of the test and reference distribution while also considering the spread of values around these centers. The sample sizes used for this study (19 large reference and 8 small reference streams) created, at times, discrete jumps between the values for adjacent quantiles. Thus, rather than choosing a single quantile for all comparisons of overlap between reference and test streams, we evaluated this overlap more broadly by considering multiple quantiles for each metric. For the 19 large reference streams, we used thresholds ranging from the lowest/highest value up through the 6^{th} lowest/highest value; these corresponded to the the 5^{th}, 11^{th}, 16^{th}, 21^{st}, 25^{th}, and 32^{nd} quantiles. For the 8 small reference streams, we used thresholds ranging from the lowest/highest value up through the 3^{rd} lowest/highest value; these correspond to the 13^{th}, 25^{th}, and 38^{th} quantiles.

We note here that the definition of an observed quantile (or percentile) for an ordered set of observations is somewhat ambiguous. By some definitions (e.g., S-plus statistical software), the lowest value corresponds to the 0^{th} observed quantile and the highest value corresponds to the 100^{th} quantile. For small sample sizes, however, the minimum and maximum will substantially overestimate and underestimate (respectively) the 0^{th} and the 100^{th} quantile of the true distribution. A more intuitive definition of a quantile, and one which minimizes the bias for small sample sizes, is to include the minimum and maximum value as data points in the tails of the distribution and to define the quantiles as the proportion of sites greater than (or less than) a given ordered value. Thus, instead of the minimum value out of 20 observations representing the 0^{th} quantile, this more intuitive definition would say that 19 of 20 observations (95%) lie above this minimum value and would therefore identify the minimum value as the 5^{th} quantile. This latter and more intuitive definition is followed throughout this report. These two definitions, of course, only represent issues of semantics. The crucial information is what specific cut-off is used (e.g., the minimum value of the reference distribution) and not the label for that cut-off. We highlight these subtle nuances in quantile definitions so that readers are clear what values we have used for cut-offs when we mention the quantile chosen, and to raise critical awareness of the ambiguity in various references to quantiles and percentiles in the literature, particularly for distributions with small sample sizes.

From the suite of tables we generated to document the overlap between test and reference streams for the different quantile thresholds, we repeatedly filtered out metrics based on three criteria in order to narrow the list to a set of core metrics which could be included in an IBI calculation. These three criteria were: 1) Power - the actual magnitude of a metric's discriminatory power (i.e., the proportion of test streams identified as impaired); 2) Consistency - the degree to which a metric provided discriminatory power for both large and small streams as well as across different quantile thresholds; and 3) Uniqueness - the extent to which a metric provided information unique among the remaining candidate metrics and which therefore provided some independent discriminatory power. As these three criteria were simultaneously evaluated and as these criteria could only be partially quantified, this selection of the best metrics for inclusion in an IBI was a somewhat subjective step in the IBI development. It is

possible that somewhat different final subsets of metrics would be selected by a different group of researchers conducting the same analysis because of this inherent subjectivity.

Creation of the IBI

The above selection procedure resulted in the selection of 15 primary metrics as the basis of a multi-metric Index of Biological Integrity (IBI). Following the recommendations in the Rapid Bioassessment Protocols manual (Barbour et al. 1999), we converted the individual scores for the different metrics to standardized scores on a continuous 0-10 scale so that they could be aggregated into a multi-metric IBI. The median value of the reference stream distribution was scored as a 10 and any value greater than or equal to this median reference value likewise obtained a 10. Similarly, the minimum value of the test stream distribution was scored as a 0 since this represented the actual worst empirical value attained in our study and thus the potential lowest value any stream might attain for that metric. Any metric score between the reference median and the test minimum value was scored by interpolating between these two numbers. For example, total taxa richness for our large reference streams using the SNARL methods had a median value of 45 taxa, and the minimum among all large test streams using the SNARL methods was 24 taxa. The standardized scoring for raw taxa richness was therefore:

$$\text{Richness Score} = \begin{cases} 0 & for\ rich = 24\ taxa \\ 10 & for\ rich \geq 45\ taxa \\ 10 \cdot \dfrac{Richness - 24}{45 - 24} & otherwise \end{cases}$$

Values for the 15 core metrics were converted to standardized scores using this procedure. The 15 standardized scores were then added and the total score divided by 1.5 to yield the full multi-metric Index of Biological Integrity (IBI). This final IBI theoretically ranged from 0 to 100 with equal weight given to each of the 15 core metrics.

In order to minimize potential redundancy among the metrics composing the IBI score, we also created another IBI based on a subset of these metrics. This subset was selected such that correlation among metrics was minimized, discrimination was maximized, and conceptual distinctness was maintained among the variables. Six

variables were selected for this reduced IBI (Biotic Index, Taxa richness, Trichoptera richness, Percent EPT richness, Percent Shredders, and Percent Tolerant Taxa), and evaluations of this model indicated comparable performance to the 15-metric IBI in terms of variability and discriminatory ability.

Multivariate Predictive Model (RIVPACS-type Model)

Grouping Reference Sites: Cluster Analysis

The first step in a RIVPACS-type predictive model is to classify reference sites into homogenous groups based solely on the biological information at those sites. A number of potential methods exist for grouping such multivariate data, but many predictive model analyses have used cluster analysis techniques because they have worked as well or better than other grouping methods (Reynoldson et al. 1995, Marchant et al. 1997, Moss et al. 1999, Hawkins et al. 2000). The results from a cluster analysis depend both on the similarity measure used and the clustering algorithm used in the analysis. We used two common methods (Bray-Curtis distance on proportional data; Sorensen's similarity on presence/absence data) and found little difference in the results. As a result, we present the final results based only on the use of Sorensen's similarity for presence/absence data.

Because little consensus has been established on the best clustering algorithms, we used a number of recommended clustering methods and evaluated the consistency of patterns across methods before establishing stream groups in the data. The clustering methods used were: 1) Ward's clustering; 2) Flexible Beta Weighted Pair Group Means with Arithmetic averaging (Flexible Beta WPGMA; Flexible Beta UPGMA was unavailable); 3) Unweighted Pair Group Means with Arithmetic averaging (UPGMA) using Average linkage; and 4) K-means clustering. Analyses were conducted using a specialized clustering procedure written by D.L. Lorentz for S-Plus and verified for select clustering outputs using established analytical procedures in S-Plus and SAS.

Using different values for Beta in the Flexible Beta WPGMA and different values of K in K-means clustering, we evaluated the grouping patterns of sites for the three primary sampling methods (CSBP, R5.USFS.USU, SNARL) and determined the grouping of reference sites that appeared most stable across clustering methodologies and

that gave relatively large numbers of sites in each individual group. We determined that the clustering analyses showed the most distinct partitioning of sites when the data were limited to 3 or 4 groups of sites. The number of stream in each group ranged from 4 to 15 streams, with different sites being grouped together for the SNARL, CSBP, and R5.USFS.USU field and laboratory methods.

As with the multimetric IBI calculations, it is important to note that this discriminant model selection involves a number of subjective decisions about which variables to use in the model, the number of variables to use, and the criteria for a successful model (e.g., apparent and cross validation error rates). Alternative models to those listed above, with either different variables and/or with different numbers of variables, achieved classification rates comparable to the final models chosen but were not selected. The final model, in each case, was that suite and number of variables that we felt provided the highest correct classification rates with appropriate variables.

Predicting Membership in Groups Based on Environment Data: Discriminant Analysis

A separate discriminant analysis model for each of the 3 methods was then constructed where only the environmental characteristics at each site were used to differentiate among the 3 groups of streams identified in the above clustering.

Sites were assigned to groups based on the results from the cluster analysis, and environmental variables were selected for the discriminant model based on two criteria. First, only variables which were unlikely to be affected by human disturbance were included in the model. The candidate variables which met this criterion were: elevation, latitude, longitude, sampling date, azimuth, distance to headwaters, watershed area, slope, depth, width, conductivity, alkalinity, percentage of boulder outcrops, and 2 climatic statistics (annual precipitation, number of days with precipitation) obtained through Climate Source Inc. (these climatic statistics are continuous coverages across the United States based on both observed and modeled climatic data).

Second, we proceeded through a series of manual variable selection procedures to determine the optimal set of variables for discriminating among groups of streams. The performance of each model was determined from both the apparent error rates and cross-validation error rates for each model. The apparent error rates are simply the number of

sites which are classified incorrectly divided by the total number of sites. The cross-validation error rates are more robust measures of model performance because they essentially construct the discriminant model with a single site removed from the data set, predict which group this omitted site would be classified to using the discriminant model, compare this classification to the actual group identity of the site, and then repeat the process for each stream. Because of the independence between model and prediction in the cross-validation error rate, we relied more heavily on this error rate estimate when evaluating competing models.

The final discriminant model was selected as that model providing the lowest apparent error rates and cross-validation error rates with the greatest number of meaningful predictor variables. The final models for each the three methods contained 4 environmental predictor variables:

R5.USFS.USU	Conductivity, Longitude, Azimuth, and Days of Precipitation
SNARL	Alkalinity, Elevation, Sampling Date, and Annual Precipitation
CSBP	Elevation, Sampling Date, Width, and Slope

As with the multimetric IBI calculations, it is important to note that this discriminant model selection essentially involves subjective decisions and the end result depends on the group of researchers conducting the analysis.

The final discriminant models were then used to predict the probability of each site belonging to one of the 3 or 4 groups of reference streams (e.g., Stream K might have a 0.65 probability of being in group 1, a 0.20 probability of being in group 2, and a 0.15 probability of being in group 3). This trio of probabilities of group membership were estimated both for test sites (not previously classified into a group) as well as the original reference streams (whose group membership had already been determined though the cluster analysis based solely on biological data). We used a proportional prior for these predictions so that 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.

Calculating the Expected Number of Common Taxa

The expected number of common invertebrate taxa for each stream (the "E" in the "O/E" metric) was then estimated by predicting the probability that each invertebrate taxon would be present at a site and then summing these probabilities for the most common species. The probability that an invertebrate would be present at a site was calculated as a weighted mean value of the observed proportion of streams where that taxon was found in each of the 3 groups of reference streams. Specifically, for each reference stream group, the proportion of streams with a given taxon was calculated (e.g., *Calineuria californica* was found at 12 of the 15 streams in group 1). The probability that a species was then present at any given stream (reference or test) was calculated by multiplying the proportion of streams in each group with that species by the probability that the stream in question belonged to each group, and then summing these quotients. This procedure for estimating the probability of finding a specific invertebrate taxon at each site was then repeated across all taxa found in this study and across the 3 field/lab methods.

A pair of examples will clarify these calculations. Suppose *Baetis* was found at 8 of 16 streams in reference group 1, 5 of 5 streams in reference group 2, and 4 of 4 streams in reference group 3. For stream X with probabilities of membership (based on environmental conditions) to groups 1, 2, and 3 of 0.75, 0.15, and 0.10, respectively, the final probability that *Baetis* will be present at stream X from this RIVPACS-style model will be 0.625:

$$\begin{aligned}\Pr(\textit{Baetis at stream } X) &= \frac{8}{16}\bullet 0.75 + \frac{5}{5}\bullet 0.15 + \frac{4}{4}\bullet 0.10 \\ &= 0.625\end{aligned}$$

For *Baetis* at stream Z, which has probabilities of being in groups 1, 2, and 3 of 0.05, 0.50, and 0.45, respectively, the final probability of *Baetis* being present at stream Z is 0.975:

$$\begin{aligned}\Pr(\textit{Baetis at stream } Z) &= \frac{8}{16}\bullet 0.05 + \frac{5}{5}\bullet 0.50 + \frac{4}{4}\bullet 0.45 \\ &= 0.975\end{aligned}$$

Thus, for stream X, the lower probability of being in groups 2 and 3 translates into a lower probability for *Baetis* being present at the site. For stream Z, the high probabilities

of being in groups 2 and 3 translates into a high probability of *Baetis* being present at the site.

All invertebrate taxa with a probability of being present at a site less than 0.50 (i.e., less than a 50% predicted probability) were then removed from the analysis. Researchers using RIVPACS-style predictive models have typically found that a moderate cut-off for excluding less common taxa from the analysis (such as 0.50) gives the clearest discrimination between reference streams and impaired streams (Marchant et al. 1997, Hawkins et al. 2000). Essentially, the predictive models perform best when restricted to predicting the occurrence of common taxa.

The final expected taxa richness ("E") for each stream in the study was then calculated as a simple sum of the predicted probabilities for those taxa with probabilities greater than 0.50.

Calculating Observed Number of Common Taxa and the O/E Metric

The observed number of common taxa (the "O" in the "O/E" metric) for each site was then calculated as the sum of the common taxa present which were used to calculate the "Expected" number of taxa. Thus, for every site, all taxa whose predicted probability of being at that site was less than 0.50 were removed from the data for that site. The "Observed" number of common taxa was then the sum of the number of remaining taxa which were actually observed at that site. As such, this observed number of taxa is only an estimate of the number of the most common invertebrate taxa which were present at a given stream. Finally, the O/E metric was calculated for each site as the simple ratio of the "Observed" number of common taxa to the "Expected" number of common taxa.

Cost-Benefit Analysis

Evaluation of alternative assessment approaches requires not only that the performance characteristics are compared but that the cost:benefit ratio of the methods are considered. Sustaining monitoring programs will require that a balance is achieved between the accuracy and utility of the assessment results and the expense in time and cumulative effort. An estimate of the relative cost of each method was obtained from field and lab observations of person-hours required to complete tasks of sample

collection, processing, sorting and identification and counting as well as habitat surveys. The data analysis phase was accounted for in this cost estimation in qualitative terms of the level of expertise and number of steps required to obtain complete results.

Performance Based Method System

A wide variety of metrics were screened for inclusion in IBI development according to how well they produced separation of test and reference sites and were least variable. Screening resulted in selection of 15 metrics, providing a standard system for comparison of the same set of indicators across all methods (see Appendix).

Methods were contrasted based on the following PBMS criteria, as described in technical guidance documents (Barbour et al. 1999, Barbour and Hill 2003, Diamond et al 1996): (1) precision, defined using the coefficient of variation (CV) among reference streams as a standardized measure of variability in the metrics used to develop the IBIs (the number of metrics meeting data quality objectives set at CV values of 20% and 25%), and the CVs of the multimetric IBIs defined for large and small stream classes, and for the observed over expected taxa ratio (O/E) criterion of the multivariate predictive model, (2) bias in applicability to different stream types, as precision differences for a given metric or score when compared between different habitats or ecoregions (defined here as the ratio in CV values between metrics from reference streams of different size class habitat types), where a ratio near 1.0 would indicate least bias, (3) discriminatory power or average sensitivity, defined as the ratio of test to reference mean values (power increased the smaller the ratio) or as the difference between reference and test means divided by the reference standard deviation (sensitivity increases with the magnitude of this value), and (4) accuracy in minimizing type II error (impaired sites identified as such) determined from (a) the proportion of test sites overlapping reference sites after successive removal of the lowest references (the proportion removed corresponding to a quantile or type I error rate), or (b) the distribution of t-statistics calculated for each test site (type II error rate here is the fraction of test sites that do not exceed the t-value at a given type I error rate and degrees of freedom) =

$$\frac{X_{i(test)} - \overline{X}_{reference}}{sd_{\bar{x}\,reference}}$$

Though we could not know *a priori* that test sites were impaired, it could be shown that the test sites were exposed to stress or disturbance and these formed a class distinct from the undisturbed or minimally exposed reference sites (refer to Table 1, listing stream classification).

Inter-calibration Among Methods

Given the inevitable differences among methods for bioassessment surveys across labs, regions, states, and countries, the greatest utility of these bioassessment data will be realized when either calibrations among methods are developed or when the methods can be demonstrated to give inter-changeable results. For the three field and laboratory methods considered in this study, a number of such differences existed that could conceivably produce different assessment results for the same stream. In order to gain the broader ability to use all data collected via these methods in unified assessments of California streams, it is therefore necessary to determine which standardizations or calibrations of data sets are needed (if any) to achieve an acceptably high level of correspondence among results for the different methods.

Given the lower costs associated with single composite samples (e.g., R5.USFS.USU method), many groups are considering adopting this field/laboratory procedure in order to save money and expand the number of sites surveyed given budget constraints. As a result, we focused on conversions of the replicated methods (CSBP, SNARL) toward the single composite samples that are likely to be used in most future monitoring efforts, particularly the SNARL conversion to R5.USFS.USU method since the CSBP to R5.USFS.USU comparison and inter-calibration has recently been completed for a large California data set (Ode et al. *in press*).

Two methods were utilized for comparing the results among methods before and after any data standardization step. First, the final multi-metric IBI scores for each stream were correlated among methods to determine the strength of the relationship between the final results from the two methods. For these analyses, small and large streams were analyzed separately, and then all streams were analyzed together. The

results were generally comparable for the different stream groupings so the results from the analysis of all streams together are therefore presented in this paper. The second method of comparing the results from different methods was to calculate a multivariate dissimilarity (Bray-Curtis distance measure) between the samples for the same stream using the two methods, and then examine the distributions of these similarities across all samples. To provide a scale for these dissimilarity values in terms of expected differences among any two samples taken from a stream, the 10 combinations of pairings of replicate SNARL samples at each stream were used to calculate 10 independent within-method dissimilarity values for replicate samples, these 10 values were averaged for each stream, and then the distribution of these within-method dissimilarities was set as the expected maximum similarity that would likely be obtained among replicate samples using the same methodology.

For data standardization, our analyses focused on three primary differences among the methods: (1) different numbers of invertebrates identified for each stream; (2) different mesh sizes used in the field; and (3) different handling of large and/or rare invertebrates not collected in the subsample. For the different numbers of invertebrates processed and identified for each stream, multiple re-sampling methodologies were thoroughly tested to determine whether standardizing the more intensive efforts (i.e., SNARL and CSBP) to the 500-counts of the R5.USFS.USU method produced more comparable data, and whether the method of re-sampling affected the final comparisons. The two re-sampling approaches most completely evaluated were: (1) a simple pooling of all identified invertebrates and random sampling without replacement from this overall pool; and (2) a stratified random sampling with equal numbers of individuals re-sampled without replacement from each replicate to more closely mimic the equal probability of any invertebrate being collected in the final 500-invertebrate subsample. For the potential bias introduced by different field mesh sizes, we determined the extent and frequency of over-sampling via the 250 µm net by comparing the absolute density estimates and the proportional abundance estimates for each stream between the 250 µm method (i.e., SNARL) and the two 500 µm methods both individually and averaged together. For common taxa showing consistent differences across streams and for both the density estimates and the proportional abundance estimates, probabilistic corrections

to the 250 µm method (i.e., SNARL) were calculated for these taxa. After first standardizing the SNARL data to a 500 invertebrate sample for each stream, the SNARL data were "corrected" for the apparent higher collection rate of these mostly small invertebrate groups with the 250 µm mesh nets. Finally, the different handling of large and/or rare invertebrate groups was evaluated by first standardizing the sampling of the SNARL data to a single 500 invertebrate composite sample for each stream, and then adding the large/rare invertebrates collected for that stream from the R5.USFS.USU samples to the standardized SNARL samples. When possible, multiple re-samplings of the data were performed for each standardization technique to evaluate the variability in the responses.

RESULTS

The average metric precision, expressed as the mean CV value for the 15 metrics selected for IBI development, showed that SNARL > R5.USFS.USU > CSBP for reference groups over all streams (Figure 1). This was also expressed in a greater number of the metrics calculated from SNARL reference data having CV values below data quality objectives set at either 20% or 25% (Figure 2). When metrics were standardized and combined into an IBI, and when the variation in O/E values were compared, little difference between methods was apparent in the CV values, and reference stream scores were all near or below a DQO of 20% (Figure 3).

Comparing the metrics and IBIs calculated from the CSBP data revealed that CV estimates for the large and small stream class types were not equally precise – there was bias in this performance characteristic (small streams are more variable) that did not appear in the SNARL or R5.USFS.USU data (with CV ratios near unity, Figure 4). Based on comparisons of reference and test means, discrimination power shows no substantial differences between methods (Figure 5) and though sensitivity shows some differences between methods, there were no consistent patterns from one measure to another (Figure 6). Correlation among the methods in rating of stream quality though IBI and O/E indicators shows close correspondence for most sites over a gradient of rankings relative to the SNARL method (Figures 7, 8 and 9). Stream reaches of high quality show the best agreement, though there were several instances where CSBP and R5.USFS.USU

methods under-rated reference streams (for both large and small streams) and some tendency for these methods to over-rate test streams relative to SNARL. The agreement among methods placing the site Owens.abovetun in the test range shows this site was misclassified as a reference, and that the WWalker.Pickel test site was unimpaired (Figure 7). Kirman and Slinkard Creeks were placed by all methods in an intermediate IBI range for small streams, but the reference (Slinkard) may have been over rated, and the test (Kirman) only moderately impaired (Figure 8). Correlations between method pairings were all high (R^2>0.80, except small stream IBI SNARL vs CSBP, R^2=0.75). Results were similar for the 6-metric IBI, but only the 15-metric IBI results are presented for brevity. The O/E values generated from the predictive modeling analysis also show that the methods produced similar ratios for each stream (Figure 9).

The accuracy in assessment is related to the discrimination and sensitivity measures, but more explicitly defines the chances of making errors in not identifying a test stream that is impaired (type II error), or in indicating that an unimpaired (or reference) stream is impaired (type I error). Another way to think about these errors is that minimizing type II error would aid in the protection of the biological integrity of aquatic habitats while minimizing type I errors serves the regulated community in providing reasonable standards in the designation of impairment because it implies a lower criterion for reference or unimpaired sites. It is also important to note that the error rates defined assuming those sites defined *a priori* as test sites truly are impaired may lead to an over-estimate of type II error if these sites were actually misclassified.

There are differing ways that a threshold criterion for designating impairment can be defined. Successive removal of the lowest scoring reference sites from the pool of references gives the quantile for type I error as that fraction of the number of sites removed. The fraction of test sites still overlapping the resulting range is the type II error rate. Alternatively, the threshold for impairment can be defined according to the standard deviation of the reference data set, with type I errors corresponding to the mean minus 2 SD (alpha=0.05), 1.65 SD (=0.10), or 1.3 SD (=0.20). Using either the quantile or the standard deviation, accuracy in assessment can be attained for type II error in the range of 80-90% at about this same level of type I error rate (Figures 10 and 11).

The cost of field and laboratory efforts for each method was evaluated from records of the effort necessary to complete all tasks related to sample collection, processing, sorting, counting and identification, and including field habitat surveys (Figure 12). The field efforts for all methods were nearly equal and comprised a smaller fraction than that required in the laboratory where the number of replicates in large part contributed to the SNARL method requiring 1.5 to 3 times the effort of CSBP and R5.USFS.USU, respectively. Data analysis efforts were more difficult to evaluate because expertise in statistical methods was more relevant than time requirements. Multivariate analysis involves a step-wise approach to model development requiring knowledge of complex methods while multimetric data analysis used only a simple combination of scaled metrics for IBI development. Predictive modeling may therefore require a greater initial investment of time or expense in statistical consultation.

Discriminating intermediate levels of impairment is an important aspect of certainty in defining the extent of impact to a site and also provides a basis for the regulatory process of assigning different categories of aquatic life use attainment. When sharp transitions exist in the distribution of IBI values from impaired test sites to unimpaired reference sites, and few reference sites are transitional, it should be possible to distinguish with greater certainty where environmental thresholds for impairment exist, and determine which test sites have reduced or intermediate levels of biological impairment. Contrasts of the ranked distribution of IBI values for all streams (small and large pooled) shows that the SNARL method has a steeper slope interval or transition, and fewer reference sites intergrading with test sites than the other methods (Figure13).

In order to examine the efficacy of IBI scores in detecting human impairment to water and habitat quality, environmental stress gradients among the sample sites were constructed from surveys of particle size composition, riparian cover, and conductivity. These measures were selected to reflect livestock grazing and channel alteration effects on erosion and sedimentation, bank exposure, vegetation loss, and agricultural irrigation return flows. The IBI scores produced by each method over all streams were then plotted for each stressor (Figures 14, 15 and 16). These graphs revealed apparent thresholds of biological impairment for each stressor, with the SNARL-derived scores exhibiting the

clearest distinction of effect level. The IBI scores were effective in detecting impairment related to the suspected sources of human disturbance.

The inter-calibration of data sets generally indicated the only data standardization needed to produce higher correspondence between the sampling methods was the standardization to equal numbers of identified invertebrates for each stream (Table 3). Specifically, the correlations between IBI scores between any two methods, before or after data standardizations, were always between 0.92 and 0.95 (0.84 and 0.91 R^2). No consistent increase or decrease in these correlations was obtained by any data standardization. By contrast, the between-method similarity improved substantially by standardizing to 500 invertebrates for each stream. The within-method Bray-Curtis distance was 0.32 as estimated by the complete set of permutations of SNARL replicates (recall that a Bray-Curtis distance of 0.00 means no difference between samples while a B-C distance of 1.00 means no shared taxa and thus no similarity in abundance between the two samples; a value of 0.32 thus indicates moderately strong similarity among replicates within a method). The SNARL to R5.USFS.USU between-method B-C distance averaged 0.38 originally, but decreased to 0.33 upon standardizing the SNARL data to 500 invertebrates per stream (Table 4). Thus, while the univariate correlation did not show any substantive change through data standardization, the multivariate similarity analyses indicated that between-method similarity could match that of within-method similarity if the data were standardized to 500 invertebrates per stream. Even these improvements, however, were modest in magnitude relative to the high original correspondence among the different methods (original R^2=0.88 between methods; original 0.38 Bray-Curtis distance between-methods vs. 0.32 within-method).

It is important to note that among these results, the correction for differential sampling by the 250 µm and 500 µm methods was not an important component of the data standardizations. Consistent differences for select taxa were observed (especially smaller taxa such as *Baetis* and various Chironomidae midges), but the most striking bias among the sampling methods was actually the substantially lower density estimates obtained using the CSBP field and lab methods. Across all streams, the CSBP methods typically under-estimated densities for all invertebrates by approximately 50%. Because this was seen only for the CSBP and not the R5.USFS.USU method, we suspect that this

undersampling was unrelated to mesh size. The more likely cause of the lower density estimates were the greater area sampling for each kick (1 ft by 2 ft compared to 1 ft by 1 ft) and both the possibility that this larger area was not sampled as thoroughly and that the greater distance from the net mouth resulted in a greater fraction of organisms dislodged from the substrate passing around rather than into the sampling net. Because the differences for specific taxa between the CSBP and SNARL methods were similar to the differences between the R5.USFS.USU and SNARL methods, there appeared to be no taxon-specific bias from this under-sampling by the CSBP method. Instead, similar fractions of all taxa appeared to be under-sampled using this method.

DISCUSSION

The use of differing methodologies to collect, process, identify and analyze samples of stream macroinvertebrates for bioassessment evaluations of water quality creates potential discrepancies in comparing results and in the assessment conclusions drawn. This study directly addressed how the combined differences between varied methods affect the comparability of results. Using a performance-based methods system to assess precision, bias, discrimination, and accuracy, three dissimilar methods were found to exhibit only small differences in performance, and closely correlated assessment scores, whether derived from multimetric IBIs or multivariate predictive models. The consistent agreement across indicators produced by different bioassessment procedures suggests that output is often directly comparable, data sharing is possible, and that we can have confidence in applying certain alternative techniques to the measurement of biological health in streams.

The need for conformity in bioassessment methods has been identified primarily as enabling data sharing among agencies. Use of uniform methods could permit assessments over broader geographic areas using data combined from different sources, decrease duplication of effort (cost savings), and minimize the potential for conflicting interpretation of results. Another benefit of having a common foundation for evaluating water quality status and trends is that the reporting of ambient conditions over broad regions can be unambiguously understood by the public without any need for adjustment. An alternative view is that data-sharing from programs that together could cover large

geographic areas is not often useful or advisable because stream communities between distant areas share less biogeographic affinity (especially in the western United States) because they come from habitats that may not have common species pools contributing to their assembly. Under these circumstances the differences between streams may have less to do with detecting impairment than natural differences in faunal composition. Furthermore, duplication of effort is probably infrequent (from different management jurisdictions) and agreement among results from different approaches may actually strengthen interpretation, making conclusions more sound through cross-confirmation. Where sharing of data can demonstrably improve bioassessment efforts, it may be sufficient to have a means of calibrating or converting results to the lowest common denominator methodology. It might also be argued that programs that have established a legacy of information through long-term data collection should maintain methodology for the sake of internal consistency rather than expensive re-sampling of existing study sites. As we evaluate the needs for data sharing we must consider not only what could be gained, but what might be lost or not effectively achieved given differing monitoring objectives.

Despite some differences in metric precision and slight bias in taxonomic representation, the PBMS contrasts showed there was broad agreement in site assessment among the methods and similar accuracy in distinguishing reference from test sites. Only in discerning intermediate levels of impaired biological condition were differences more apparent. The sigmoidal response function of ranked IBI scores suggest that as habitat conditions grade from impaired to unimpaired, the threshold for the changes (slope and inflection points) may be more clearly defined by the SNARL method than for CSBP or R5.USFS.USU (Figure 13). In addition, fewer reference sites in the SNARL data set fall within this transitional range of intermediate impairment. As with classic dose-response curves, these properties may permit an improved distinction of lethal from sub-lethal conditions and critical toxicity levels. Plots of the relation of IBI scores to stream degradation did indeed reveal loss of biological integrity over environmental stress gradients at apparent threshold levels (Figures 14, 15, 16). Sedimentation measured by the dominance of small particle sizes (fines, sand, gravel) above 60% of total composition was associated with IBI scores reduced below a reference level of 80 (Figure

14). This dose-response threshold was more clearly defined by SNARL samples than the other methods (fewer sub-threshold points fell below an IBI of 80 and most test sites were distributed above the stress threshold and showed a general graded decrease with dosage). Similarly, thresholds above a conductivity of 200 µSiemens (Figure 15), and below a riparian vegetation cover of 30% (Figure 16), were most obviously exhibited by the SNARL data set. Distinct sigmoid data distributions provide clearer patterns for identifying ecotoxicity levels for environmental stressors, and for establishing management practices and biological criteria targets for recovery. The designation of aquatic life use attainment categories also depends on being able to define levels of impairment. The terms "not supporting" and "partially supporting" might be defined according to whether a test site falls into the lowest or transitional portions of the distribution (respectively), providing a means of prioritizing control and cleanup of pollution sources as directed by the Clean Water Act.

The greater laboratory effort required for the 5-replicate SNARL samples, even with the benefit of improved resolution of intermediate levels of impairment, may not be practical in many cases, but existing data may still be used to supplement samples collected using other methods. This potential to interchange data sets is highlighted by the results from the comparisons among standardized data sets. Univariate IBI correlations were as high initially with raw assessment scores as they were following a broad suite of data standardizations and conversions. Although the multivariate characterization of the community became more consistent following these data conversions (particularly standardization to equal numbers of invertebrates processed in the lab), the initial high correlations among IBI scores for the three methods without standardization and the lack of improvement of these correlations indicate that the methods produce generally consistent results and that the data collected by one method can be used in data sets composed primarily of data derived from different methods. The only substantive data conversion that should be considered is standardization to equal numbers of invertebrates for all samples. Although not critical for the univariate methods evaluated here, all three methods considered had high counts (typically 500-1500 invertebrates per stream) and the differences may have attenuated at this count along the species accumulation curves (Vinson and Hawkins 1996). In general, standardizing data

sets to equal numbers of individuals makes intuitive sense, can lead to improved correspondence among assessment methods, eliminates the possibility of variable counts influencing the final results, and is relatively easy to implement using various statistical and database software.

Though the use of RIVPACS predictive models typically involves large data sets (hundreds of sites, e.g. Hawkins et al. 2000), the intent of the contrast of methods presented here was to evaluate relative effectiveness given the same data set size. An added cost of the RIVPACS approach may then be that effective assessment using the recommended data requirements might only be attained with more extensive sampling and laboratory effort. Other sample effort issues that could be considered are any changes in method performance that might accompany (1) reducing the SNARL replication from 5 to 3, (2) using genus-level taxonomic resolution of midges and mites for CSBP, (3) pooling of all samples in replicated methods (SNARL and CSBP) and comparing fixed and equal counts of 500 (count for R5.USFS.USU), (4) replicating sampling from composites collected over the entire reach rather than composites within the same riffle in each reach.

Misclassifications of two sites initially assigned to reference and test groups for large streams were indicated by accordance of all three methods (see low and high IBIs for Owens.abovetun and WWalker.Pickel, respectively, in Figure 7). The Owens River site was used as a local upstream control to evaluate conditions above and below inflow from a diversion tunnel but otherwise is a channel exposed to flow modification and bank disturbance associated with a long history of livestock grazing; and the West Walker River site was in a livestock grazing allotment but showed stable channel habitat. This rationale argues that the bioassessment data was not in error, but initial site classification was wrong, so the type I/II error rates (Figures 10 and 11) were overestimated.

The composite IBI scores and predictive model O/E values yield comparable assessments over all streams examined despite their differences in computation. Both multimetric and multivariate approaches to contrasting test sites with reference sites use procedures, however, that are not consistent from one data set to another. Multimetric calculation of a single index of biotic integrity involves selection, standardization and summation of those metrics that produce the best separation of reference from test sites or

best correlations with stressor gradients. This means that the number and type of metrics used to compute the index 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). Multivariate approaches also involve somewhat subjective decisions for identifying reference groups that depend on the clustering techniques and measures of community similarity used, and also employ differing combinations of environmental variables and weightings to assign a test site to some probability of belonging to a reference group. These predictor variables and their coefficients change from one data set to another such that test sites are evaluated only in the context of a circumscribed group of reference sites. Notwithstanding this lack of uniformity and other potential biases and limitations of both multimetric indices and multivariate predictive models (reviewed by Karr and Chu 2000, Norris and Hawkins 2000, Suter 1993), the results presented here suggest that similar assessments of impairment are obtained using either of these analytical tools for data sets derived from differing field and laboratory bioassessment methods.

CONCLUSIONS

This study showed broad equivalence among bioassessment methods in application to distinguishing impaired from unimpaired biotic conditions in streams. Using different field and laboratory methods and analytical tools, the complementary results 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 biological monitoring of water quality. Although conclusions regarding the separation of reference from test sites were seldom in conflict, some differences between methods were apparent with regard to distinguishing moderate levels of degradation and environmental stress thresholds. For the purpose of choosing methodologies in California biomonitoring programs, following are some considerations and options for proceeding:

- Monitoring of research studies with long-term data sets and project-specific sampling designs, where changes in the composition of localized biological communities are being followed, require consistent methodologies for

comparing trends and taxon-specific or localized changes. In these cases it is important to retain pre-existing methods so that data are strictly comparable.

- For ambient monitoring and biocriteria development:
    1. Continue using existing methods since assessments will usually be in agreement (high correlations of IBIs and O/Es suggest data may be shared directly or converted if necessary)
    2. Use the most cost-effective method since the results showed equal outcomes in impaired/unimpaired assessment conclusions (R5.USFS.USU is lowest cost method)
    3. Adopt one uniform method with the best potential for data-sharing in biocriteria development (CSBP has the most extensive statewide database, but R5.USFS.USU contains data from most western states)
    4. Use the method with the most precision and best potential for distinguishing moderate levels of stress response to impairment (UC-SNARL has the best performance characteristics in this regard)
    5. Include an alternative method to improve upon impairment resolution, taxonomic bias, replication if needed, and multihabitat representation

Though the methods compared had substantial differences in protocols, they were equivalent in accuracy of discriminating pre-defined reference from test sites. These results therefore support the conclusion that data from multiple sources can be used in broad, integrated analyses of stream conditions, and can be converted to a uniform data structure provided simple standardizations are employed to enhance data comparability.

REFERENCES

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. Second edition. U.S. Environmental Protection Agency; Office of Water; Washington, D.C. EPA 481-B-99-002.

Barbour, M.T. and C. Hill. 2003. The status and future of biological assessment for California streams. Report to the California State Water Resources Control Board, Division of Water Quality.

Cao, Y., C.P. Hawkins, and A.D. Storey. 2004. A method for measuring the comparability of different sampling methods used in biological surveys: implications for data integration and synthesis. Manuscript submitted for publication.

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.

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.

Faith, D.P., P.R. Minchin, and L. Belbin. 1987. Compositional dissimilarity as a robust measure of ecological distance. Vegetatio 69: 57-68.

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. U.S. Geological Survey, Open File Report 94-490.

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.

Karr, J.R. 1998. Rivers as sentinels: using the biology of rivers to guide landscape management. In Naiman, R.J. and R.E. Bilby (editors), River Ecology and Management: Lessons from the Pacific Coastal Ecoregion. Springer, New York. Pages 502-528.

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.

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 (3): 664-681.

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.

Ode, P.R., A.C. Rehn, and J.T. May. 2004. A quantitative tool for assessing the integrity of southern coastal California streams. Environmental Monitoring and Management, In Press.

Resh, V.H. and E.P. McElravy. 1993. Contemporary quantitative approaches to biomonitoring using benthic macroinvertebrates. In D.M. Rosenberg and V.H. Resh (editors), Freshwater Biomonitoring and Benthic Macroinvertebrates. Chapman and Hall, New York. Pages 159-194.

Resh, V. H. and J.K. Jackson. 1993. Rapid assessment approaches to biomonitoringusing benthic macroinvertebrates. In D.M. Rosenberg and V.H. Resh (editors), Freshwater

Biomonitoring and Benthic Macroinvertebrates. Chapman and Hall, New York. Pages 195-233.

Reynoldson, T.B., R.C. Bailey, K.E. Day, and R.H. Norris. 1995. Biological guidelines for freshwater sediment based on BEnthic Assessment of SedimenT (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., II. 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.

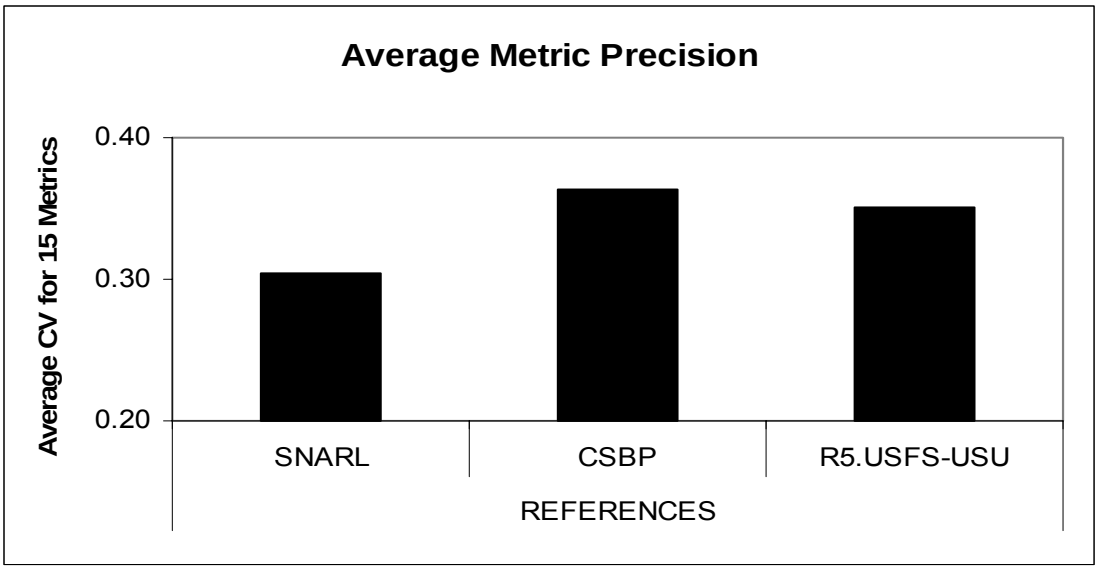


Figure 1. Precision contrast among methods, standardized as the coefficient of variation (CV) for the 15 metrics used to prepare the IBI from reference streams sampled.

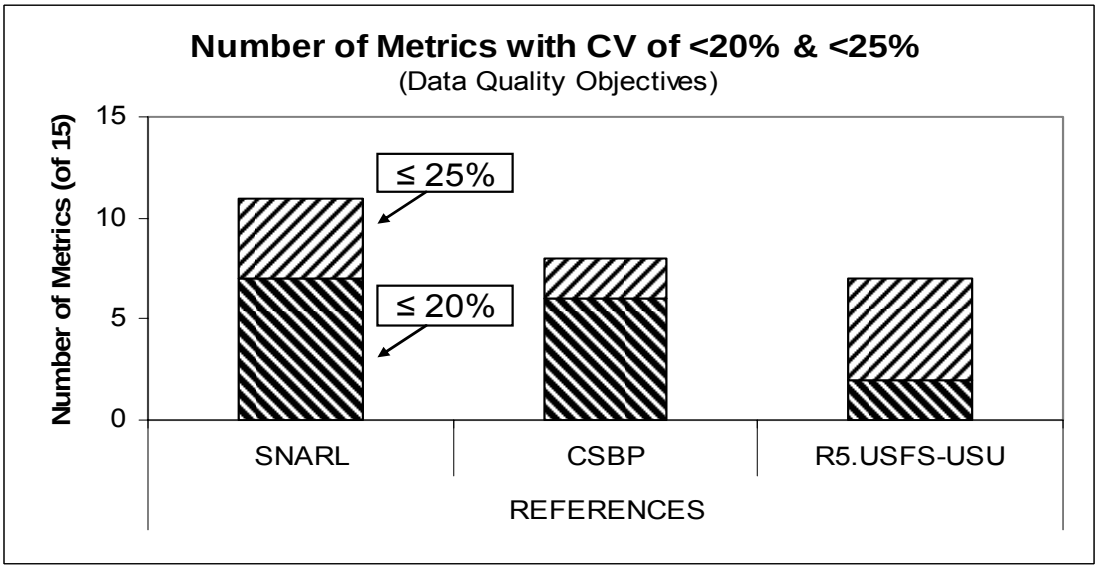


Figure 2. Number of metrics achieving data quality objectives of variability of less than 20% and 25% CV values for the 15 metrics used to prepare the IBI from reference streams sampled.

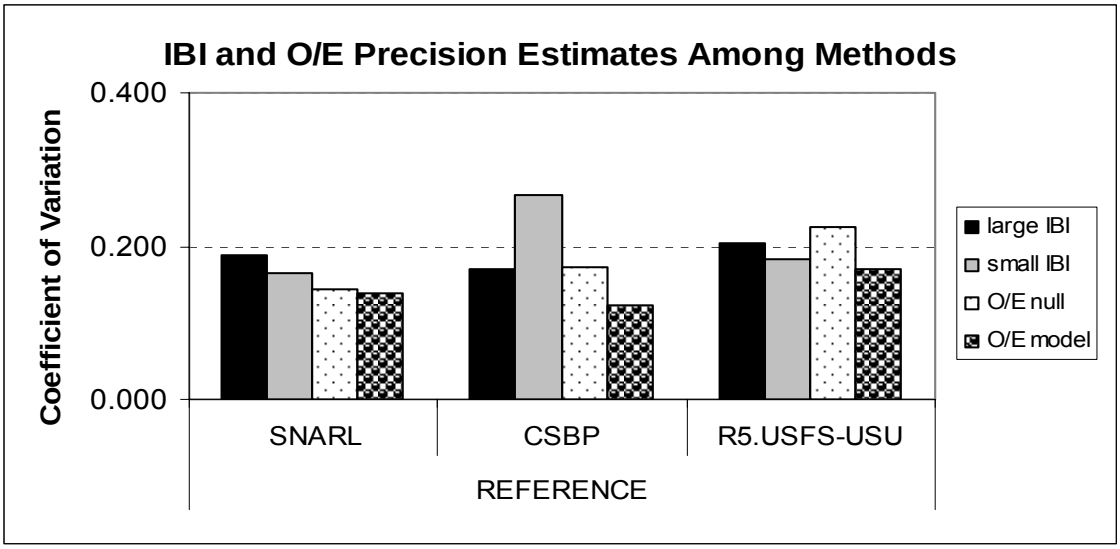


Figure 3. Precision contrast among methods for overall index indicators of integrity and impairment: IBIs for large and small stream types (multimetric) and O/E (observed over expected) values (multivariate or predictive models). O/E null refers to simple calculation of this community similarity comparison without predictive classification of reference sites, and O/E model refers to the full model where the expected values for test sites are based on the weighted prediction from reference site clusters. Note that all methods and models show reference condition near or below DQO of 20% CV.

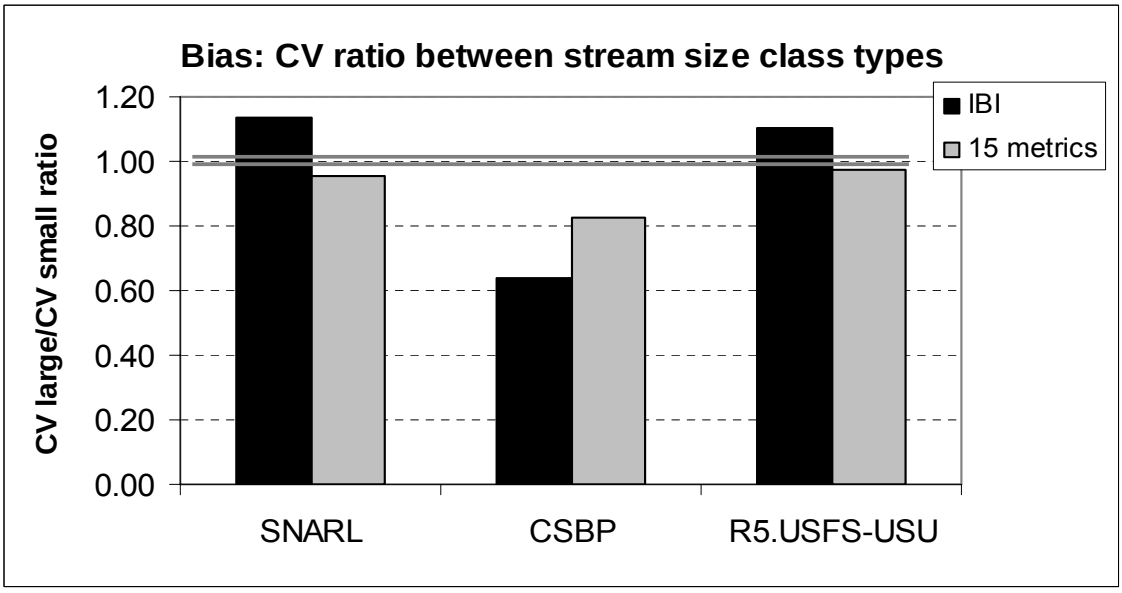


Figure 4. Ratio of CV for large vs small stream types (reference sites only) as an indicator of bias among methods. Deviation from ratio near 1.0 shows bias in applicability of metrics or index to different stream types sampled here (or regions, stream class categories).

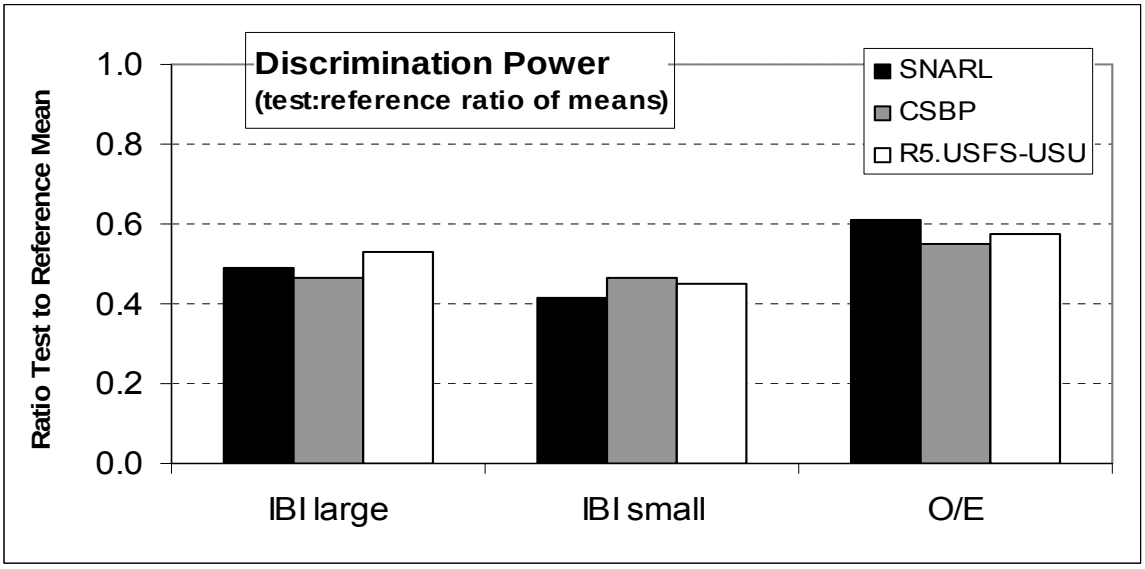


Figure 5. Potential for discrimination of test sites from reference condition is indicated by reduction in the ratio of test mean from reference mean, shown here for multimetric IBI for large and small stream groups, and for the multivariate O/E indicator. As the ratio approaches 1.0 there is less and less power for the indicator to discriminate impairment.

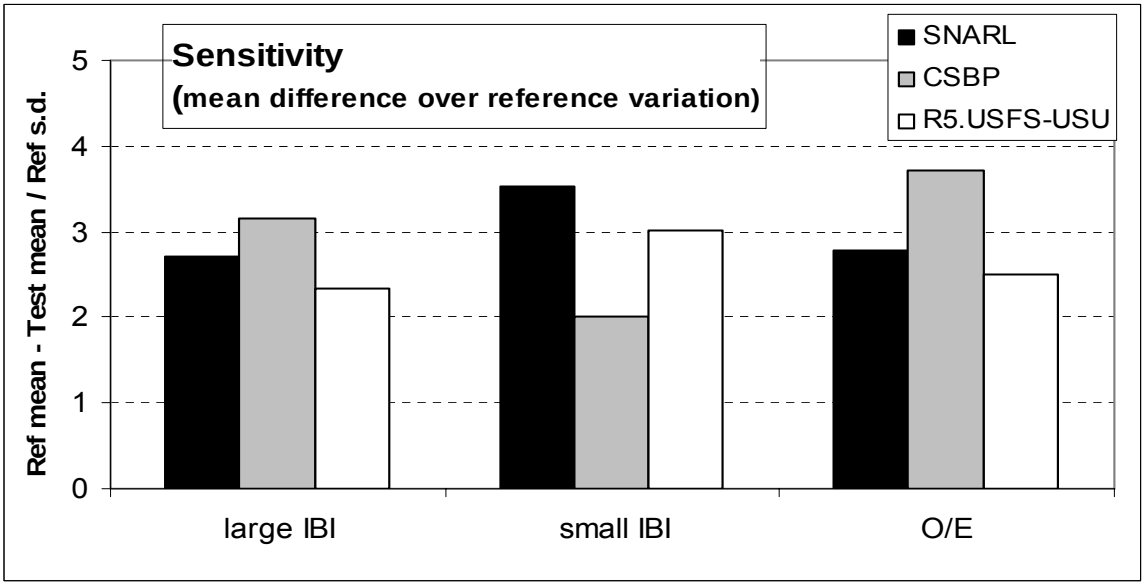


Figure 6. Sensitivity in distinguishing test sites on average relative to the reference mean and its variation in different cases of index development. Used as a t-statistic, all methods and cases would yield significant values for maintaining the probability of making a type I error at less than 0.05.

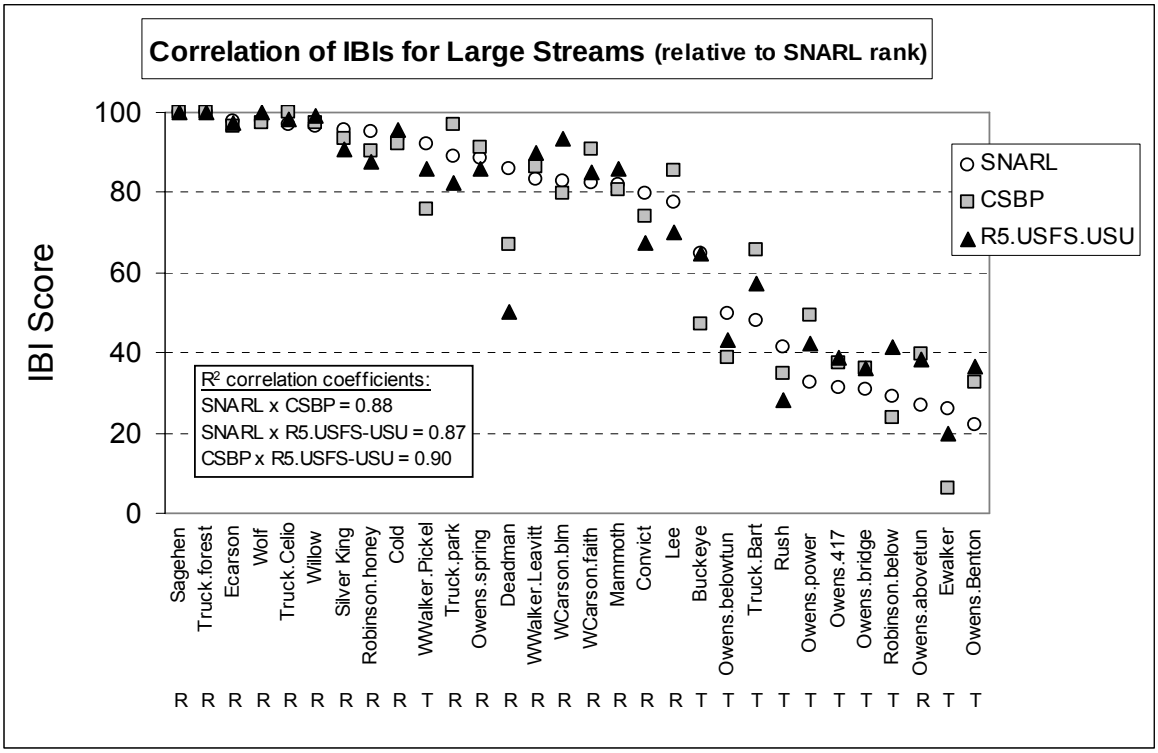


Figure 7. Co-plots of IBI scores derived from each method for large streams relative to the ranking of SNARL scores (reference = R, test = T).

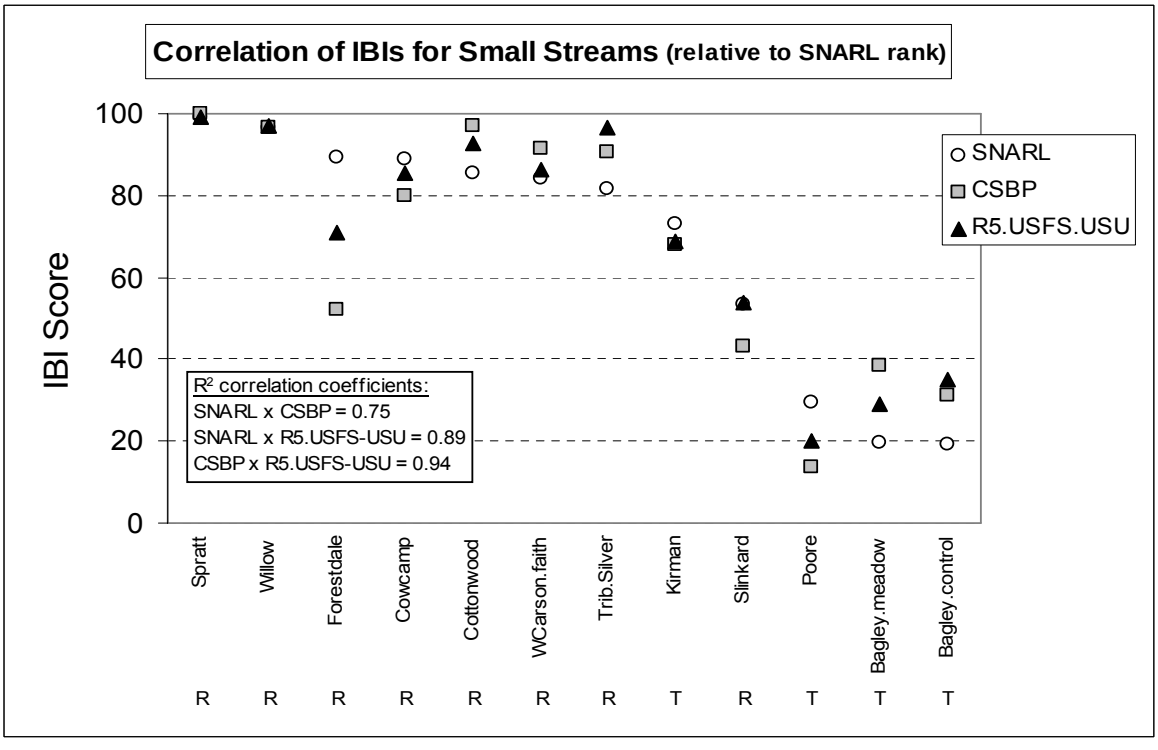


Figure 8. Co-plots of IBI scores derived from each method for small streams relative to the ranking of SNARL scores (reference = R, test = T).

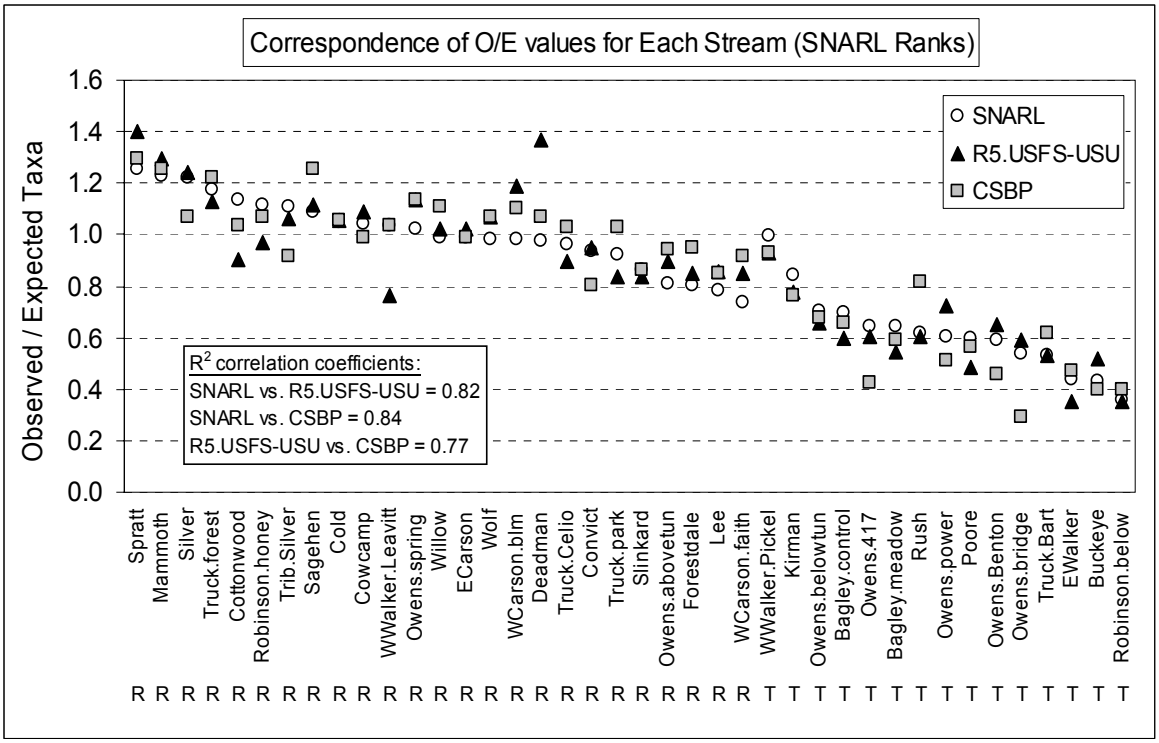
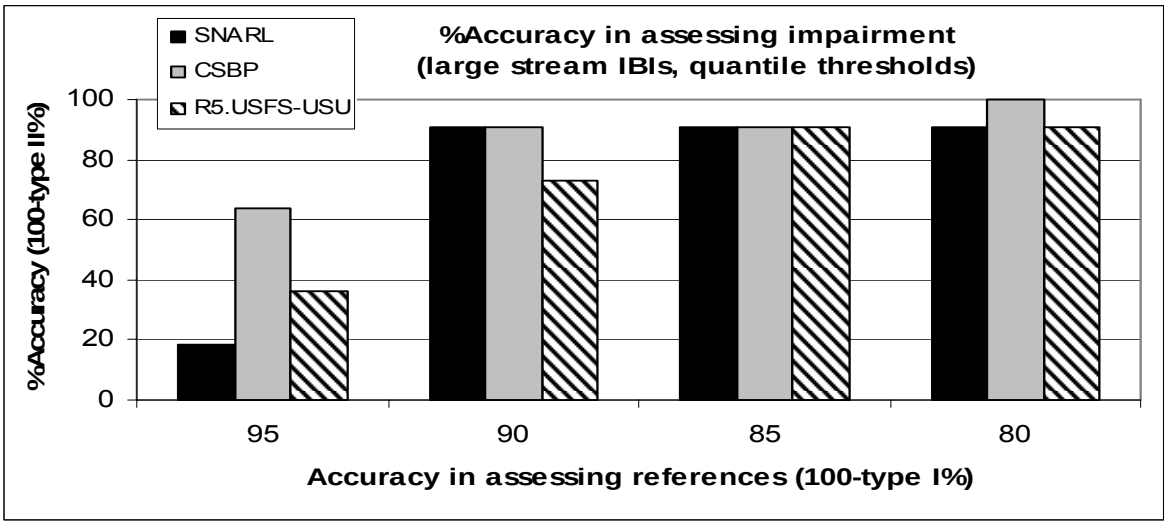
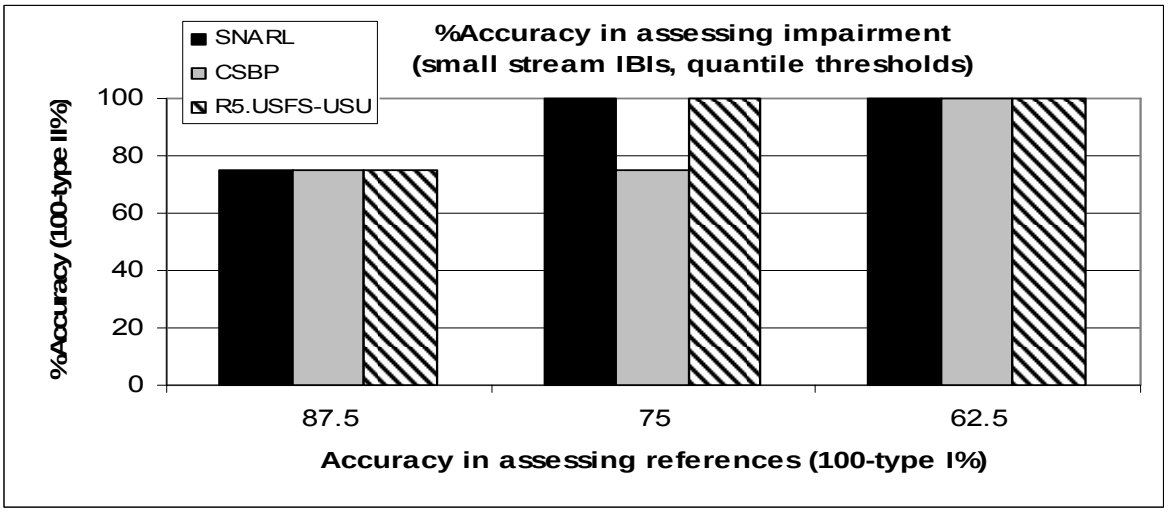


Figure 9. Co-plots of O/E values derived from each method for all streams relative to the ranking of SNARL scores (reference = R, test = T).





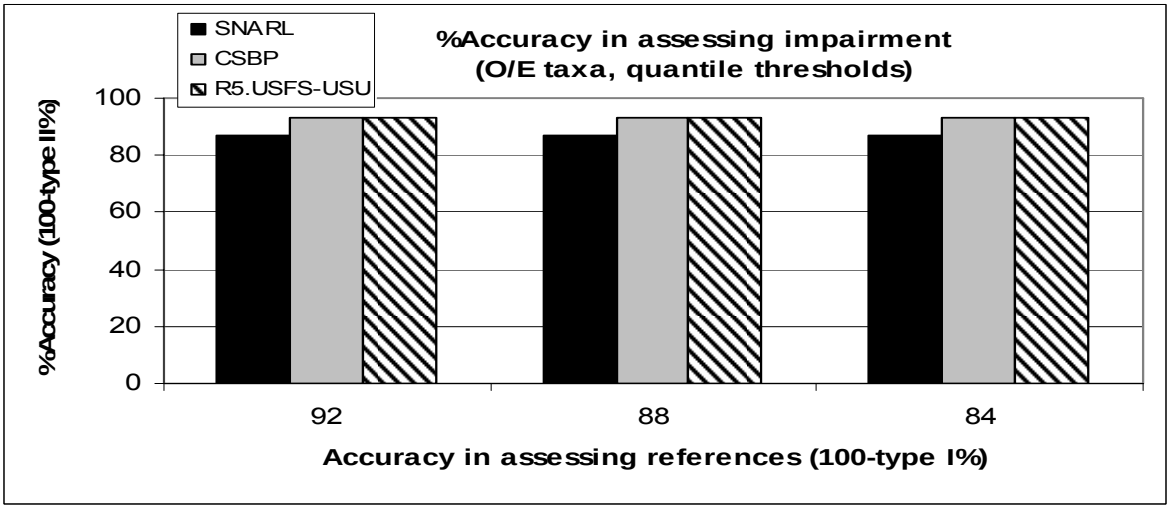
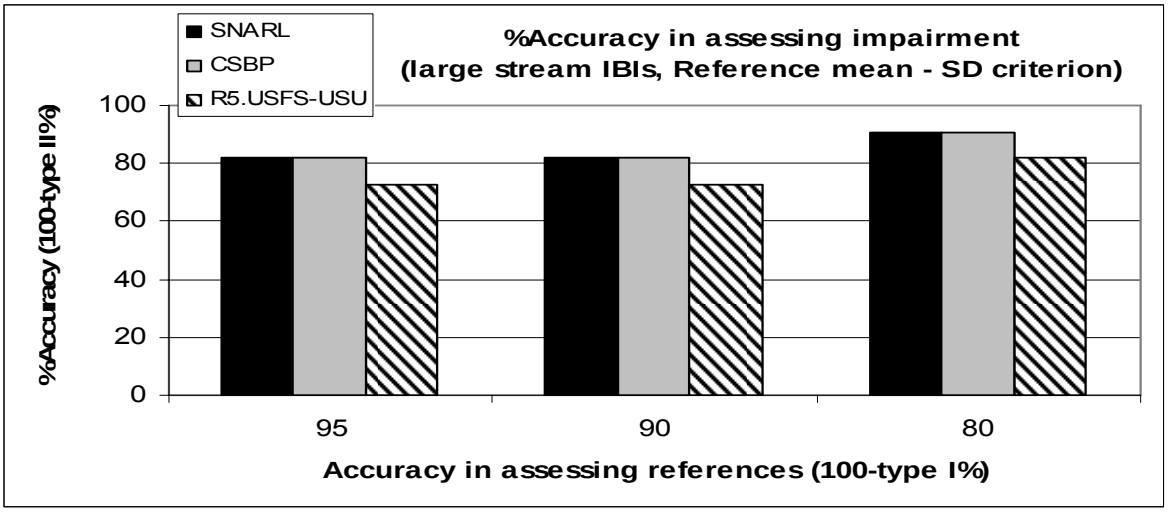
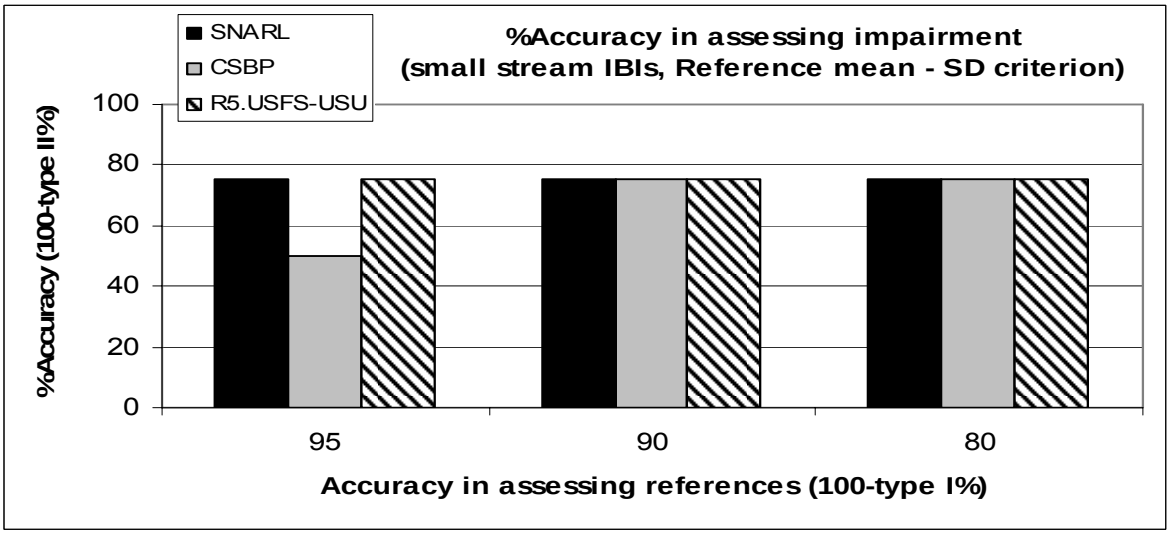


Figure 10. Quantile thresholds for assessing the accuracy of identifying test site impairment for different methods and different measures of biological integrity.





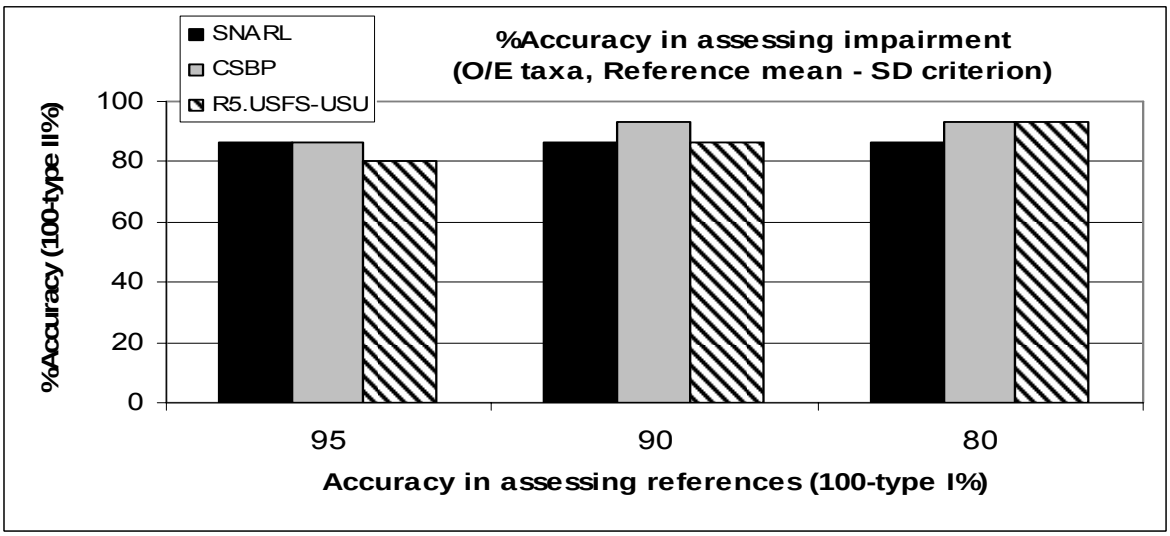


Figure 11. Standard deviation thresholds for assessing the accuracy of identifying test site impairment for different methods and different measures of biological integrity.

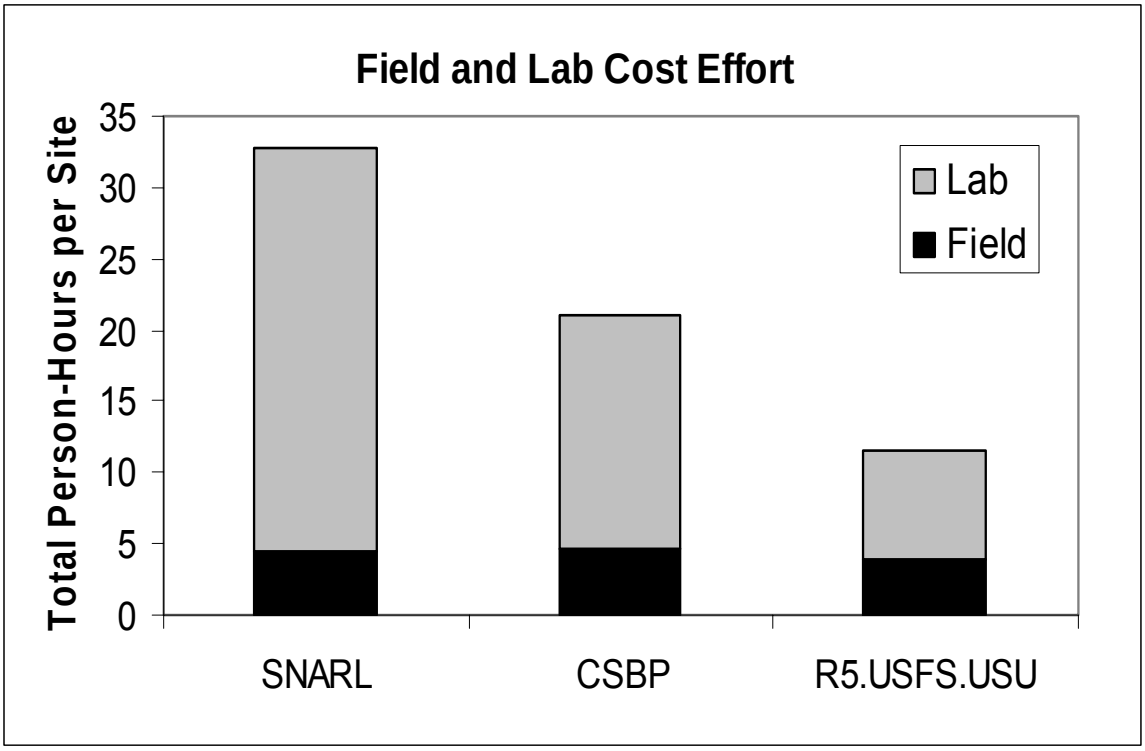


Figure 12. Total person-hours of effort spent in completing field and laboratory tasks for a single site or reach bioassessment survey (sample collection, habitat survey, sample processing, sorting, identifications and counts) for each of three methods.

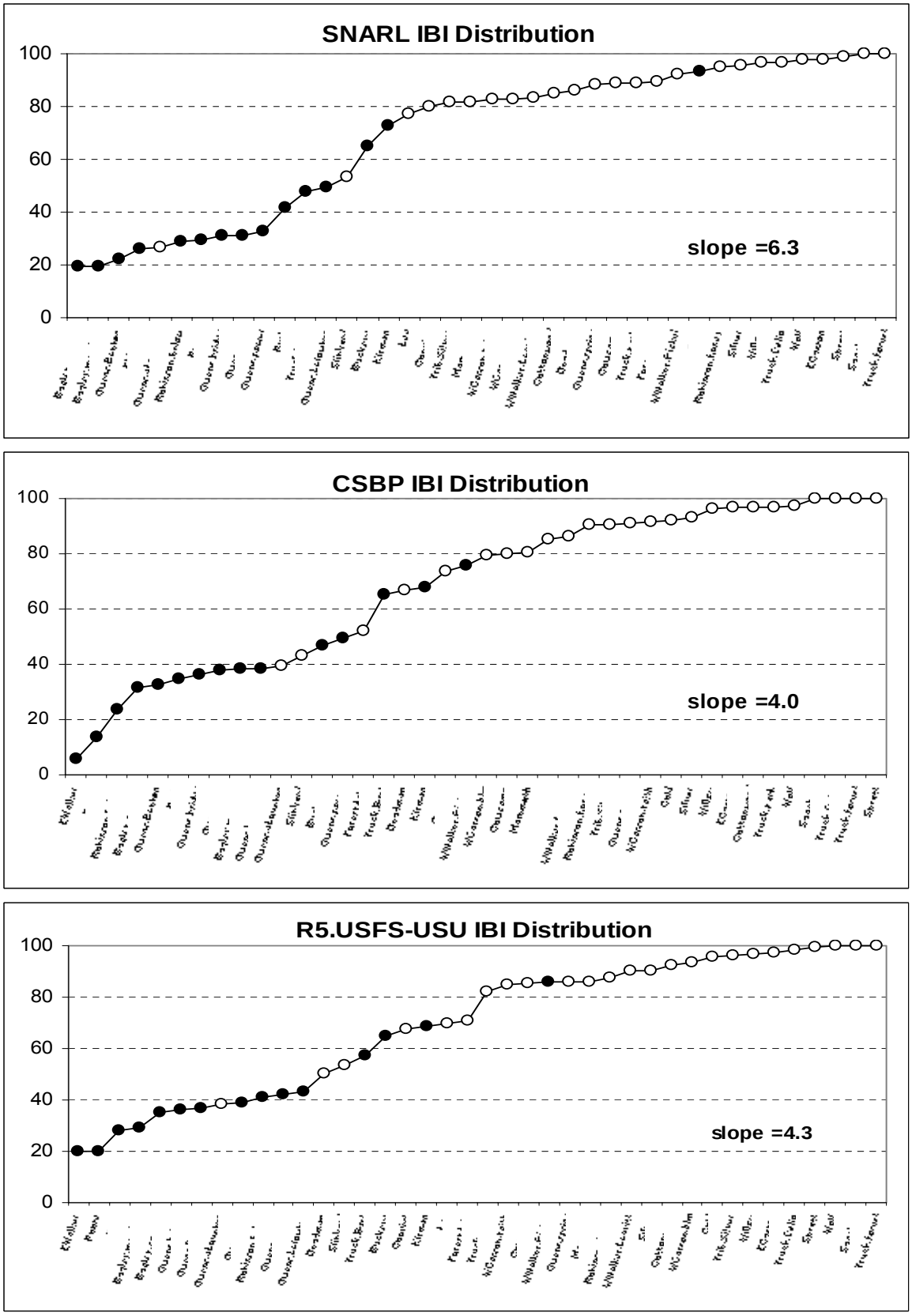


Figure 13. Ranked IBI scores for reference (open) and test (black) sites for each method.

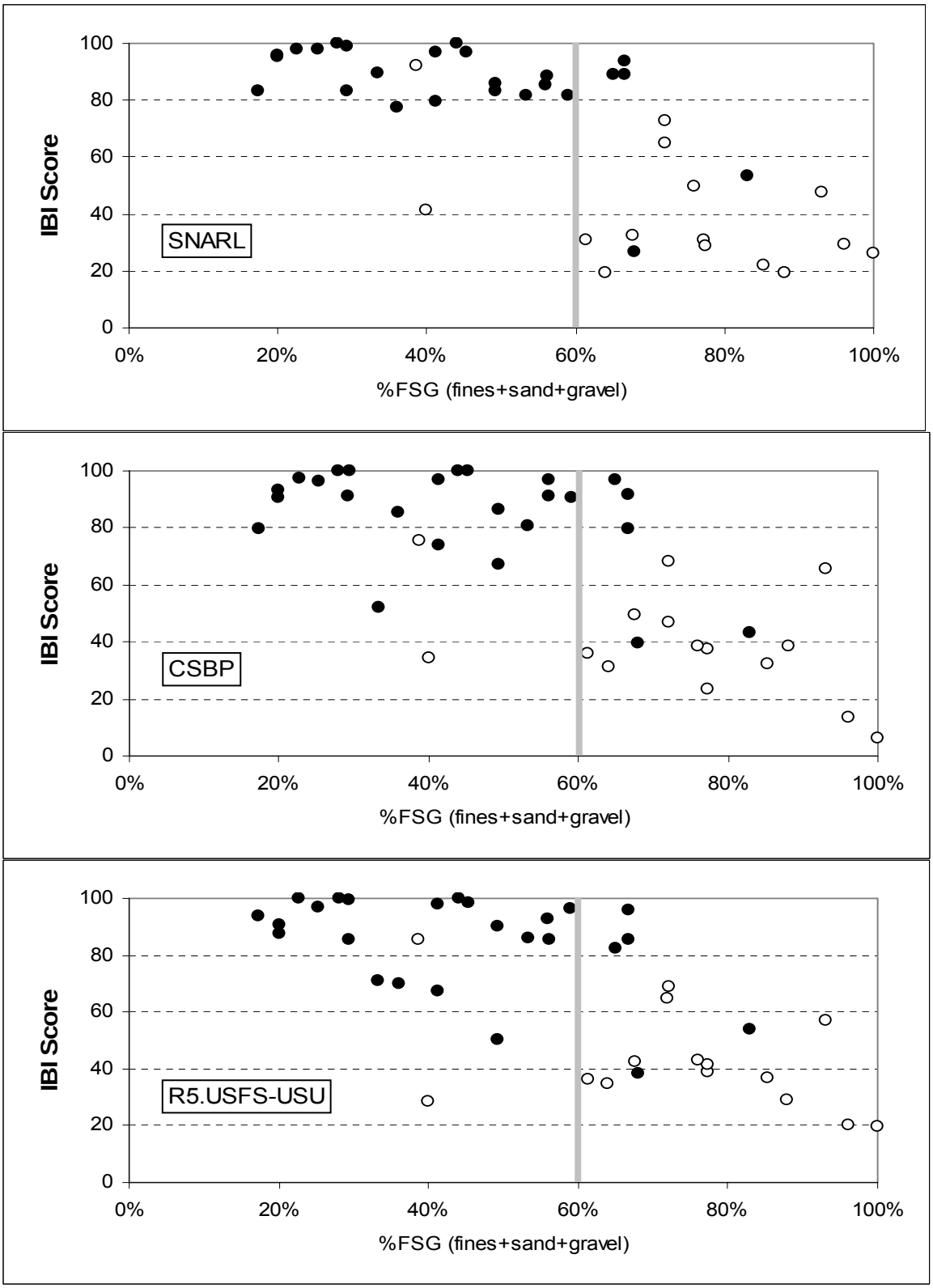


Figure 14. Threshold values for loss of biological integrity associated with substrate degradation (by small particle sizes = % FSG, fines + sand + gravel) for different bioassessment methods. Filled symbols = reference sites, open symbols = test sites.

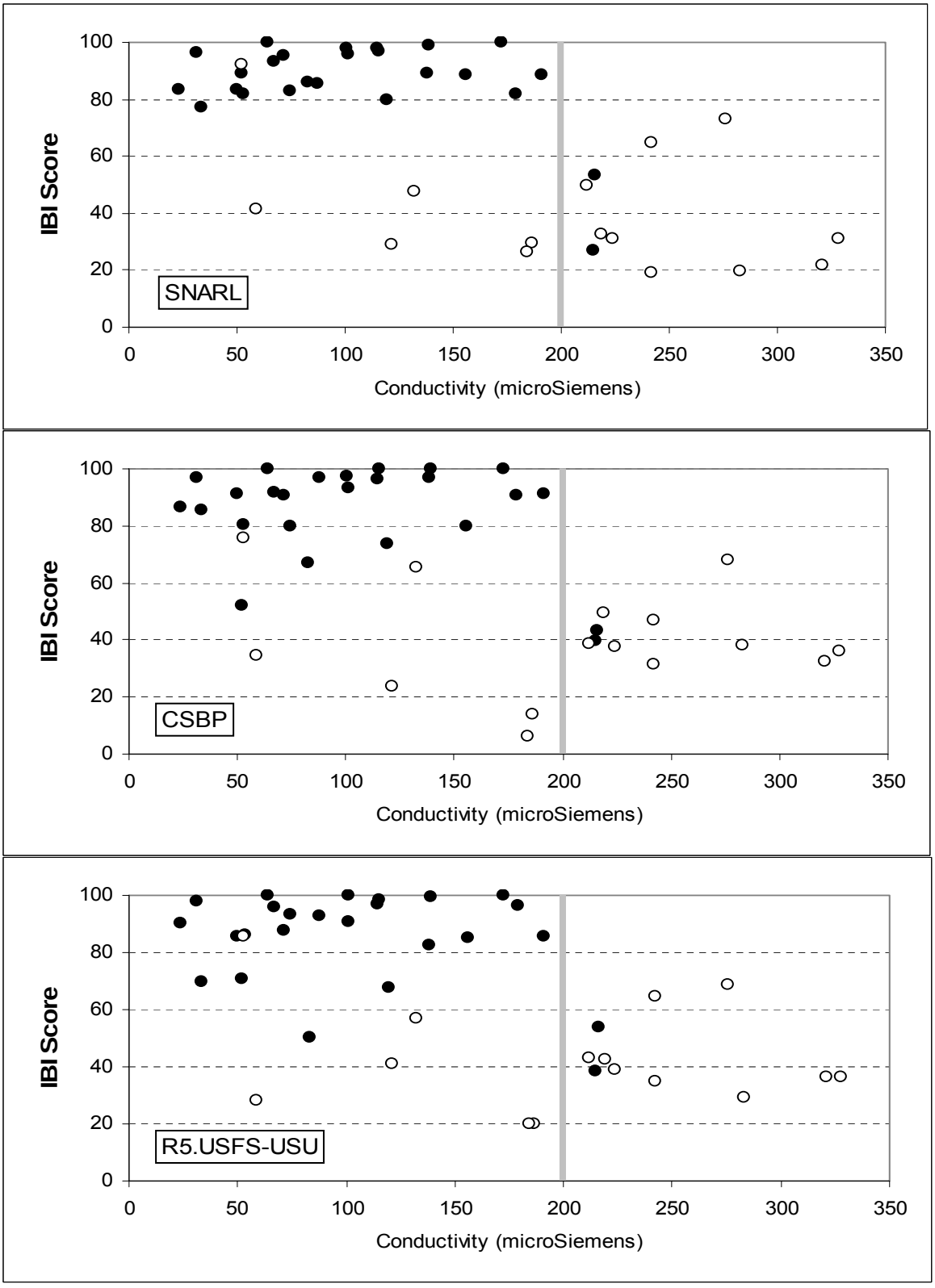


Figure 15. Threshold values for loss of biological integrity associated with conductivity (low flows, agricultural return flows) for different bioassessment methods. Filled symbols = reference sites, open symbols = test sites.

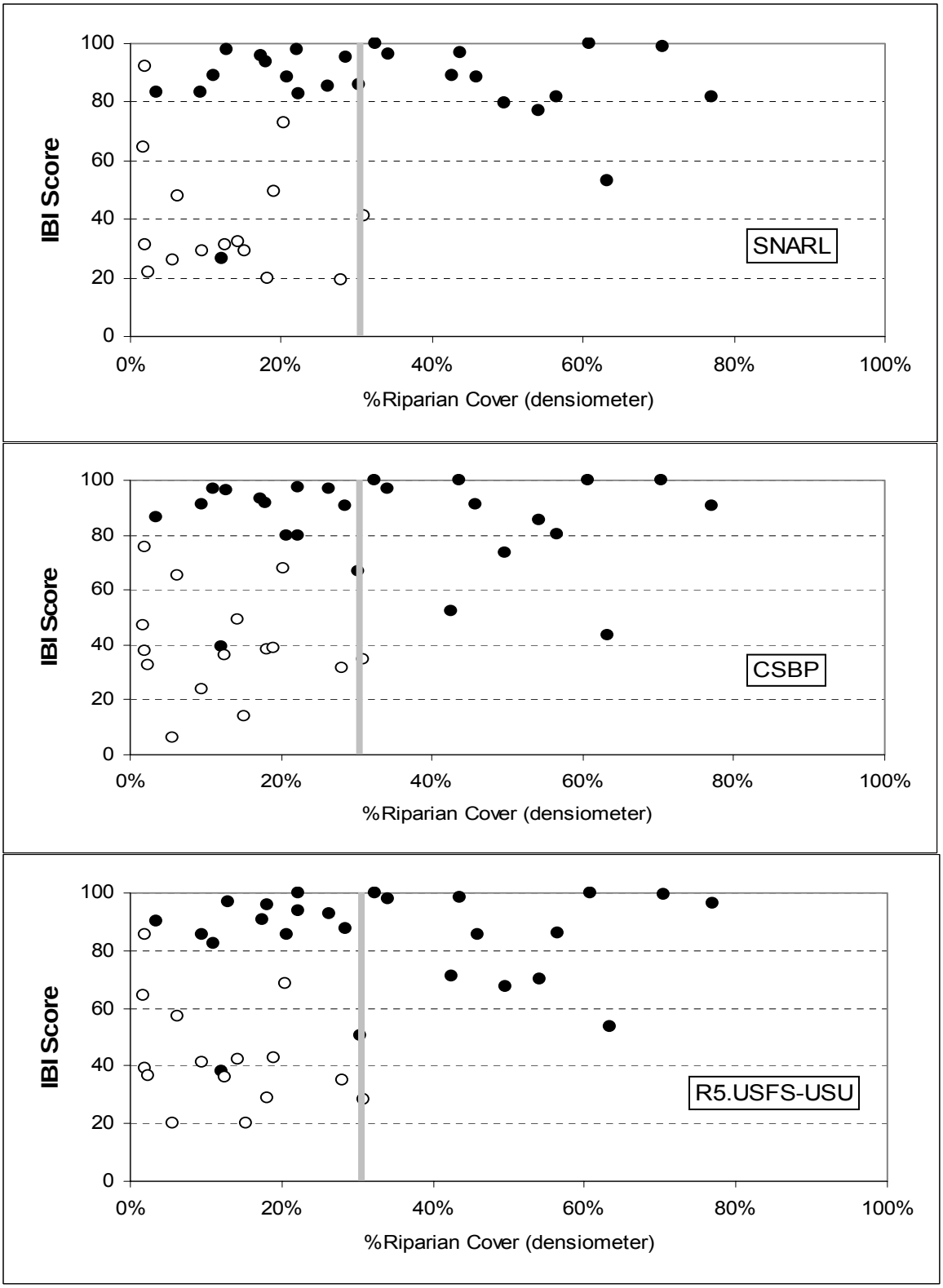


Figure 16. Threshold values for loss of biological integrity associated with riparian cover (reduced canopy vegetation) for different bioassessment methods. Filled symbols = reference sites, open symbols = test sites.

Table 1. Stream classification and reference site selection criteria.

Table 1. Stream Size and Degradation Classification:				Reference = <0.2 xings /km	Or <25% erosion	Local or upstream habitat degradation (grazing, altered channel structure)		
Large Stream Class	Mean Width	Upstream Length (km)	Upstream Area (km^2)	Road Xings per upstream km	Local bank erosion %	minimal or absent	severe or extensive	R or T
Upper Truckee - Forest	737	11.3	32.5	0.000	0.0%	☑		R
Willow Creek - lower	307	10.4	28.1	0.000	3.3%	☑		R
East Carson – above Bagley	1484	37.3	237.7	0.000	3.3%	☑		R
Silver King Creek – above valley	711	22.2	100.1	0.000	10.0%	☑		R
West Walker – upper Leavitt	1253	24.6	120.4	0.000	40.0%	☑		R
Convict Creek – lower SNARL	415	16.6	61.7	0.043	0.0%	☑		R
Wolf Creek – above trailhead	636	12.8	40	0.076	20.0%	☑		R
West Walker - middle Pickel	1464	27.84	145.9	0.102	33.3%		☑	T
Robinson Creek – Honeymoon flat	817	23.1	106.9	0.112	26.7%	☑		R
Buckeye Creek – below WRID	422	30.3	168	0.122	76.7%		☑	T
Sagehen Creek – below field station	382	6	11.3	0.123	3.3%	☑		R
Robinson Creek – below WRID	672	34.8	211.7	0.134	63.3%		☑	T
Lee Vining Creek – moraine campground	951	12.6	38.9	0.145	10.0%	☑		R
Rush Creek - bottomlands	963	30.3	168	0.170	26.7%		☑	T
Deadman Creek - above Big Springs cg.	489	17.3	66.1	0.174	13.3%	☑		R
Upper Owens - below Mono Tunnel	1008	23.9	113.2	0.188	0.0%		☑	T
Upper Owens - above Mono Tunnel	644	23.3	108.5	0.189	0.0%	☑		R
Upper Owens - below Big Springs cg.	753	19.2	78.6	0.191	0.0%	☑		R
West Carson – upper Faith	479	4.3	6.5	0.195	3.3%	☑		R
East Walker – WRID	919	24.6	120.4	0.221	90.0%		☑	T
Upper Owens – Ebasco 417s	964	27.8	145.6	0.225	26.7%		☑	T
Upper Owens – Ebasco Powerline	994	32.4	188	0.235	3.3%		☑	T
Upper Truckee - Celio lower	736	12.8	40	0.280	6.7%	☑		R
West Carson - lower/BLM	1255	33.4	197.7	0.312	6.7%	☑		R
Upper Truckee - state park	921	13.7	44.8	0.315	7.0%	☑		R
Upper Truckee - Barton lower	885	21.8	97.1	0.327	33.0%		☑	T
Upper Owens - above Bridge	1556	42.2	292	0.389	16.7%		☑	T
Upper Owens - below Benton xing	1132	44.2	315.4	0.395	33.3%		☑	T
Mammoth Creek - substation	660	17	64.2	0.560	10.0%	☑		R
Cold Stream Creek – upper gravel pit	523	6.9	14.3	0.565	20.0%	☑		R
Small Stream Class								
Trib.1 Silver King – above SKC	75	2	1.8	0.000	0.0%	☑		R
Forestdale Creek – upper	318	1.98	1.8	0.000	3.3%	☑		R
Willow Creek - lower	307	10.4	28.1	0.000	3.3%	☑		R
Spratt Creek – above rd xing	174	7.2	15.4	0.132	10.0%	☑		R
West Carson – upper Faith	479	4.3	6.4	0.195	3.3%	☑		R
Kirman Creek - lower	96	2.75	3.1	0.232	10.0%		☑	T
Cottonwood Creek - Sweetwater meadow	153	8	18.3	0.269	0.0%	☑		R
Cowcamp Creek – lower Schoettler	114	3.37	4.3	0.286	10.0%	☑		R
Slinkard Creek - restoration area	66	8	18.1	0.365	0.0%	☑		R
Bagley Valley Creek - meadow	133	2	1.9	0.629	10.0%		☑	T
Bagley Valley Creek - lower	136	2.7	3	0.862	10.0%		☑	T
Poore Creek - 1/3 grazed	207	4.36	6.64	0.890	3.3%		☑	T

Upstream area calculation: estimated from length of single longest channel length from site location to headwater as $L = 1.4\ A^{0.6}$ (square kilometers)

Reference site selection: initial screen as sites with upstream road density less than 0.2 crossings per km channel, OR local bank erosion <25% cover AND no known degradation source exists

Table 2. Summary of Differences Between Methods

PROTOCOL:	UC-SNARL Lahontan	CSBP Dept. Fish and Game	RIVPACS Forest Service
Net type and mesh	D-frame, 250 µm	D-frame, 500 µm	D-frame, 500 µm
Replication	5 composites of 3	3 composites of 3	1 composites of 8
Area sampled	1.39 m^2 (1'x1')	1.67 m^2 (1'x2')	0.74 m^2 (1'x1')
Subsampling	Drum splitter	Grid Tray	Grid Tray
Enumeration	250-500 count	300 fixed count	500 fixed count
Taxonomic Resolution	Genus/species (including midges and mites) plus large and rare	Genus/species (midges and mites to subfamily/family) plus large and rare	Genus/species (including midges and mites) plus large and rare

Similarities: riffle habitat, physical habitat surveys, water chemistry, QA/QC to <5% sort error (20% checked) and 100% IDs checked, multimetric and multivariate analyses used for each

Table 3. Correlations among final IBI scores for the different field/lab methods before and after data standardizations. Correlations are calculated from all streams analyzed simultaneously. For additional details on the data standardizations, see Methods section.

	Correlation between IBI Scores (r)	R^2 (% of Variation Explained)
Original Correlations		
SNARL vs. RIVPACS	0.94	88%
SNARL vs. CSBP	0.92	84%
RIVPACS vs. CSBP	0.95	91%
Post-Standardization Correlations		
(1) SNARL vs. RIVPACS - SNARL raw counts pooled and 500 individuals re-sampled without replacement for each stream	0.93 [a]	87% [a]
(2) SNARL vs. RIVPACS - same as (1) above plus large/rare individuals added to the re-sampled data	0.93	86%
(3) SNARL vs. RIVPACS - same as (1) above but with SNARL data corrected for apparent over-sampling from using 250 µm mesh	0.92	85%
(4) SNARL vs. RIVPACS - each SNARL replicate re-sampled for same number of individuals to yield ~500 individuals per stream	0.93 [a]	87% [a]
(5) SNARL vs. RIVPACS - each SNARL replicate re-sampled based on fraction of sample originally processed	0.94	88%

[a] - *These are the mean value for 3 independent randomized re-samplings of the data*

Table 4. Similarity Comparisons within and between field/lab methods. Calibration alternatives for converting SNARL method to 500 fixid-count R5.USFS-USU data form. Similarities are calculated as Bray-Curtis Distance (0 value indicating identical sample) using proportional data.

	Original Comparisons		Standardized SNARL Data [a]			
			(1) 500 resampling	(2) 500 & Large/Rare	(3) 500 & Bias Correction	(4) Same # from each Replicate
Stream	**Within-Method [a]**	SNARL vs. RIVPACS	SNARL vs. RIVPACS	SNARL vs. RIVPACS	SNARL vs. RIVPACS	SNARL vs. RIVPACS
Bagley.control	0.52	0.49	0.31	0.32	0.31	0.34
Bagley.meadow	0.34	0.29	0.18	0.20	0.19	0.20
Buckeye	0.25	0.55	0.52	0.53	0.52	0.53
Cold	0.31	0.37	0.36	0.34	0.34	0.35
Convict	0.31	0.37	0.33	0.33	0.31	0.35
Cottonwood	0.35	0.37	0.32	0.35	0.32	0.34
Cowcamp	0.21	0.37	0.36	0.35	0.35	0.37
Deadman	0.28	0.44	0.43	0.43	0.41	0.42
ECarson	0.31	0.36	0.31	0.31	0.29	0.31
EWalker	0.34	0.31	0.23	0.26	0.22	0.26
Forestdale	0.31	0.30	0.26	0.23	0.24	0.24
Kirman	0.36	0.33	0.25	0.24	0.25	0.25
Lee	0.25	0.27	0.23	0.25	0.24	0.24
Mammoth	0.31	0.31	0.25	0.29	0.30	0.27
Owens.417	0.23	0.30	0.27	0.28	0.27	0.28
Owens.abovetun	0.19	0.18	0.35	0.17	0.39	0.16
Owens.belowtun	0.30	0.38	0.16	0.36	0.15	0.35
Owens.Benton	0.37	0.41	0.32	0.40	0.35	0.39
Owens.bridge	0.43	0.36	0.19	0.27	0.26	0.25
Owens.power	0.31	0.40	0.39	0.36	0.35	0.37
Owens.spring	0.26	0.34	0.30	0.31	0.32	0.30
Poore	0.27	0.41	0.40	0.38	0.38	0.39
Robinson.below	0.35	0.32	0.23	0.27	0.25	0.24
Robinson.honey	0.38	0.32	0.23	0.23	0.20	0.22
Rush	0.21	0.42	0.41	0.42	0.40	0.43
Sagehen	0.34	0.43	0.39	0.38	0.36	0.38
Silver	0.34	0.39	0.36	0.29	0.30	0.32
Slinkard	0.22	0.50	0.49	0.49	0.49	0.49
Spratt	0.42	0.35	0.24	0.27	0.26	0.25
Trib.Silver	0.40	0.50	0.45	0.46	0.45	0.47
Truck.Bart	0.29	0.43	0.37	0.40	0.39	0.41
Truck.Celio	0.30	0.38	0.35	0.36	0.36	0.35
Truck.forest	0.35	0.60	0.58	0.57	0.58	0.58
Truck.park	0.26	0.39	0.36	0.37	0.36	0.36
WCarson.blm	0.38	0.42	0.41	0.33	0.31	0.33
WCarson.faith	0.33	0.33	0.26	0.26	0.27	0.28
Willow	0.34	0.32	0.36	0.26	0.38	0.25
Wolf	0.26	0.31	0.44	0.26	0.45	0.25
WWalker.Leavitt	0.39	0.43	0.25	0.37	0.24	0.37
WWalker.Pickel	0.38	0.52	0.26	0.45	0.26	0.46
Average B-C Distance =	0.32	0.38	0.33	0.33	0.33	0.33

[a] - data standardizations are explained in detail in the methods and in the Appendix XX table.

[b] - Within-Method similarity is measured as the mean among the 10 similarity permutations among original SNARL replicates

[c] - *These are the mean value for 3 independent randomized re-samplings of the data*

[d] - these analyses compared SNARL and RIVPACS methods because they shared the same refined taxonomic resolution for mites and midges

APPENDIX. List of Metrics Tested for IBI Development (*selected for IBI) and Metric Calculations

Number of individuals per square meter
Total number of individuals identified per sample
***Raw taxa richness per sample**
Taxa richness per sample using CSBP taxa resolution
Raw taxa richness rarefied to 217 bugs per sample
***CSBP taxa richness rarefied to 217 bugs per sample**
Raw taxa richness rarefied to the same number of bugs for each site (but variable across sites)
CSBP taxa richness rarefied to same level for each site
Raw taxa richness rarefied to 243 bugs per sample (but removed 2 SNARL samples)
CSBP taxa richness rarefied to 243 bugs per sample (but removed 2 SNARL samples)
Total number of individuals identified across all samples for each method
***Raw Composite Richness - total number of taxa found in all replicates for a method**
Composite Richness for CSBP taxa resolution - total number of taxa in all replicates
Composite Richness but rarefied to 437 bugs per sample
***Composite Richness but rarefied to 437 bugs per sample and using CSBP taxa resolution**
***Shannon Diversity (H')**
Shannon Diversity for samples standardized to CSBP taxonomic resolution
Community Evenness
Community Evenness standardized to CSBP taxa res.
Simpson's Diversity Measure
Simpson's for CSBP taxa resolution
Hurlbert's "Pi" Diversity Measure
Hurlbert's "Pi" Diversity Measure for CSBP taxa resolution
***Total EPT Taxa per sample**
***Total Ephemeroptera taxa per sample**
***Total Plecoptera taxa per sample**
***Total Trichoptera taxa per sample**
Total Dipteran taxa per sample (includes Chironomids)
Total Dipteran taxa per sample but with CSBP taxa resolution for each sample
Total number of Chironomidae taxa
Total Number of Chironomidae subfamilies (i.e., CSBP taxa resolution)
Total Number of Non-Insect Taxa
Total Number of Non-Insect Taxa but with mites just as "mites"
Percent of Individuals in a Sample which were EPT taxa
***Percent of Taxa which were EPT taxa**
Percent of Individuals in a Sample which were EPT taxa but excluding Baetis and Hydropsychidae
Percent of Individuals which were Chironomidae
Percent of taxa which were Chironomidae
Percent of taxa which were Chironomidae (ID'd only to subfamily)
Ratio of Chironominae abundance to Orthocladiinae abundance
Ration of EPT Richness to Chironomidae Percent Abundance (ept.rich / perc.chiro.abund)
The Percent Abundance of the Most Abundant Taxon in a Sample (Dominance)
The Percent Abundance of the Most Abundant Taxon in a Sample but at CSBP taxa resolution
***The Percent Abundance for the 3 most abundant taxa (Dominance 3)**

Appendix. List of Metrics Tested and Selected for IBI Development (continued)

The Percent Abundance for the 3 most abundant taxa but at CSBP taxa resolution
D-50 Dominance = # taxa to get to 50% abundance of sample
D-50 Dominance = # taxa to get to 50% abundance of sample but using CSBP taxa resolution
***Simple Biotic Index**
Biotic Index but using CSBP taxa resolution
***Number of Taxa which were Intolerant (TV=0,1,2)**
Percent of Taxa which were Intolerant
Percent Abundance of Taxa which were Intolerant
Number of Taxa which were Intolerant (TV=0,1,2) but using CSBP taxa resolution
Percent of Taxa which were Intolerant but using CSBP taxa resolution
Percent Abundance of Taxa which were Intolerant using CSBP taxa resolution
***Percent of Taxa which were Tolerant (TV=7,8,9,10)**
Percent Abundance of Tolerant Taxa
Percent of Taxa which were Tolerant using CSBP taxa resolution
Percent Abundance of Tolerant Taxa using CSBP taxa resolution
Percentage of Collectors
Percentage of Scrapers
Percentage of Filterers
***Percentage of Shredders**
Percentage of Predators
Percentage of Piercers
Percentage of Collectors using CSBP taxa resolution
Percentage of Scrapers using CSBP taxa resolution
Percentage of Filterers using CSBP taxa resolution
Percentage of Shredders using CSBP taxa resolution
Percentage of Predators using CSBP taxa resolution
Percentage of Piercers using CSBP taxa resolution

Note: In addition to IBI development using the 15 metrics listed, an IBI with nearly the same properties was developed using only 5 of these metrics [richness rarefied to 217 count using CSBP taxonomy, EPT richness, biotic index, number of intolerant (TV 0,1,2) taxa, and percent of the taxa that were tolerant (TV 7,8,9,10)].

Appendix (continued) Metric Calculations:

Metric	Formula	Calculation
1. Richness	$S = \sum_{all\ taxa} I_i$	The sum of non-zero unique taxa in each sample.
2. Standardized Richness	$E(S_r) = \sum_{all\ taxa} \left(1 - \frac{\binom{N-N_i}{r}}{\binom{N}{r}} \right)$	The sum of non-zero unique taxa after making 2 adjustments: 1) leaving all midges at subfamily and mites as just "mites"; and 2) standardizing the number of individuals in each sample using Hurlbert's rarefaction formula (Hurlbert 1971).
3. Composite Richness	$S = \sum_{all\ taxa} I_i$	The sum of non-zero unique taxa after pooling all replicate samples at a stream together for each method.
4. Standardized Composite Richness	$E(S_r) = \sum_{all\ taxa} \left(1 - \frac{\binom{N-N_i}{r}}{\binom{N}{r}} \right)$	The sum of non-zero unique taxa after pooling all replicate samples at a stream together for each method and after making the same 2 adjustments outlined above for Standardized Richness.
5. Shannon Diversity	$H' = \sum_{all\ non-zero\ taxa} \frac{N_i}{N} \ln\left(\frac{N_i}{N} \right)$	The standard diversity measure described by Shannon & Weaver.
6. EPT Richness	$EPT_{rich} = \sum_{all\ EPT\ taxa} I_i$	The sum of non-zero unique taxa in each sample belonging to the orders Ephemeroptera, Plecoptera, and Trichoptera.
7. Ephemeroptera Richness	$E_{rich} = \sum_{all\ E\ taxa} I_i$	The sum of non-zero unique taxa belonging to the order Ephemeroptera.
8. Plecoptera Richness	$P_{rich} = \sum_{all\ P\ taxa} I_i$	The sum of non-zero unique taxa belonging to the order Plecoptera.
9. Trichoptera Richness	$T_{rich} = \sum_{all\ T\ taxa} I_i$	The sum of non-zero unique taxa belonging to the order Trichoptera.

10. Percent EPT Richness	$\%\ \text{EPT} = \frac{EPT_{rich}}{S} \cdot 100$	EPT Richness divided by Richness.
11. Dominance of 3 Top Taxa	$\text{Dom} = \sum_{i=(n),(n-1),(n-2)} \frac{N_i}{N}$	Proportional abundance of the 3 most abundant invertebrate taxa in each sample.
12. Biotic Index	$\text{BI} = \sum_{all\ taxa} \frac{N_i \cdot TV_i}{N}$	Average of the abundance of each taxon weighted by that taxon's pollution tolerance score.
13. Number of Intolerant Taxa	$\text{Intol} = \sum_{taxa\ with\ TV \le 2} I_i$	Number of non-zero unique taxa whose pollution tolerance score equaled 0, 1, or 2, on a scale of 0 to 10.
14. Percent Tolerant Taxa	$\text{Tol} = \frac{\sum_{taxa\ with\ TV \ge 7} I_i}{S} \cdot 100$	Percent of total Richness composed of tolerant taxa; Tolerant Taxa defined as the number of non-zero unique taxa whose pollution tolerance score equaled 7, 8, 9, or 10, on a scale of 0 to 10.
15. Percent Shredder	$\text{Shred} = \sum_{all\ shredder\ taxa} \frac{N_i}{N} \cdot 100$	Percent of total invertebrate abundance composed of shredder individuals.

Where N_i=Abundance of i^{th} taxon; N=Total Abundance across taxa ($\sum N_i$); I_i is an indicator variable which takes a value of 1 when $N_i>0$ and which equals 0 when $N_i=0$; r=number of individuals in the standardized rarefaction sample (e.g., 100 individuals if the set level for rarefaction richness were 100 individuals in every sample); $N_{(n)}$= n^{th} ordered abundance value (i.e., most abundant invertebrate out of n taxa); $N_{(n-1)}$= $(n-1)^{th}$ ordered abundance value (i.e., second most abundant invertebrate out of n taxa); TV_i = Tolerance Value of each taxon, which is a value between 0 and 10 and reflects (for higher numbers) increasing ability to tolerate severe natural or anthropogenic environmental conditions.