J. N. Am. Benthol. Soc., 2007, 26(2):000–000
© 2007 by The North American Benthological Society

Comparisons of targeted-riffle and reach-wide benthic macroinvertebrate samples: implications for data sharing in stream-condition assessments

Andrew C. Rehn[1] AND Peter R. Ode[2]

California Aquatic Bioassessment Laboratory, Office of Spill Pollution Response, Department of Fish and Game, 2005 Nimbus Road, Rancho Cordova, California 95670 USA

Charles P. Hawkins[3]

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

Abstract. Recent comparisons of benthic macroinvertebrate (BMI) sampling protocols have shown that samples collected from different habitat types generally produce consistent stream classifications and assessments. However, these comparisons usually have not included biological endpoints used by monitoring agencies, such as multimetric indices (e.g., benthic index of biotic integrity [B-IBI]) or observed-to-expected (O/E) indices of taxonomic completeness, as target variables, and estimates of method precision are rarely provided. Targeted-riffle (TR) and reach-wide (RW) benthic samples have been collected at thousands of sites across the western USA, but little guidance is available for understanding 1) the extent to which raw data sets can be combined in regional or large-scale analyses, 2) the degree of precision afforded by each method, or 3) the efficacy of cross-application of biological indicators derived from one sample type to the other. To address these issues, we used data from 193 sites in California where the Environmental Monitoring and Assessment Program (EMAP) collected the 2 samples side by side. We also conducted a separate study wherein 3 replicates of each sample type were collected from 15 streams to estimate minimum detectable difference (MDD) as a measure of each method's precision. Metrics calculated from TR and RW samples showed similar dose–response relationships to stressors gradients and similar raw scoring ranges. Biological indices (B-IBI, O/E_0, and O/E_{50}) derived from RW samples were more precise than those derived from TR samples, but precision differences were not substantial. On average, pairwise differences in any index between TR and RW sample types were much less than the MDD associated with either sampling method. We observed a weak but consistent bias toward higher O/E_{50} scores from TR samples than from RW samples at the highest elevations and in the largest watersheds. Broad-scale condition assessments were nearly identical when B-IBI and O/E_0 were used as endpoints, and assessments based on O/E_{50} were only slightly less similar. Our analyses indicate that raw data sets and biological indicators derived from TR and RW samples may be generally interchangeable when used in ambient biomonitoring programs.

Key words: benthic macroinvertebrates, bioassessment, sample habitat, index of biotic integrity, predictive models, EMAP, California.

Benthic macroinvertebrates (BMIs) are the most commonly used organisms in freshwater biomonitoring programs (Bonada et al. 2006). Numerous multimetric indices (e.g., benthic index of biotic integrity [B-IBI]), observed-to-expected (O/E) indices of taxonomic completeness, and various other tools have been developed in many parts of the world, including North America (Klemm et al. 2003, Hawkins 2006), Australia (Simpson and Norris 2000), Europe (Moss et al. 1987, De Pauw et al. 1992), New Zealand (Stark 1993), South Africa (Chutter 1972), and Indonesia (Sudaryanti et al. 2001). These biological indicators aid in the interpretation of complex BMI assemblage data and help classify the ecological condition of test sites relative to regional reference conditions (Hughes 1994).

[1] E-mail addresses: arehn@ospr.dfg.ca.gov

[2] pode@ospr.dfg.ca.gov

[3] chuck.hawkins@usu.edu

Recently, B-IBI- and O/E-based assessments have been used in conjunction with probability survey designs to estimate the ecological condition of entire resource populations, such as all mapped wadeable stream lengths within large geographic regions (Herlihy et al. 2000, Stevens and Olsen 2004, Stoddard et al. 2005).

Despite their popular use in biomonitoring, there is no commonly agreed upon method for sampling BMIs or for processing samples (Carter and Resh 2001, Houston et al. 2002). Debates continue regarding which habitat is best to sample (Parsons and Norris 1996), what subsample size of organisms is best (Ostermiller and Hawkins 2004, Cao and Hawkins 2006), and what taxonomic resolution is sufficient to detect anthropogenic impairment (Lenat and Resh 2001, Waite et al. 2004). Decisions about where to sample frequently have been driven by the assumption that index values obtained at sites will be influenced by the types or mixture of habitats sampled rather than by water-quality differences among sites (Chessman 1995), or that certain disturbances (e.g., sedimentation) may have a more pronounced effect on biota in certain habitats and might go undetected if only a single habitat were sampled (Kerans et al. 1992, Parsons and Norris 1996). These assumptions seem to be supported by observations that like habitats can have more similar BMI assemblages among streams than different habitats within a stream (e.g., McCulloch 1986, Parsons and Norris 1996). Nonetheless, growing evidence suggests that BMI samples collected from different habitat types generally produce similar stream classifications and assessments (Hewlett 2000, Ostermiller and Hawkins 2004, Gerth and Herlihy 2006).

Thorough comparison of sampling methods requires evaluation of multiple performance characteristics, including precision, accuracy, bias, and sensitivity (Diamond et al. 1996). Quantitative performance characteristics aid in determinations of whether raw data sets derived from independent programs with different sampling techniques can be combined for larger analyses, and whether biological endpoints (i.e., B-IBI or O/E scores) derived from those programs can be compared directly. To date, comparisons of sampling methods that target different habitats usually have not included estimates of method precision (but see Stark 1993, Houston et al. 2002). Replicate samples are required to estimate the variance associated with sampling error in biological assessments (Barbour et al. 1996, Fore et al. 2001), and documentation of precision has been advocated as an essential component of any performance-based monitoring system (PBMS; Diamond et al. 1996).

We compared the 2 sampling methods (targeted-riffle [TR] and reach-wide [RW]) used by the US Environmental Protection Agency's Environmental Monitoring and Assessment Program (EMAP) survey of wadeable streams in the western USA. First, we evaluated whether the responses of several BMI metrics to gradients of anthropogenic stressors varied if the metrics were calculated from different sample types. Second, we determined whether within-site precision of B-IBI and O/E indices varied with sampling method, and we used within-method precision as a context for evaluating between-method differences in index scores. Third, we assessed whether systematic biases in B-IBI or O/E in relation to several natural gradients (elevation, watershed area, etc.) occurred between sampling methods. Last, we assessed whether sampling method affected site-specific and regional condition assessments based on B-IBI and O/E. If the 2 sampling methods produce comparable data and biological endpoints, raw TR and RW samples could be combined for large-scale analyses, and indicators developed from one sample type could be applied with reasonable confidence to data sets collected with the other.

Methods

Data sets

Data for pairwise comparisons of TR and RW sample types were obtained from 193 sites sampled in California (Fig. 1) during 2000 to 2003 by the western EMAP probability stream survey (Stoddard et al. 2005). Sampling sites were selected randomly from the digitized stream network depicted on 1:100,000-scale US Geological Survey topographic maps to ensure a spatially balanced, representative survey (Herlihy et al. 2000, Stevens and Olsen 2004). At each site, a sampling reach was defined as 40× the average stream width at the center of the reach, with a minimum reach length of 150 m and maximum length of 500 m. Eleven equidistant transects were established, and an RW sample was taken by sampling 0.09 m^2 of substrate with a kick net at each transect. Sampling points alternated among 25%, 50%, and 75% of stream width (thus, RW samples often contained at least some riffle components), and all 11 kick samples were composited into a single sample (Peck et al. 2004). A TR sample was taken from within the same reach by sampling 0.09 m^2 of substrate with a kick net from each of 8 randomly chosen riffle or fastest-water habitat units (Peck et al. 2004). All 8 kick samples were composited into a single sample.

Data for estimates of within-site precision, or sampling error, associated with each method were obtained from 29 streams in northern coastal Califor-

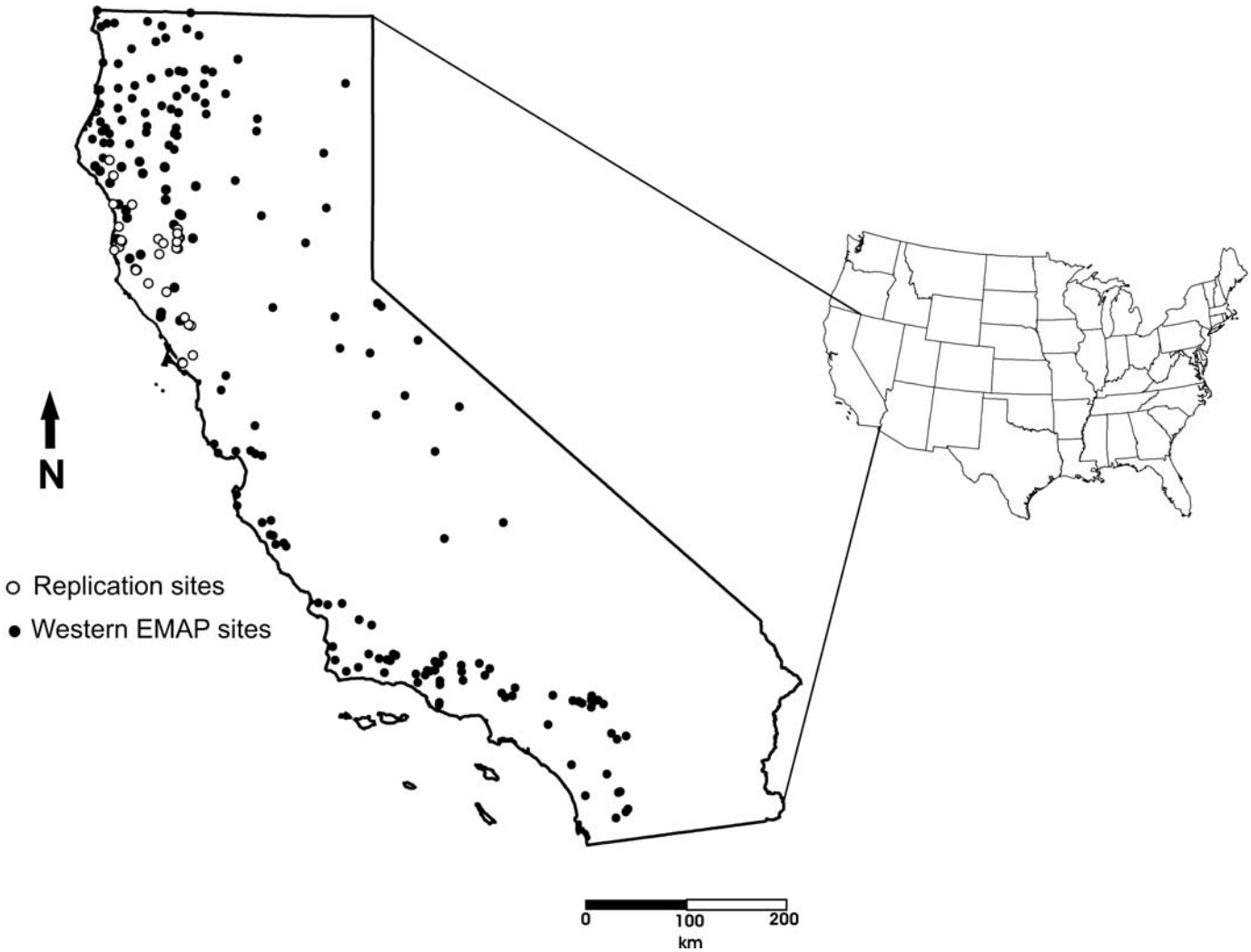


FIG. 1. Map of 193 sampling locations in California where targeted-riffle and reach-wide samples were collected for pairwise comparisons (Environmental Monitoring and Assessment Program [EMAP] sites) and 29 locations where replicate samples were collected for precision estimates.

nia (Fig. 1). Sites were sampled in September 2004 and were selected to represent the range of stream conditions found in the region. Four of the sites had been sampled in previous years by EMAP. At each site, a 150-m sampling reach was established. At 15 sites, 3 TR replicates were collected following the protocol described above after randomly assigning each fastest-water habitat unit in the reach to 1 of 3 bins (Rep 1, 2, or 3). At 15 other sites (except Mark West Creek, where the 2 methods were replicated in adjacent sampling reaches), 3 RW replicates were collected from within the sampling reach following the protocols described above by alternating the sampling position along each transect for each replicate.

In the laboratory, each BMI sample was rinsed carefully in a 0.5-mm-mesh sieve before being transferred to a 20 × 25-cm tray subdivided into a grid of 20 squares. Organisms were subsampled from randomly chosen squares until 500 individuals were picked from each sample. Insects were identified to genus with standards of taxonomic effort defined by the California Aquatic Macroinvertebrate Laboratory Network (www.dfg.ca.gov/cabw/camlnetste.pdf). Chironomid genera were lumped at the subfamily level for analyses described below.

Data analyses

Metrics comparisons.—Dose–response relationships of 11 biological metrics to 5 anthropogenic or human-influenced stressors (% sand and fines, conductivity, total N, qualitative channel alteration, and local road density) known to be associated with

biological degradation were examined to determine whether the relationships differed for TR and RW sample types. The evaluated metrics were chosen because they are used currently in California B-IBIs that were developed from TR sample data (Ode et al. 2005, Rehn et al. 2005). Percent sand and fines, qualitative channel alteration, conductivity, and total N were measured at study reaches with EMAP protocols (Klemm and Lazorchak 1994, Peck et al. 2004). Local road densities were obtained through geographical information system (GIS) analyses. First, a polygon delineating the area drained within a 1-km radius upstream of each study reach was defined. Then the ArcView® (version 3.2; Environmental Systems Research Institute, Redlands, California) extension ATtILA (version 3.0; US Environmental Protection Agency, Washington, DC) was used to calculate road densities within polygons with a road network obtained from the US Forest Service Remote Sensing Lab (http://fsweb/gis/gis_data/calcovs/fs/nwctran03_2.html).

Linear regression was used to quantify the strength of each metric–stressor relationship for each sample type. In cases where relationships were clearly wedge shaped (i.e., had distinct ceilings or floors), upper-bound (or lower-bound) regression was used to quantify the limiting slope of the relationship (Blackburn et al. 1992). For this analysis, the stressor axis was divided into 10 equal-interval bins and either the 3 highest or 3 lowest metric values were selected from each bin. Ordinary least-squares regressions were then calculated for the subsets of data to estimate the upper- or lower-bound slopes of wedge-shaped polygons. As an approximate Bonferroni correction for a large number of correlations, only relationships with a p-value ≤ 0.0001 were considered significant. Box plots and Mann–Whitney U tests were used to evaluate whether raw metrics differed between TR and RW samples and might require different scaling in a B-IBI.

Minimum detectable difference (MDD).—Replicate samples allow estimation of the variance in metric or composite indicator values associated with sampling error. We were interested in the variance of actual endpoint indicators used by water-quality managers in California. Northern coastal California B-IBI scores (Rehn et al. 2005) were calculated for each TR and RW replicate from the 29 replication sites. The replicate samples also were assessed with a recently developed California O/E index (CPH, unpublished data). The index was based on TR samples and generates 2 O/E taxa ratios, one based on taxa with modeled site-specific probabilities of capture >0 (O/E_0) and another based on taxa with site-specific probabilities of capture $\geq$0.5 (O/E_{50}; see Ostermiller and Hawkins 2004 for further explanation). Nested analyses of variance (ANOVAs) with replicate samples nested within sites were used to estimate the average within-site variance (as mean squared error [MSE] with 30 df) for both B-IBI and O/E values. These estimates of MSE were then applied in 2-sample t-tests ($\alpha = 0.05$, $\beta = 0.10$) to calculate the MDD for each indicator (Zar 1999, Fore et al. 2001). The MDD provides a measure of how different B-IBI or O/E values must be before they are considered significantly different.

Pairwise comparisons of B-IBI and O/E scores.—Pairwise differences were evaluated between recently developed California B-IBI (Ode et al. 2005, Rehn et al. 2005) scores calculated from TR and RW sample types. Two sites were eliminated from B-IBI comparisons because of low sample counts (<450). Pairwise differences between O/E scores were evaluated for a subset of 187 statewide sites where sample counts were sufficiently large ($n \geq 300$) after taxon lists were reduced to those operational taxonomic units (OTUs) used in the index.

Average pairwise differences in B-IBI and O/E scores between TR and RW sample types and the number of cases where the pairwise differences in these 2 indicator values exceeded the MDD for each sampling method were calculated. The degree to which B-IBI and O/E discriminated between reference and test sites depending on whether they were calculated from TR or RW samples was also evaluated. A principal components analysis (PCA) of the 5 stressors used in metrics comparisons was done, and the responsiveness of B-IBI and O/E to the first PCA axis (PCA1) was plotted. Our purpose was not to compare responsiveness between indicators, but rather to evaluate whether each indicator showed different responses when calculated from TR and RW sample types. Last, to determine whether the effect of sampling method on indicator values was influenced by natural gradients or by the extent of human influence, pairwise differences in TR- and RW-derived indicator values were plotted against watershed area, elevation, mean channel slope, % fast-water habitat in the sample reach, and PCA1.

Condition assessments.—Use of a spatially balanced probability process for site selection in regional stream surveys is well documented (Herlihy et al. 2000, Stevens and Olsen 2004). In short, each EMAP site in California represented a portion of the total perennial wadeable stream length in the state, and the status of the total stream population was inferred from the sample data. Our purpose here was not to report on the condition of wadeable streams in California per se, but rather to present a comparison of condition assessments based on TR and RW sample types and

TABLE 1. r^2 values from mean and upper- (or lower-) bound regressions between metrics used in California benthic indices of biotic integrity (B-IBI) and example stressor gradients used in metric screening. TR = targeted-riffle samples, RW = reach-wide samples, TN = total N, EPT = Ephemeroptera, Plecoptera, and Trichoptera taxa. – indicates relationships that did not have ceilings or floors. Significant values ($p \leq 0.0001$) are shown in bold.

	EPT richness		Coleoptera richness		Diptera richness		Predator richness		% collector individuals		% intolerant individuals		% nongastropod scrapers		% noninsect taxa		% predator individuals		% shredder taxa		% tolerant taxa	
	TR	RW	TR	RW	TR	RW	TR	RW	TR	RW	TR	RW	TR	RW	TR	RW	TR	RW	TR	RW	TR	RW
Mean																						
% sand and fines	**0.33**	**0.42**	**0.14**	**0.21**	0.04	0.06	**0.13**	**0.23**	**0.11**	**0.07**	**0.16**	**0.18**	**0.14**	**0.16**	**0.21**	**0.28**	0.006	0.005	**0.13**	**0.1**	**0.28**	**0.33**
Conductivity	**0.25**	**0.25**	0.05	0.05	0.02	0.03	**0.16**	**0.16**	0.06	0.05	**0.17**	**0.15**	**0.10**	**0.09**	**0.12**	**0.20**	0.005	0.001	**0.12**	**0.12**	**0.23**	**0.30**
Log_{10}(TN)	**0.34**	**0.31**	0.04	**0.08**	**0.09**	**0.09**	**0.23**	**0.25**	0.003	0.003	**0.13**	**0.10**	0.05	0.04	**0.33**	**0.39**	0.001	0.03	**0.18**	**0.15**	**0.27**	**0.31**
Channel alteration	**0.14**	**0.21**	**0.07**	**0.11**	**0.08**	**0.09**	**0.08**	**0.18**	0.001	0.005	0.03	0.06	0.03	0.04	**0.24**	**0.32**	**0.07**	0.01	0.04	0.04	**0.19**	**0.24**
Local road density	**0.11**	**0.14**	0.06	**0.10**	0.06	0.06	0.06	**0.12**	0.006	0.002	**0.08**	0.06	0.03	0.04	**0.18**	**0.23**	0.03	0.01	**0.07**	**0.09**	**0.18**	**0.14**
Upper (or lower) bound																						
% sand and fines	**0.79**	**0.84**	**0.77**	**0.80**	–	–	**0.41**	**0.76**	**0.48**	0.38	**0.69**	**0.61**	**0.65**	**0.65**	–	–	–	–	**0.51**	**0.58**	–	–
Conductivity	**0.76**	**0.77**	0.53	0.53	**0.76**	**0.71**	**0.60**	**0.61**	0.30	0.26	**0.61**	**0.55**	0.53	**0.55**	–	–	0.01	0.16	**0.55**	**0.63**	–	–
Log_{10}(TN)	**0.67**	**0.7**	–	–	**0.45**	**0.41**	**0.70**	**0.71**	–	–	**0.57**	**0.51**	–	–	–	–	–	–	**0.52**	**0.56**	–	–
Channel alteration	**0.77**	**0.82**	**0.8**	**0.84**	**0.73**	**0.73**	**0.69**	**0.82**	–	–	**0.64**	**0.69**	**0.76**	**0.76**	**0.73**	**0.81**	**0.55**	0.38	**0.50**	**0.53**	**0.57**	**0.74**
Local road density	**0.63**	**0.63**	**0.67**	**0.66**	**0.47**	**0.53**	**0.52**	**0.72**	0.28	0.33	**0.59**	**0.59**	–	–	–	–	–	–	**0.60**	**0.64**	**0.69**	**0.60**

to evaluate how robustly TR-derived indicators could be used to assess RW-derived samples. The R statistical program (R Foundation for Statistical Computing, Vienna, Austria; http://www.R-project.org) and an R contributed library (psurvey.analysis, www.epa.gov/nheerl/arm) were used to plot the cumulative distribution of B-IBI and O/E scores in the population of wadeable streams in California. Cumulative distribution functions (CDFs) and their 95% confidence intervals were used to evaluate whether assessments derived from different combinations of sample type and indicator produced similar stream-condition assessments in California.

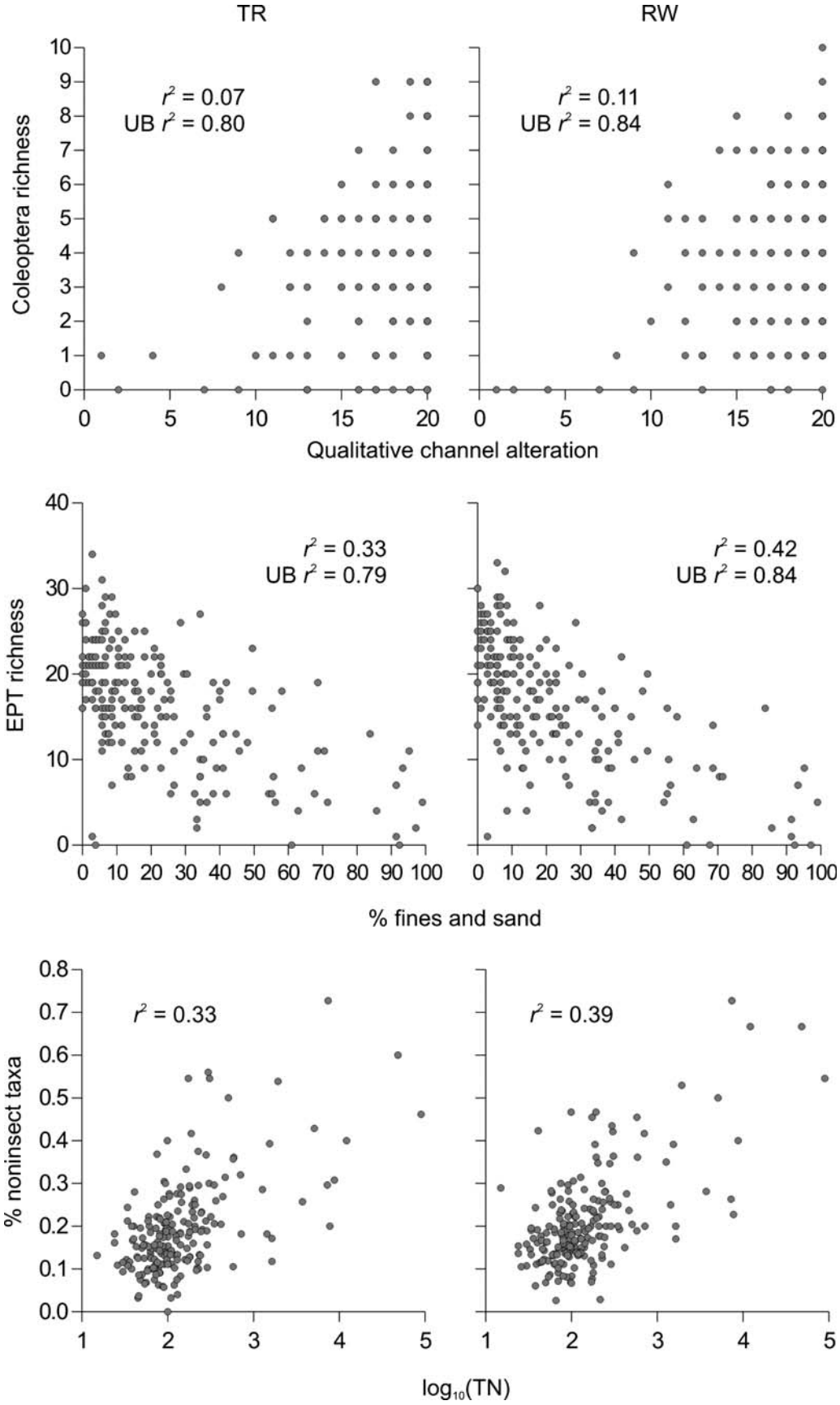
Results

Metrics comparisons

Metrics showed similar responses to stressor gradients regardless of whether they were calculated from TR or RW samples (Table 1, Fig. 2). In most cases, relationships were slightly tighter (r^2) when metrics were calculated from RW samples. Interquartile ranges of TR and RW samples were strongly overlapping (Fig. 3). Of the 4 metrics for which medians differed significantly different between sample types (Mann–Whitney U, $p < 0.05$), adjustments in scoring ranges to account for sample-type differences had little or no effect on resulting B-IBI scores. For example, predator richness was most different between TR and RW sample types ($p < 0.0001$; Fig. 3). This metric is used in the southern coastal California B-IBI where scoring ceilings were set as the 80th percentile of the reference-site distribution (Ode et al. 2005). The 80th percentile of reference-site predator richness was 13 for TR samples and 15 for RW samples. Therefore, consequent adjustments in overall metric and B-IBI scoring were minute. Current California B-IBIs were used as the biological endpoints in within-site precision comparisons even though the B-IBIs were developed with data from TR samples because of the similar responses of TR- and RW-derived metrics to stressors and similar ranges of raw metric values.

MDD

The MDD for B-IBI values adjusted to a 100-point scale was 15.5 for the RW sampling method and 19.7 for the TR sampling method (Figs 4A, B). Thus, we have a 90% chance of detecting a 15.5-point difference between RW-based B-IBI scores or a 19.7-point difference between TR-based B-IBI scores at a p-value <0.05. The RW method was slightly more precise than the TR method, but the difference in MDD between the 2 methods was small.


TR
RW
Coleoptera richness
r^2 = 0.07
UB r^2 = 0.80
r^2 = 0.11
UB r^2 = 0.84
Qualitative channel alteration
EPT richness
r^2 = 0.33
UB r^2 = 0.79
r^2 = 0.42
UB r^2 = 0.84
% fines and sand
% noninsect taxa
r^2 = 0.33
r^2 = 0.39
$\log_{10}$(TN)

Six sites were excluded from MDD estimates for O/E scores because of low sample counts in at least one of the replicates after reduction of taxon lists to OTUs used by the index, so our estimate of average within-site variance in O/E scores was slightly less robust than for B-IBI. The O/E MDD ranged from 0.19 to 0.31, depending on sample type and probability-of-capture threshold (O/E_0 vs O/E_{50}; Figs 5A–D).

Pairwise comparisons of B-IBI and O/E scores

B-IBI scores calculated from TR and RW sample types were highly correlated (Fig. 6A), as were O/E values (Figs 6B, C). Pairwise differences between TR and RW B-IBI and O/E scores were usually less than the corresponding within-method MDD (~83–92% agreement depending on the indicator and sampling method; Table 2). When pairwise differences exceeded MDD, values for TR samples were more often higher than those for RW samples when B-IBI and O/E_{50} were used as biological endpoints, but this pattern was not observed when O/E_0 was used as the endpoint (Table 2, Fig. 7).

TR- and RW-derived indices discriminated equally between reference and test sites (Fig. 8). Discrimination between reference and test sites was illustrated separately for northern and southern coastal California because the large number of high-quality EMAP test sites in the north coast obscured otherwise good discrimination observed in the south coast when all data were plotted together. TR- and RW-derived indices also showed similar responses (sensitivity) to a multivariate stressor axis (PCA1; Table 3, Fig. 9).

In general, little or no systematic bias was observed in pairwise differences between indicator scores in relation to watershed area, elevation, mean slope, % fast-water habitat in the sample reach, or PCA1 (Fig. 7). At the highest elevations, at sites with the largest watersheds, and where the sampling reach was predominantly slow water (>80%), O/E_{50} scores usually were higher if calculated from TR samples rather than RW samples (see ellipses in Fig. 7). However, many of these pairwise differences did not exceed the MDD for each combination of indicator and sampling method, and the trends were based on few data points. In no case was the pairwise difference in B-IBI or O/E_0 scores related to the natural or disturbance gradients we tested.

Condition assessments

Condition assessments for perennial streams in California based on TR and RW sample types collected at probability-survey sites were nearly identical for B-IBI and O/E_0 (Figs 10A, B). CDFs of indicator scores derived from each sample type were strongly overlapping, and each sampling method's CDF was within the 95% confidence interval of the other. Agreement in condition assessments based on TR and RW sample types was lower when O/E_{50} was used as the biological indicator, but the RW curve was still almost always within the 95% confidence interval of the TR curve (Fig. 10C). This greater difference implies that it may be less appropriate to apply a TR-derived O/E_{50} index than a B-IBI or an O/E_0 index to RW samples because only the most common riffle taxa (i.e., taxa with site-specific probabilities of capture $\geq$0.5) are included.

Discussion

As the popularity of BMI-based bioassessment has grown, interest also has grown in comparability between benthic data sets collected with different sampling protocols and in the precision associated with these protocols. Targeted-riffle and reach-wide BMI samples have been collected at thousands of sites across the western USA, but little guidance is available for understanding 1) the extent to which raw data sets can be combined in regional or large-scale analyses, 2) the degree of precision afforded by each method, or 3) the efficacy of cross-application of biological indicators derived from one sample type to the other. We used several approaches to address these issues and noted only minor systematic differences in indicator values between sample types across a range of stream types and levels of impairment. In addition, our documentation of performance characteristics for TR and RW sampling may help agencies establish assessment (condition) criteria that reflect true differences in assessment scores.

Sensitivity to stressor gradients

Few studies have compared the responses of metrics calculated from different sample types to stressor gradients. Klemm et al. (2003) found that riffle metrics were significantly correlated with more stressors than were pool metrics in the EMAP survey of Mid-Atlantic Highland streams. Even so, Klemm et al. (2003) were

←

FIG. 2. Example dose–response relationships of benthic macroinvertebrate (BMI) metrics to stressor gradients. Metrics calculated from targeted-riffle (TR) samples are shown on the left, and the same metrics calculated from reach-wide (RW) samples are shown on the right. r^2 values are from ordinary linear and upper-bound (UB) regressions. TN = total N, EPT = Ephemeroptera, Plecoptera, and Trichoptera taxa.

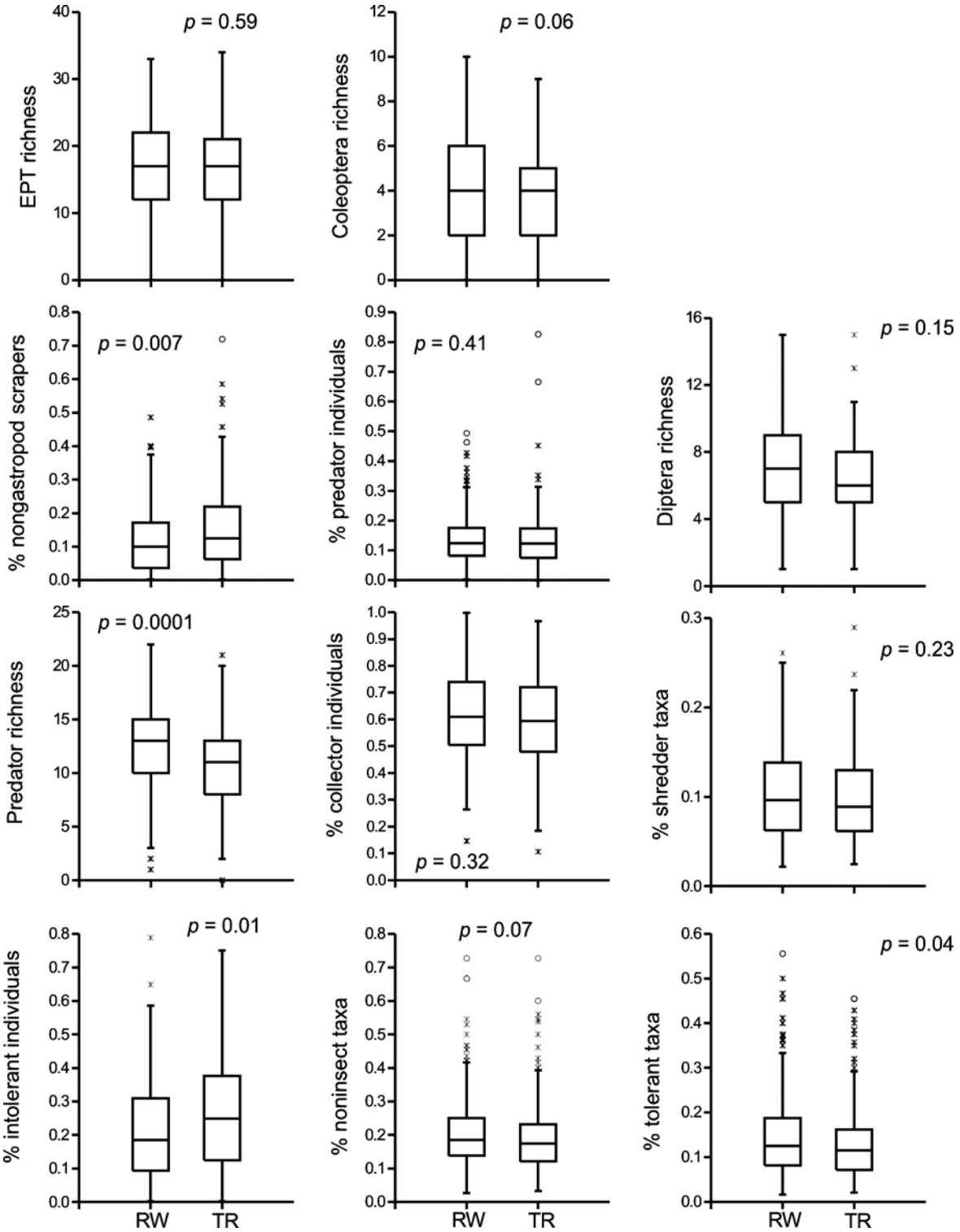


FIG. 3. Comparisons of raw benthic macroinvertebrate (BMI) metric values calculated from targeted-riffle (TR) and reach-wide (RW) samples. Boxes indicate median values and interquartile ranges, whiskers indicate 95th percentiles, outliers are indicated by an x or a circle. EPT = Ephemeroptera, Plecoptera, and Trichoptera taxa. $n = 201$, p-values from Mann–Whitney U tests are indicated.

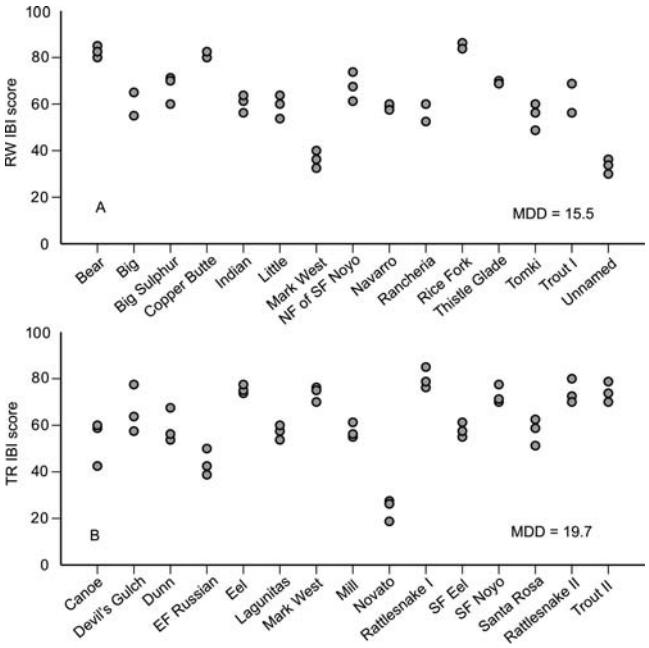


FIG. 4. Replicate benthic index of biotic integrity (B-IBI) scores calculated for 3 reach-wide (RW) samples collected at 15 sites (A) and 3 targeted-riffle (TR) samples collected at 15 sites (B). Replicates were used in estimation of minimum detectable difference (MDD) for each method.

able to use identical metrics for separate riffle and pool samples to develop a regional B-IBI, and had to adjust only the metric scoring scales to account for habitat differences. Using the same data set, Gerth and Herlihy (2006) found considerable differences between BMI assemblages in riffle and pool samples and found that Ephemeroptera, Plecoptera, and Trichoptera (EPT) richness and taxon richness were higher in riffles than in pools. Despite these overall differences, assessments (i.e., percentages of sites in either good or poor biological condition based on EPT richness) were not substantially influenced by sample type.

In our study, metrics calculated from TR and RW showed similar responsiveness to various stressors and similar scoring ranges, indicating that raw data from these 2 sample types can be combined in development of regional B-IBIs. We presented only a few examples of individual metric responses to stressors, but we conducted similar comparisons for >70 BMI metrics and found no consistent differences in metric sensitivity to stressor gradients depending on whether they were derived from TR or RW samples. Parsons and Norris (1996) did not evaluate metric responsiveness, but found considerable data redundancy between riffle and edge samples collected in wadeable streams in the Australian Capital Territory, and that O/E indices based on either sample type (or

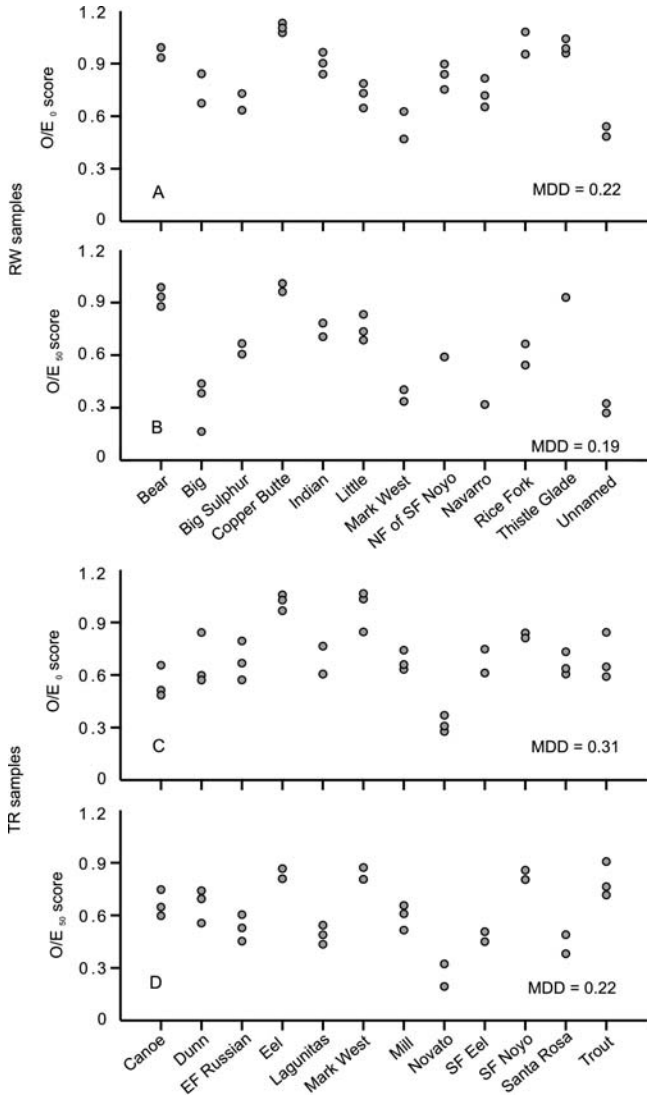
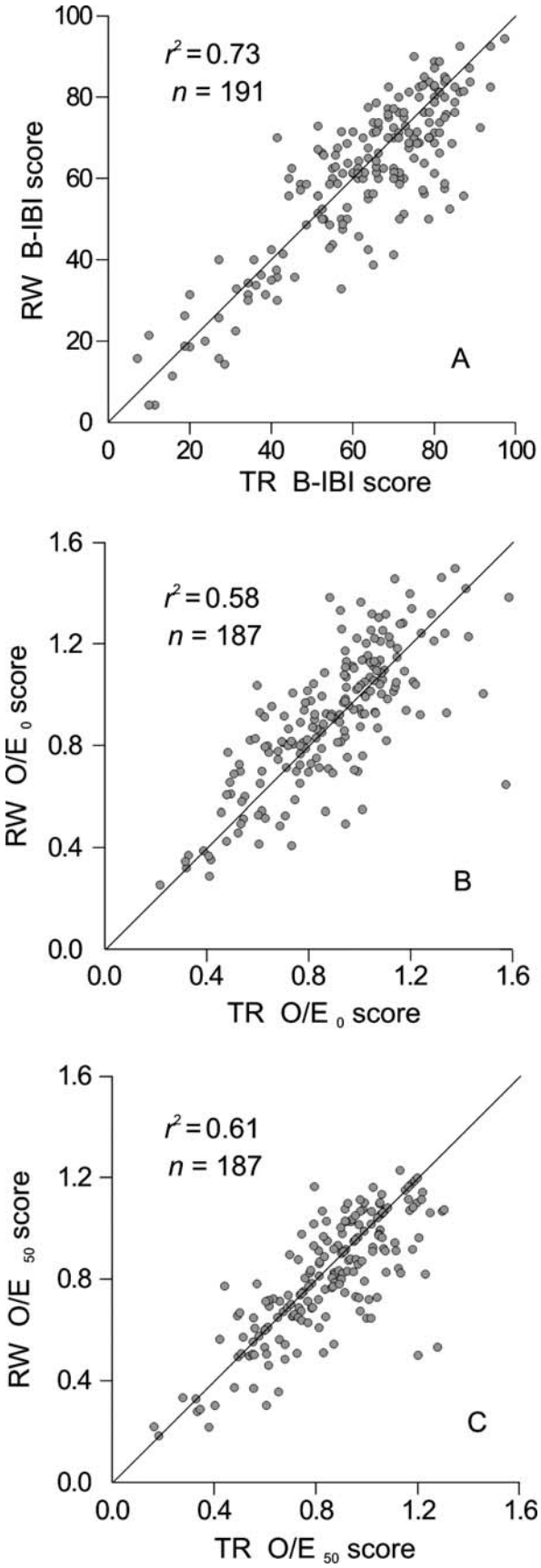


FIG. 5. Replicate observed-to-expected (O/E) index of taxonomic completeness scores calculated as O/E_0 (A, C) and O/E_{50} (B, D) for 3 reach-wide (RW) samples collected at 12 sites (A, B) and 3 targeted-riffle (TR) samples collected at 12 sites (C, D). Replicates were used in estimation of minimum detectable difference (MDD) for each method. Subscripts on O/E ratios indicate site-specific probabilities of capture >0 or ≥0.5 (O/E_0 and O/E_{50}, respectively).

combined samples) were equally capable of detecting biological impairment. Together, these results do not support the hypothesis that certain disturbances have a more pronounced effect on biota in certain habitats that might go undetected were only a single habitat sampled. However, these results might not extend beyond wadeable streams. For example, Blocksom and Flotemersch (2005) found that metrics significantly correlated with stressor gradients varied among 5 sampling methods for nonwadeable streams in Ken-



tucky and Ohio and concluded that raw data were not interchangeable.

Method precision

Indicators derived from different sampling methods may have equal precision, but may not necessarily produce identical site assessments (Cao and Hawkins 2006, Hawkins 2006). We chose MDD as the measure of method precision because it provided a statistical criterion to evaluate whether indicators calculated from TR and RW samples produced equivalent site assessments. Classification strength (Van Sickle 1997) or sampling-method comparability (Cao et al. 2005) can be used to quantify the comparability of raw taxa lists collected with different sampling methods, but similarity analyses provide no statistical criterion to determine whether assessment endpoints differ between sampling methods. Moreover, low taxonomic similarity does not necessarily result in disagreement between metric or B-IBI scores derived from different sample types. The coefficient of variation (CV) of indicator values among reference sites also has been used to estimate sampling-method precision, but has the disadvantage that it incorporates among-site variation in addition to sampling error.

Estimates of all indicator values (B-IBI, O/E_0, O/E_{50}) derived from RW samples were slightly more precise than those derived from TR samples (Figs 4, 5). Between-method differences in MDD were usually small, but RW-derived indicators (B-IBI, O/E_0, or O/E_{50}) were capable of detecting ~1 more condition category than TR-derived indicators (as determined by dividing the indicator scoring range by MDD). Contrary to bioassessment dogma, targeted-habitat sampling did not reduce within-site sampling error relative to multihabitat sampling, and thus, RW sampling may provide water-resource agencies with slightly more sensitive indicators. We suggest the following potential explanations for this observation: 1) the RW protocol sampled an additional 0.27 m^2 of substrate compared to the TR protocol, and the added sampling effort may have been sufficient to produce slightly more precise indicators; 2) the RW protocol, in which sampling was more systematic and spatially balanced, may have reduced sampling error compared

←

FIG. 6. Correlations between benthic index of biotic integrity (B-IBI) scores (A), observed-to-expected (O/E) index of taxonomic completeness O/E_0 (B) and O/E_{50} (C) scores calculated from targeted-riffle (TR) and reach-wide (RW) sample types. Subscripts on O/E ratios indicate site-specific probabilities of capture >0 or ≥0.5 (O/E_0 and O/E_{50}, respectively). r^2 values are from ordinary linear regressions.

TABLE 2. Summary of pairwise differences in biological indicator scores calculated from targeted-riffle (TR) and reach-wide (RW) sample types, and the percentage (number) of sites where pairwise differences exceeded minimum detectable difference (MDD); $n = 191$ for benthic index of biotic integrity (B-IBI) comparisons, $n = 187$ for observed-to-expected (O/E) index of taxonomic completeness comparisons. Subscripts on O/E ratios indicate site-specific probabilities of capture >0 or ≥0.5 (O/E_0 and O/E_{50}, respectively).

Summary of pairwise differences in indicator scores	B-IBI	O/E_0	O/E_{50}
Range in absolute differences	0–31.4	0–0.93	0–0.75
Mean absolute difference	7.8	0.13	0.1
% of sites exceeding TR MDD:			
RW scored higher	1.5% (3)	3.7% (7)	2.7% (5)
TR scored higher	6.8% (13)	4.3% (8)	11.2% (21)
% of sites exceeding RW MDD:			
RW scored higher	2.6% (5)	9.1% (17)	3.7% (7)
TR scored higher	8.9% (17)	6.9% (13)	13.4% (25)

to the TR protocol, in which eligible sample habitats were chosen by field crews; 3) riffle taxa may have had patchier distributions than taxa in other habitats in the streams, making TR-derived indicators more susceptible to sampling error and, therefore, less precise. In any case, TR and RW sample types may have sufficiently similar precision from a PBMS perspective (Diamond et al. 1996) for comparable assessment

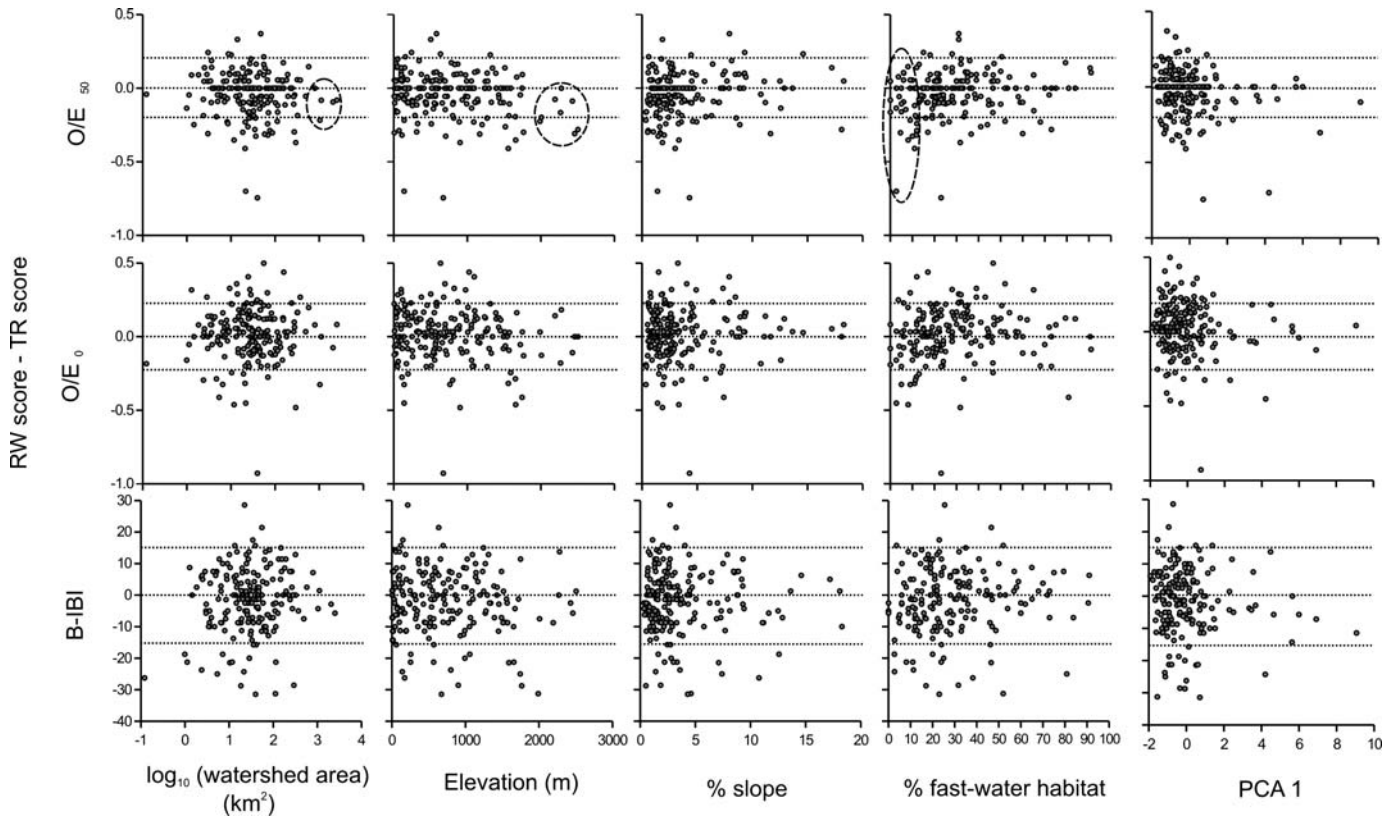


FIG. 7. Pairwise differences in benthic index of biotic integrity (B-IBI) and observed-to-expected (O/E) index of taxonomic completeness scores calculated from targeted-riffle (TR) and reach-wide (RW) sample types in relation to selected natural and disturbance gradients. Subscripts on O/E ratios indicate site-specific probabilities of capture >0 or ≥0.5 (O/E_0 and O/E_{50}, respectively). Horizontal dashed lines show the lowest minimum detectable difference (MDD) for each biological indicator. Pairwise differences between 0 and either the lower or upper MDD lines are not statistically significant. Ellipses were drawn subjectively and show potential conditions where indicator scores from TR samples are consistently higher than scores from RW samples, although many points in the ellipses do not represent statistically significant pairwise differences. PCA1 = principal components analysis axis 1.

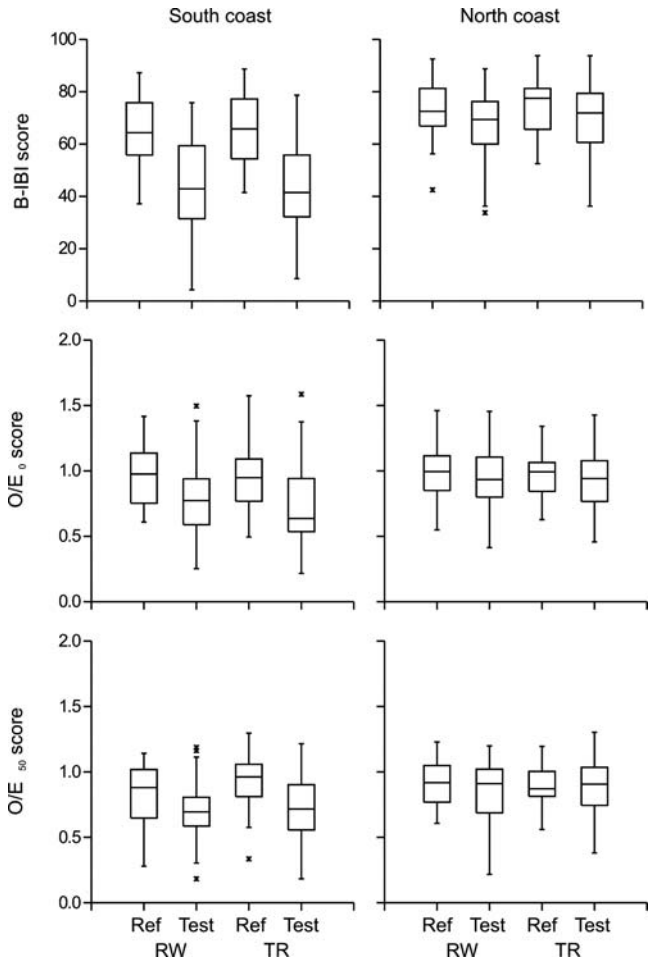


FIG. 8. Discrimination of benthic index of biotic integrity (B-IBI) and observed-to-expected (O/E) index of taxonomic completeness scores between reference (ref) and test sites based on reach-wide (RW) and targeted-riffle (TR) sample types. Subscripts on O/E ratios indicate site-specific probabilities of capture >0 or ≥0.5 (O/E_0 and O/E_{50}, respectively). Discrimination is illustrated by region because of the high frequency of good-quality test sites in northern coastal California. Symbols are as in Fig. 3.

results derived from either method (but see Cao and Hawkins 2006 for a fuller treatment of comparability issues).

Pairwise comparisons

On average, pairwise differences between TR and RW sample types for any indicator were much less than either method's MDD (Table 2). Our preliminary evaluations of raw metrics and the relatively high assemblage similarity between TR and RW sample types (Gerth and Herlihy 2006) indicated that riffle biases may not be present. The slight tendency for TR-derived indicators to overestimate impairment if applied to RW samples (Fig. 7) may be because riffles tend to have more taxa than other habitats. Given equal sampling effort, taxa should accrue more rapidly in TR samples than in RW samples. However, in the western EMAP survey, EPT richness did not differ between riffle and reach-wide samples and taxon richness was higher, on average, in reach-wide samples than in riffle samples (Gerth and Herlihy 2006). Therefore, the small riffle bias we observed may be partly because we used TR-derived indicators for comparisons.

TABLE 3. Loadings of stressor variables on the first principal components axis (PCA1; 55% of total variance explained).

Variable	Axis 1
% sand and fines	0.45
Conductivity	0.46
Log_{10} total N	0.53
Qualitative channel alteration	−0.43
Local road density	0.36

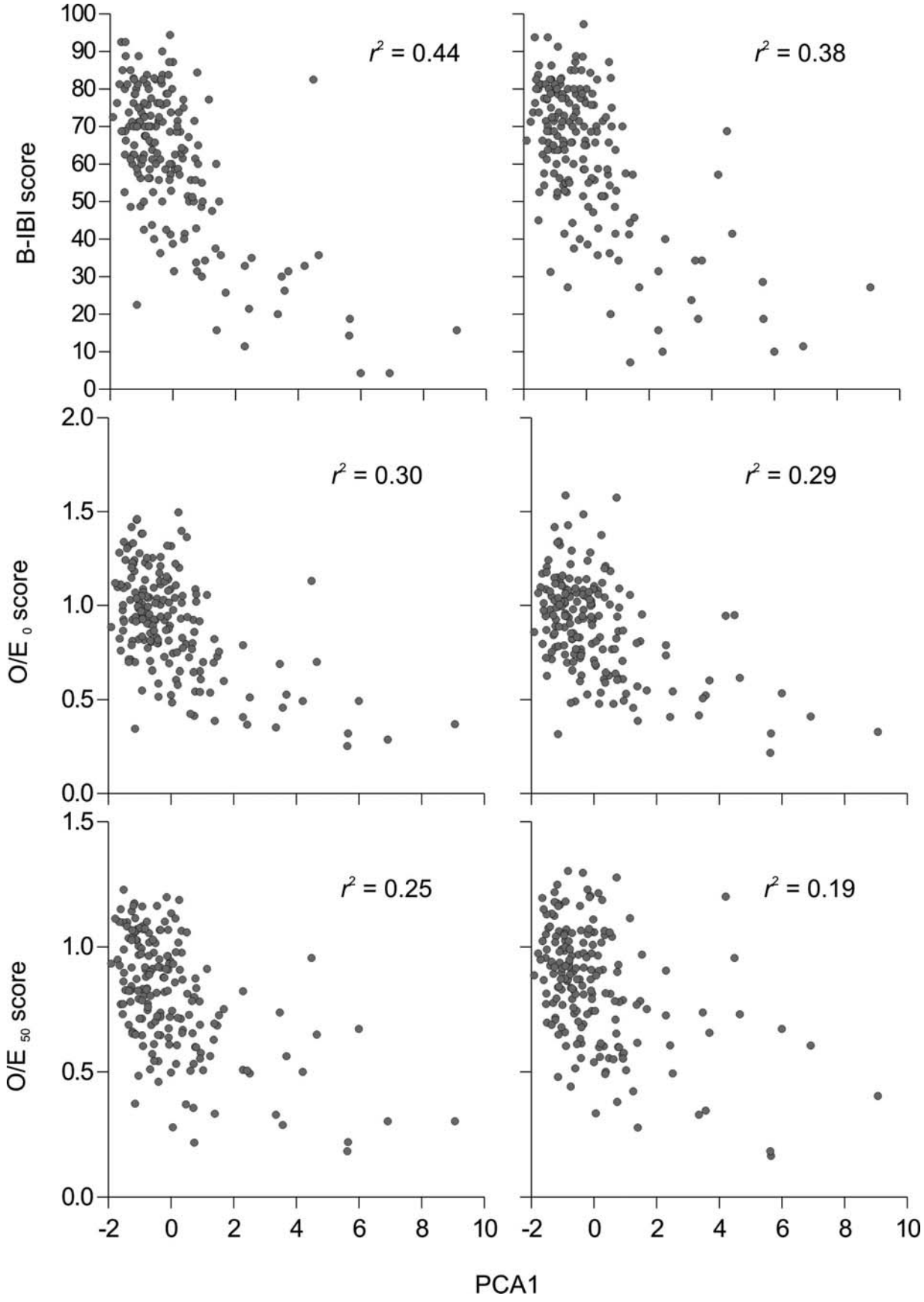
Gerth and Herlihy (2006) observed decreasing Bray–Curtis similarity between TR and RW sample types as % fast-water habitat in the sampling reach decreased. However, we did not observe substantial increases in pairwise differences in indicator scores as % fast-water habitat decreased, even for O/E (which is more akin to Bray–Curtis similarity than B-IBI). At the highest elevations, at sites with the largest upstream watersheds, and at sites with the most human influence, O/E_{50} scores were almost always higher when calculated from TR samples than from RW samples, but the pairwise differences usually did not exceed within-method sampling error (MDD). Therefore, evidence for systematic biases in relation to natural and disturbance gradients is not strong.

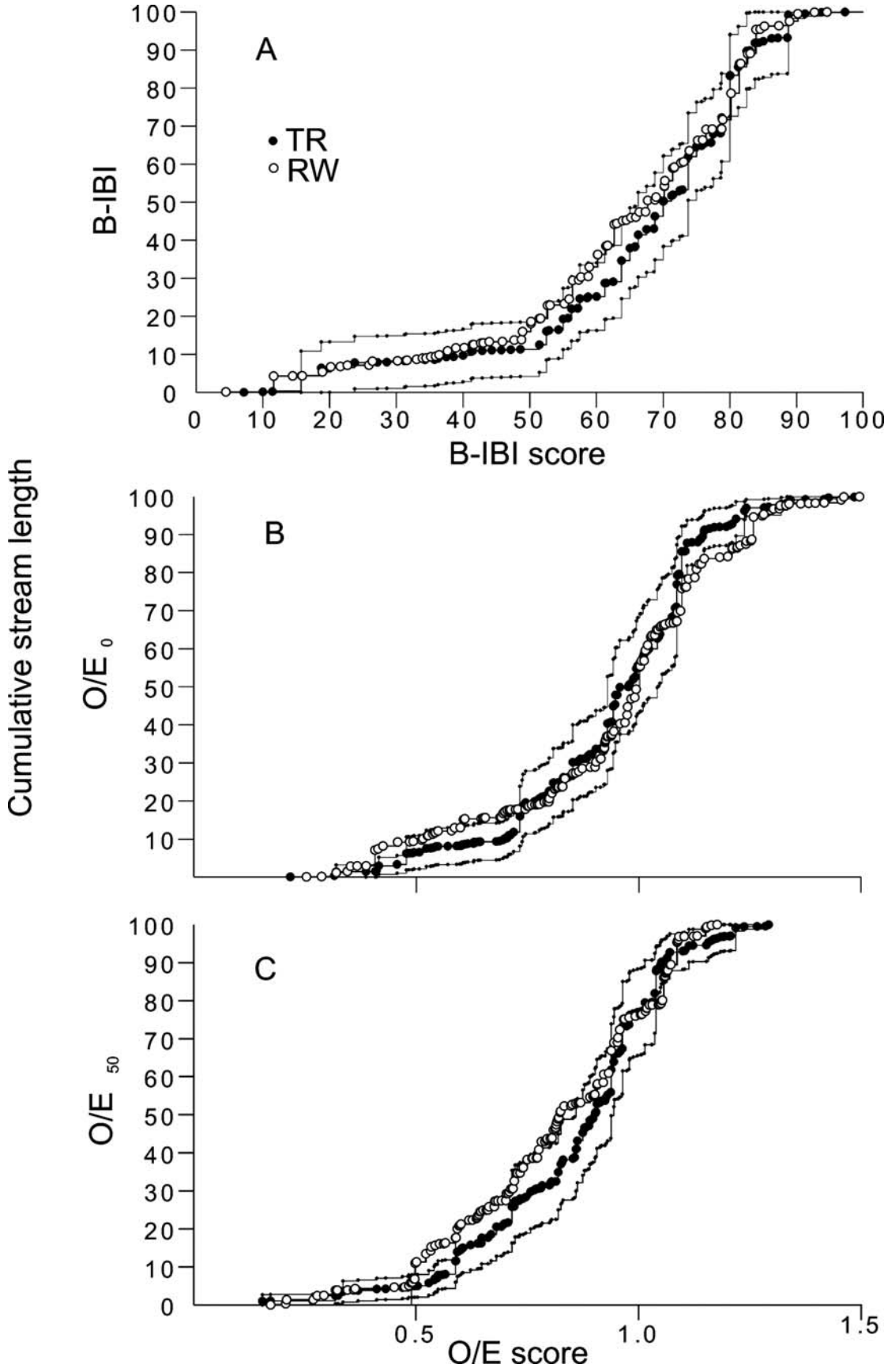
Condition assessments

Condition assessments were nearly identical when based on B-IBI and O/E_0 (Figs 10A, B), but were less similar when based on O/E_{50} (Fig. 10C). Therefore,

→

FIG. 9. Responsiveness of benthic index of biotic integrity (B-IBI) and observed-to-expected (O/E) index of taxonomic completeness based on targeted-riffle (TR) and reach-wide (RW) sample types to a composite stressor axis from principal components analysis (PCA1). Composite axis includes 5 stressor gradients: % sand and fines, conductivity, total N, qualitative channel alteration, and local road density. r^2 values are from ordinary linear regressions. Subscripts on O/E ratios indicate site-specific probabilities of capture >0 or ≥0.5 (O/E_0 and O/E_{50}, respectively).


B-IBI score
O/E 0 score
O/E 50 score
PCA1
r2 = 0.44
r2 = 0.38
r2 = 0.30
r2 = 0.29
r2 = 0.25
r2 = 0.19


A
TR
RW
B-IBI
B-IBI score
B
O/E 0
C
O/E 50
O/E score
Cumulative stream length
0
10
20
30
40
50
60
70
80
90
100
0.5
1.0
1.5

cross-application of indicators may be most appropriate when analyses are based on entire taxa lists. Winnowing taxonomic data sets to include only the most common taxa from a single habitat (riffles) may exaggerate differences between sample types, although it does produce more precise models. Therefore, a tradeoff may exist between greater accuracy and precision in models that exclude rare taxa and greater sample-type comparability in models that include rare taxa. Compromise models (e.g., models in which taxa with a predicted probability of occurrence $\geq$25% define expected conditions) may balance the tradeoff between model precision and cross-application of biological indicators.

Our results also are generally consistent with the results of other studies, including those of Hewlett (2000) who found that riffle, edge, and combined-habitat samples produced similar classifications of 165 sites in Victoria, Australia, and that taxonomic resolution was the most influential feature affecting patterns in reference-site data. Ostermiller and Hawkins (2004) found that O/E indices generated from targeted-riffle cf. timed multihabitat samples collected from wadeable streams in western Oregon and Washington were approximately equally precise. Ostermiller and Hawkins (2004) did show that assessments based on different habitat types sometimes resulted in different site-specific inferences of impairment, but that agreement improved as subsample size increased. For example, the percentage of test sites classified as impaired differed by only 1% when sample counts were $\geq$400 individuals.

In sum, broad-scale methods comparisons have consistently shown that analyses of BMI assemblages are robust to habitat differences and generally produce consistent stream-condition assessments and classifications. Therefore, the potential advantages of combining TR and RW samples for large-scale analyses, or of directly comparing assessment results based on either sample type, may greatly outweigh the apparently small problems associated with data compatibility. Development of accurate method-specific performance characteristics requires substantial data, but agencies may wish to conduct within-site repeatability analyses in ecoregions other than northern coastal California before they determine that combined data sets are appropriate for their program-specific needs.

←

FIG. 10. Estimated cumulative distributions of benthic index of biotic integrity (B-IBI) scores (A), observed-to-expected (O/E) index of taxonomic completeness O/E_0 (B), and O/E_{50} (C) scores in perennial wadeable streams in California. Subscripts on O/E ratios indicate site-specific probabilities of capture >0 or ≥ 0.5 (O/E_0 and O/E_{50}, respectively). Biological indicators were calculated from both targeted-riffle (TR) and reach-wide (RW) sample types. 95% confidence intervals of the TR curves are shown for comparison.

Acknowledgements

The US EPA Environmental Monitoring and Assessment Program funded statewide data collection under Assistance Agreement CR82823801. The views expressed in this paper are those of the authors and do not represent the views of the US Environmental Protection Agency. The California Energy Commission funded our replication study under project #500-03-017. CPH's contributions to this study were supported, in part, by EPA Science to Achieve Results (STAR) grant R-82863701. Region 5 of the US Forest Service provided funding and conducted field work that supported most of the O/E index development. Thanks go to Tony Olsen (US EPA) for help with condition assessments. Alan Herlihy (Oregon State University) and 2 anonymous referees provided helpful reviews of earlier drafts of the manuscript. Thanks go to Jason May (US Geological Survey) for help with literature searches, and to our field crew and taxonomists (Mike Dawson, Shawn McBride, Dan Pickard, Doug Post, Brady Richards, Glenn Sibbald, Joe Slusark, and Jennifer York).

Literature Cited

BARBOUR, M. J., J. GERRITSEN, G. E. GRIFFITH, R. FRYDENBORG, E. MCCARRON, J. S. WHITE, AND M. L. BASTIAN. 1996. A framework for biological criteria for Florida streams using benthic macroinvertebrates. Journal of the North American Benthological Society 15:185–211.

BLACKBURN, T. M., J. H. LAWTON, AND J. N. PERRY. 1992. A method of estimating the slope of upper bounds of plots of body size and abundance in natural animal assemblages. Oikos 65:107–112.

BLOCKSOM, K. A., AND J. E. FLOTEMERSCH. 2005. Comparison of macroinvertebrate sampling methods for nonwadeable streams. Environmental Monitoring and Assessment 102: 243–262.

BONADA, N., N. PRAT, V. H. RESH, AND B. STATZNER. 2006. Developments in aquatic insect biomonitoring: a comparative analysis of recent approaches. Annual Review of Entomology 51:495–523.

CAO, Y., AND C. P. HAWKINS. 2006. Comparability and integration of data used in aquatic bioassessments: a critical review of definitions, approaches, and methods. Technical report to the Office of Water and Watersheds, US Environmental Protection Agency (Available from; Internet URL or agency name, complete postal address, city, state, zip code USA.) ?1

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

CARTER, J. L., AND V. H. RESH. 2001. After site selection and before data analysis: sampling, sorting, and laboratory procedures used in stream benthic macroinvertebrate monitoring programs by USA state agencies. Journal of the North American Benthological Society 20:658–682.

CHESSMAN, B. C. 1995. Rapid assessment of rivers using macroinvertebrates: a procedure based on habitat-specific sampling, family-level identification and a biotic index. Australian Journal of Ecology 20:122–129.

CHUTTER, F. M. 1972. An empirical biotic index of the water quality in South African streams and rivers. Water Research 6:19–30.

DE PAUW, N., P. F. GHETTI, P. MANZINI, AND S. SPAGGIARI. 1992. Biological assessment methods for running waters. Pages 217–249 *in* P. J. Newman, M. A. Piavaux, and R. A. Sweeting (editors). River water quality ecological assessment and control. Commission of the European Communities, Bruxelles, Belgium.

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

FORE, L. S., K. PAULSEN, AND K. O'LAUGHLIN. 2001. Assessing the performance of volunteers in monitoring streams. Freshwater Biology 46:109–123.

GERTH, W. J., AND A. T. HERLIHY. 2006. The effect of sampling different habitat types in regional macroinvertebrate bioassessment surveys. Journal of the North American Benthological Society 25:501–512.

HAWKINS, C. P. 2006. Quantifying biological integrity by taxonomic completeness: evaluation of a potential indicator for use in regional- and global-scale assessments. Ecological Applications. In press. ?2

HERLIHY, A. T., D. P. LARSEN, S. G. PAULSEN, N. S. URQUHART, AND B. J. ROSENBAUM. 2000. Designing a spatially balanced, randomized site selection process for regional stream surveys: the EMAP Mid-Atlantic pilot study. Environmental Monitoring and Assessment 63:95–113.

HEWLETT, R. 2000. Implications of taxonomic resolution and sample habitat for stream classification at a broad geographic scale. Journal of the North American Benthological Society 19:352–361.

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.

HUGHES, R. M. 1994. Defining acceptable biological status by comparing with reference conditions. Pages 31–47 *in* W. S. Davis and T. P. Simon (editors). Biological assessment and criteria: tools for water resource planning and decision making. Lewis Press, Boca Raton, Florida.

KERANS, B. L., J. R. KARR, AND S. A. AHLSTEDT. 1992. Aquatic invertebrate assemblages: spatial and temporal differences among sampling protocols. Journal of the North American Benthological Society 11:377–390.

KLEMM, D. J., K. A. BLOCKSOM, F. A. FULK, A. T. HERLIHY, R. M. HUGHES, P. R. KAUFMANN, D. V. PECK, J. L. STODDARD, W. T. THOENY, AND M. B. GRIFFITH. 2003. Development and evaluation of a macroinvertebrate biotic integrity index (MBII) for regionally assessing Mid-Atlantic highland streams. Environmental Management 31:656–669.

KLEMM, D. J., AND J. M. LAZORCHAK. 1994. Environmental monitoring and assessment program, surface water and Region 3 regional monitoring and assessment program, 1994 pilot laboratory methods manual for streams. EPA/62/R-94/003. Office of Research and Development, US Environmental Protection Agency, Washington, DC.

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.

MCCULLOCH, D. L. 1986. Benthic macroinvertebrate distributions in the riffle-pool communities of two east Texas streams. Hydrobiologia 135:61–70.

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

ODE, P. R., A. C. REHN, AND J. T. MAY. 2005. A quantitative tool for assessing the integrity of southern coastal California streams. Environmental Management 35:493–504.

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

PARSONS, M., AND R. H. NORRIS. 1996. The effect of habitat-specific sampling on biological assessment of water quality using a predictive model. Freshwater Biology 36: 419–434.

PECK, D. V., J. M. LAZORCHAK, AND D. J. KLEMM, (EDITORS). 2004. Environmental monitoring and assessment program—surface waters: western pilot study field operations manual for wadeable streams. Office of Research and Development, US Environmental Protection Agency, Corvallis, Oregon. In press. ?3

REHN, A. C., P. R. ODE, AND J. T. MAY. 2005. Development of a benthic index of biotic integrity (B-IBI) for wadeable streams in northern coastal California and its application to regional 305(b) reporting. Unpublished technical report for the California State Water Quality Control Board, Sacramento, California. (Available from; http://www.swrcb.ca.gov/swamp/docs/northc1.pdf)

SIMPSON, J. C., AND R. H. NORRIS. 2000. Biological assessment of river quality: development of AUSRIVAS models and outputs. Pages 125–142 *in* J. F. Wright, D. W. Sutcliffe, and M. T. Furse (editors). Assessing the biological quality of fresh waters: RIVPACS and other techniques. Freshwater Biological Association, Ambleside, UK.

STARK, J. D. 1993. Performance of the Macroinvertebrate Community Index: effects of sampling method, sample replication, water depth, current velocity, and substra-

tum size on index values. New Zealand Journal of Marine and Freshwater Research 27:463–478.

STEVENS, D. L., AND A. R. OLSEN. 2004. Spatially balanced sampling of natural resources. Journal of the American Statistical Association 99:262–278.

STODDARD, J. L., D. V. PECK, A. R. OLSEN, D. P. LARSEN, J. VAN SICKLE, C. P. HAWKINS, R. M. HUGHES, T. R. WHITTIER, G. LOMNICKY, A. T. HERLIHY, P. R. KAUFMANN, S. A. PETERSON, P. L. RINGOLD, S. G. PAULSEN, AND R. BLAIR. 2005. Environmental Monitoring and Assessment Program (EMAP): western streams and rivers statistical summary. EPA620/R-05/006. Office of Research and Development, US Environmental Protection Agency, Washington, DC.

SUDARYANTI, S., Y. TRIHADININGRUN, B. T. HART, P. DAVIES, C. HUMPHREY, R. NORRIS, J. SIMPSON, AND L. THURTELL. 2001. Assessment of the health of the Brantas River, East Java, Indonesia using the Australian River Bioassessment Method (AUSRIVAS). Aquatic Ecology 35:135–146.

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

WAITE, I. R., A. T. HERLIHY, D. P. LARSEN, N. S. URQUHART, AND D. J. KLEMM. 2004. The effect of macroinvertebrate taxonomic resolution in large landscape bioassessments: an example from the Mid-Atlantic Highlands, U.S.A. Freshwater Biology 49:474–489.

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

Received: 8 June 2006
Accepted: 19 October 2006