Comparability of biological assessments derived from predictive models and multimetric indices of increasing geographic scope

Peter R. Ode[1], Charles P. Hawkins[2] and Raphael D. Mazor

Abstract

As the use of bioassessment techniques expands, the demand for tools that can score biological condition from aquatic community data has spurred the creation of a large number of predictive models (e.g., observed over expected (O/E) indices) and multimetric indices (MMIs). The geographic and environmental scopes of these indices vary widely and coverages often overlap. If indices developed for large, environmentally heterogeneous regions provide results that are equivalent to those developed for smaller regions, then regulatory entities could adopt indices developed for larger regions rather than fund the development of multiple local indices. This potential was evaluated by comparing the performance (precision, bias, responsiveness, and sensitivity) of benthic macroinvertebrate O/E and MMIs developed for California (CA) with indices developed for two large-scale condition assessments of United States (US) streams: the US Environmental Protection Agency's Western Environmental Monitoring and Assessment Program's (WEMAP) stream project and the western portion of the national Wadeable Streams Assessment (WSA-West). Both WSA-West and WEMAP O/E scores were weakly correlated with CA O/E index scores, had lower precision than the CA index, were influenced by two related natural gradients (percent slope and percent fast water habitat) for which the CA index was not, and disagreed with 21 - 22% of impairment decisions derived from the CA index. The WSA-West O/E index produced many fewer impairment decisions than the CA index. In the MMI comparisons, both WEMAP and WSA-West MMI scores were much more strongly associated with CA MMI scores than those found in the O/E comparisons. However, the WSA-West and WEMAP MMIs produced many fewer impairment determinations than the CA MMI. Because the WEMAP and WSA-West indices were biased and differed in responsiveness compared with CA indices, they could produce different estimates of regional condition compared with indices that are calibrated to local conditions. Furthermore, the lower precision of the WEMAP and WSA-West indices compromises their use in site-specific assessments where both precision and accuracy are important. However, because the magnitude of differences in impairment decisions was very sensitive to the thresholds used to define impaired conditions, it may be possible to adjust for some of the systematic differences among the models, thus rendering the larger models more suitable for local application. Future work should focus on identifying the geographic and environmental scale that optimizes index performance, determining the factors that most strongly influence index performance, and identifying ways of more accurately specifying reference condition from geographically extensive sets of reference site data.

Introduction

The widespread adoption of bioassessment techniques for assessing the ecological condition of waterbodies has generated an abundance of indices available to water resource managers (Reynoldson *et al*. 1997, Hughes *et al*. 1998, Barbour and Yoder

[1] *California Department of Fish and Game, Water Pollution Control Laboratory, Aquatic Bioassessment Laboratory, Rancho Cordova, CA*
[2] *Utah State University, Department of Watershed Sciences, Western Center for Monitoring and Assessment of Freshwater Ecosystems, Logan, UT*

2000, Hawkins *et al* 2000a, Van Sickle *et al*. 2005, Bonada *et al*. 2006). Because these tools were generated to meet different needs, their geographic scopes differ widely and often overlap.

As the proliferation of new indices continues, end-users (e.g., regulatory entities developing numeric biocriteria, Yoder and Rankin 1995) will need guidance for selecting among these different indices and evaluating how many different indices a region needs for effective bioassessment. If local and regional assessments based on indices developed for broad geographical areas are equivalent to assessments based on indices developed for smaller areas, then regulatory entities could profit by adopting the large-scale indices and abandoning the development and maintenance of multiple, smaller-scale indices. This potential is attractive because indices that apply to large geographic areas have already been developed for many regions of the world, including: Great Britain (Moss *et al*. 1987), Australia (Simpson and Norris 2000), Europe (Statzner *et al*. 2001), and the United States (Stoddard *et al*. 2006, 2008; Yuan *et al*. 2008). Widespread use of common indices would facilitate consistency in data interpretation among the variety of users of ecological condition indices (Bonada *et al*. 2006, Hawkins 2006).

However, indices developed for large geographic regions may have limitations that could restrict their value for both site and regional assessments. Most notably, such indices must account for natural variation that occurs within large regions. Performance characteristics of both multimetric and predictive model indices are limited by their capacity to account for variability among the reference sites used to develop indices (Moss *et al*. 1987, Hughes 1995, Reynoldson *et al*. 1997, Karr and Chu 1999, Hawkins *et al*. 2000a, Bailey *et al*. 2004, Bonada *et al*. 2006).

It is a central principle of ecology that biological assemblages naturally vary along many environmental gradients (Andrewartha and Birch 1954, Hutchinson 1959, Hynes 1970). The precision and accuracy of any index will therefore depend on how well the mechanics of index calculation account for the effects of these natural gradients on assemblage structure (Johnson *et al*. 2004, Johnson *et al*. 2007, Van Sickle *et al*. 2005, Hawkins 2006, Heino *et al*. 2007, Mykrä *et al*. 2007, 2008). If biological variation associated with local environmental gradients (e.g., reach slope or substrate size) is masked by environmental factors that vary over large spatial scales (e.g., climatic factors and geology), then indices developed from more spatially restricted datasets may be required for site-specific assessments.

Recently derived biological indices developed for the EPA's national WSA and the WEMAP project (Stoddard *et al*. 2005, 2006; EPA 2006) presented an opportunity to evaluate this idea by comparing performance metrics (precision, bias, responsiveness, and sensitivity) of these indices with those of indices developed specifically for California (Ode *et al*. 2005, Rehn *et al*. 2005). The comparability of both site-specific and regionally aggregated biological assessments, where CA indices <WEMAP indices <WSA-West indices in geographic extent and geoclimatic heterogeneity, were evaluated. For these comparisons, assessments of an independent set of evaluation (test) sites that had not been used in developing any of the indices were conducted. To the extent that the test dataset permitted, parallel analyses for both MMI and O/E indices of benthic macroinvertebrate (BMI) assemblage condition were performed.

METHODS

O/E Development

Three sets of predictive models were used to produce the O/E index values for comparison. All the O/E models were developed following a standardized process (Clarke *et al*. 2003, Hawkins *et al*. 2000a, Moss *et al*. 1987) described in the EMAP Western Streams and Rivers Statistical Summary (Stoddard *et al*. 2006). The process included: 1) sampling a set of environmentally diverse sites for BMIs, 2) specifying which of these sites would be used as reference sites, 3) applying a standard taxonomy (operational taxonomic units; OTUs) to all samples, 4) clustering of reference sites according to their similarity in BMI assemblage composition, 5) calculating and screening candidate predictor variables, and 6) calibrating linear discriminant functions models for predicting assemblage composition at new sites. All models were developed with map-level predictor variables (with the exception that field measured reach slope was used in one model) to allow more universal applicability of models (Table 1). Aside from the specific combination of predictor variables used in the models, the major difference among models was the range of environmental heterogeneity or geographic extent encompassed by the reference sites used in each model. Models were based on data from either targeted-riffle benthic samples (CA models) or a combination of targeted-

Table 1. Predictor variables used for all predictive models.

	California Models	WEMAP Models	WSA Model (no sub-models)
Sub-model 1			
	Watershed area Longitude Latitude Temperature	No predictors (null models)	Watershed area Longitude Day of year Minimum temperature Elevation Precipitation Slope
Sub-model 2			
	Longitude Preipitation Day of year Watershed area	Watershed area Longitude Elevation Precipitation	
Sub-model 3			
	Watershed area Temperature	No predictors (null models)	

riffle and reach-wide, multiple-habitat samples (WEMAP and WSA-West models). These two types of samples appear to be generally comparable for CA streams (Rehn *et al*. 2007). Other aspects of model development were similar (Table 2).

WSA-West model

A single western US model (WSA-West) developed during the national wadeable streams assessment (Yuan *et al*. 2008) encompassed the most heterogeneous environmental conditions and the largest geographic scope (~2,500,000 km^2; Figure 1). The WSA-West model was developed for all mountainous and xeric regions of the western United States and excluded only plains ecoregions (Figure 1; see Environmental Protection Agency 2006). To produce the WSA-West O/E index, 519 reference sites were clustered into 31 groups, and 7 predictor variables were selected to predict group membership (Table 1).

WEMAP models

The same data used to construct the WSA-West model had been previously used to develop five separate ecotype-specific submodels (Stoddard *et al*. 2006, 2008). All sampled sites (reference and non-reference) were assigned to one out of five broad ecotypes based on a k-means classification (MacQueen 1967) of long-term climatic (temperature and precipitation), geographic variables (latitude, longitude and elevation), and topographic variables (watershed area and channel slope). This pre-classification of sites was mainly designed to reduce the range of environmental heterogeneity encompassed by each model. The geographic scope of the resulting submodels ranged from ~200,000 km^2 to ~1,800,000 km^2 (Figure 2). Of the five submodels developed for the WEMAP study area (Stoddard *et al*. 2005, 2006), four submodels applied to geoclimatic conditions found in California. One model used predictor variables, whereas the other three were null models that predicted the same biota at all

Table 2. Comparison of BMI collection method, taxonomic effort levels and organism counts used both to build models and score test sites. See methods for definitions.

Indicator	Model	Field Method	Taxonomic Effort	Organism Count
O/E	WEMAP	RWB	Some species, but mostly genus (including Chironomidae)	300 (after removal of ambiguous individuals)
	WSA	RWB		
	2 CA sub-models	TRB		
MMI	WEMAP	RWB	Some species, but mostly genus (including Chironomidae)	300
	WSA	RWB	Some species, but mostly genus (including Chironomidae)	300
	CA model (NCIBI / SCIBI)	TRB	Genus, Chironomidae to family	500

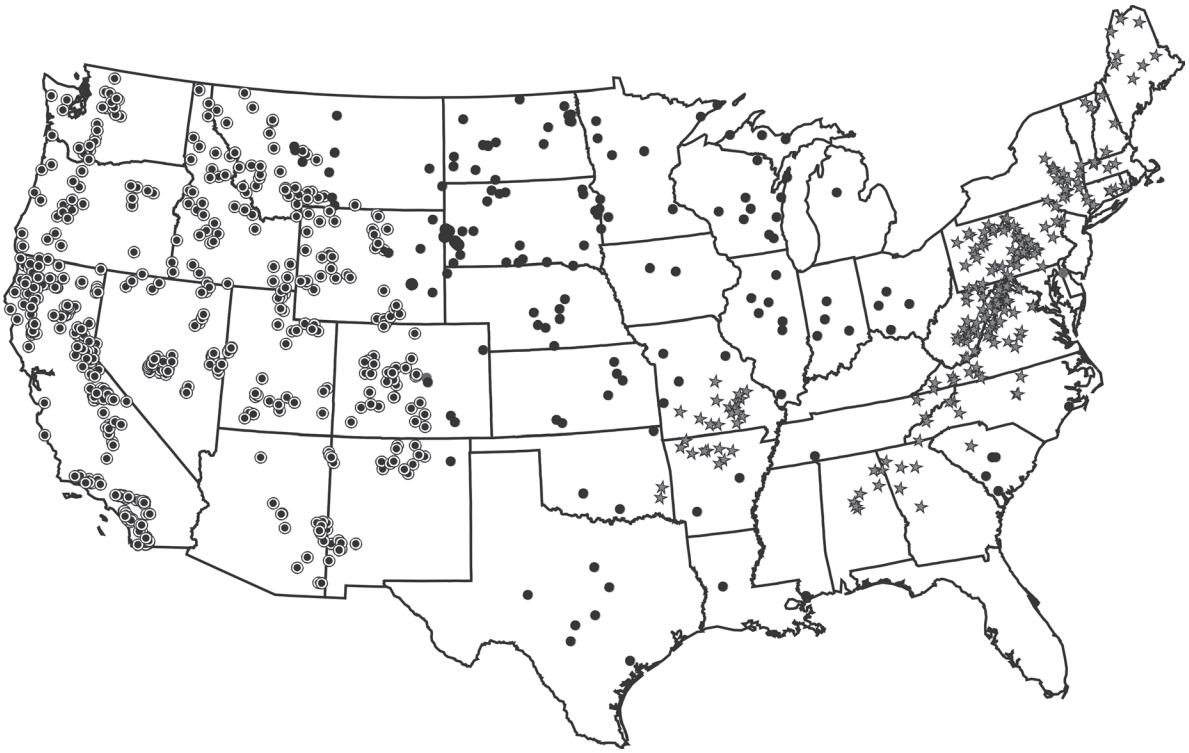

Figure 1. Location of reference sites used to create the three WSA predictive sub-models. Only the western sub-model applies to California sites. Each symbol represents a different sub-model.

sites within a geoclimatic region (Van Sickle *et al*. 2005; Table 1).

CA models

The third model set included three submodels that were developed for three types of climatic conditions in CA: cool-wet sites (mean monthly temperature (MMT) >9.9ºC and mean monthly precipitation (MMP) >895 mm), warm-dry sites (MMT >9.9ºC and MMP <895 mm), and cold-mesic sites (MMT <9.9ºC; Figure 3). The three CA submodels were calibrated from data collected at 209 reference sites, 179 of which were also used in calibrating WEMAP and WSA-West models (the other 30 sites were used as validation samples in the WEMAP and WSA projects). The spatial extent of the reference sites for these submodels was ~150,000 km^2 each (Figure 3). These three submodels also used unique combinations of predictor variables (Table 1).

MMI Development

The WSA, WEMAP, and CA MMIs were developed following similar methods as first developed by Karr (1981) and extended by others (Kerans and Karr 1994, Hughes *et al*. 1998, McCormick *et al*. 2001, Klemm *et al*. 2003): 1) assignment of a large pool of sites to either reference or test sets based on their degree of anthropogenic stress, 2) division of the site pool into calibration and validation sets, 3) using the calibration set to screen biological response metrics based on their responsiveness to important stressor gradients, their signal-to-noise ratios, and their non-redundancy with other metrics,

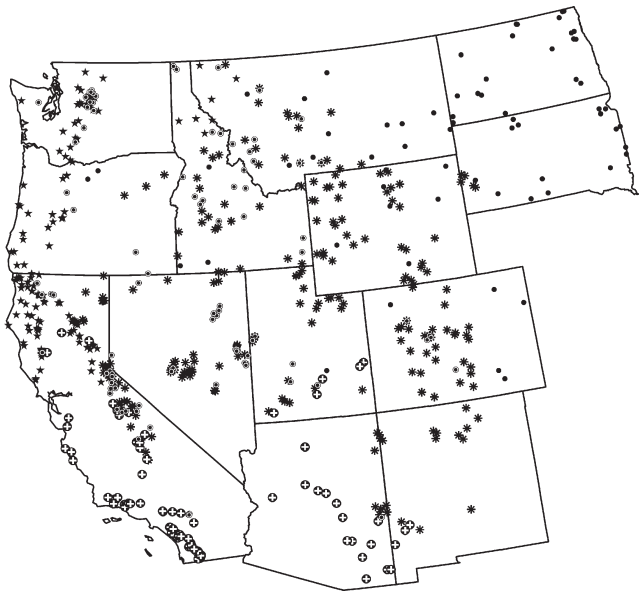

Figure 2. Location of reference sites used to create the five WEMAP predictive sub-models. Note that four of the five sub-models apply to California.

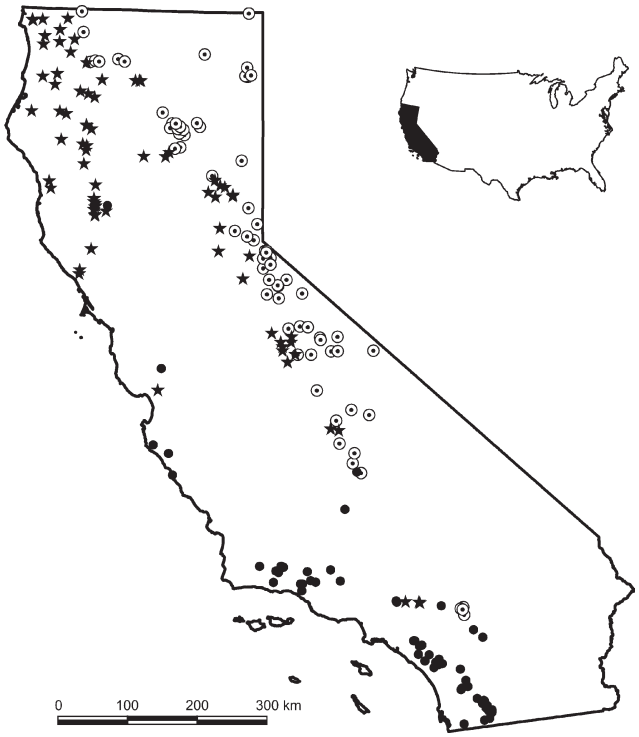

Figure 3. Location of reference sites used to create the three CA predictive sub-models. Each symbol represents a different sub-model.

4) establishing scoring ranges for selected metrics, 5) assembling a composite MMI from the component metrics, 6) establishing impairment thresholds for the index, and 7) evaluating index performance against the validation dataset (Herlihy *et al*. 2008, Stoddard *et al*. 2008).

These MMIs differed in a few important respects (Tables 2 and 3). The CA indices were based on subsamples of 500 organisms collected from targeted-riffle habitats (TRB) and identified primarily to genus level, but the WSA-West and WEMAP indices were based on subsamples of 300 organisms collected from multiple reach-wide composite habitats (RWB) with some individuals identified to species level (see text below for details on field and lab methods).

WSA-West MMIs

The EPA's WSA program developed two MMIs (xeric and western mountain ecoregions; Omernik 1987) to support its assessments of western streams using a calibration dataset of 775 sites (235 xeric and 540 mountain; Stoddard *et al*. 2008, EPA 2006). Both indices used six metrics, five of which were in common (Table 3). Scoring ranges for both WSA-West MMIs were scaled from 0 to 100 (Van Sickle and Paulsen 2008).

WEMAP MMIs

WEMAP developed three MMIs (xeric, plains and mountain ecoregions) for its analyses (Stoddard *et al*. 2005, 2006), two of which (xeric and mountain) applied to CA sites. The calibration dataset was comprised of 244 xeric and 565 mountain sites, nearly all of which (754 of 809) were used in WSA-West MMI development. As in the WSA-West index, the xeric and mountain versions of the WEMAP MMI consisted of six metrics, but shared fewer metrics in common (Table 3). Index values for both WEMAP MMIs were scaled from 0 to 100 (Stoddard *et al*. 2005).

CA MMIs

Two MMIs were developed for use in coastal California: the Southern Coastal California Index of Biotic Integrity or SCIBI (Ode *et al*. 2005) and the Northern Coastal California Index of Biotic Integrity or NCIBI (Rehn *et al*. 2005). The two CA MMIs included both the mountain and xeric aggregate ecoregions used for the WSA and WEMAP MMIs, and separate metric scoring ranges were established for the Omernik Level III (1987) ecoregions within each CA MMI development area (Figure 4). Of the 502 sites used to develop the CA MMIs, 119 were also used in WEMAP and WSA-West MMI development. The NCIBI consisted of eight metrics, whereas the SCIBI consisted of seven metrics, with four metrics in common (Table 3). The CA MMIs were also scaled from 0 to 100 (Ode *et al*. 2005, Rehn *et al*. 2005).

Test Site Data

These analyses incorporate BMI data collected for two large-scale probability surveys of CA streams. For clarity, use of the term "test sites" was restricted to refer only to these probabilistic samples of evaluation sites and not to non-reference sites used to calibrate MMIs, which are sometimes referred to as "test sites" in MMI development. For the O/E comparisons, data collected from 127 sites during the WEMAP 2000-2003 survey were used. For the MMI comparisons, data from 68 sites sampled by the California State Monitoring and Assessment Program (CMAP) between 2004 and 2006 were used. It was necessary to use different test sets for the O/E and MMI analyses because: 1) the restricted geographic boundaries of the CA MMI models limited the number of sites shared between O/E and MMI data sets, and 2) the MMI calibration datasets were

Table 3. BMI metrics comprising the multimetric indices. EPT = Ephemeroptera, Plecoptera, and Trichoptera.

Metric	California		WEMAP		WSA	
	NCIBI	SCIBI	Mountain	Xeric	Mountain	Xeric
EPT Richness	X	X	X	X	X	X
% Individuals in Top 5 Taxa			X		X	X
% Non-Insect Taxa	X	X		X		
Clinger % Taxa				X	X	X
% Intolerant Individuals	X	X				
% Non-Insect Individuals			X			X
% Tolerant Taxa		X	X			
Coleoptera Richness	X	X				
Scraper Richness					X	X
Tolerant % Taxa					X	X
% Burrower Individuals			X			
% Collector Individuals		X				
% EPT Taxa					X	
% Intolerant Taxa				X		
% Non-Gastropod Scraper Individuals	X					
% Omnivore Taxa			X			
% Predator Individuals	X					
% Shredder Taxa	X					
Diptera Richness	X					
Predator Richness		X				
Shannon Diversity				X		
Shredder Richness				X		

partially comprised of sites used for the O/E test set. The 127 sites used to evaluate predictive models were distributed throughout California (Figure 4a), whereas the 68 sites used to evaluate MMI models were restricted to coastal watersheds (Figure 4b). Most MMI test sites were concentrated in the northern half of the state (61 sites north of Monterey Bay), and the majority of these sites (40) were located within the boundaries of the NCIBI calibration sites (Figure 4b). The remaining 21 northern California sites were concentrated in the San Francisco Bay and Santa Cruz Mountains regions, which lie between the development regions of the two CA MMIs (Figure 4b). We used the NCIBI to score sites located between the NCIBI and SCIBI regions for the cross-index comparisons because this region is ecologically more similar to the North Coast than the South Coast and because reference conditions for this area were better represented in the NCIBI (Rehn *et al*. 2005). SCIBI scores were used for another 14 sites located within the region defined by the SCIBI calibration sites. Although the different geographic distributions in test sites may affect comparisons between MMIs and O/E indices, they do not affect comparisons of the performance of each type of index among the three geoclimatic scales.

Test site, field, and laboratory methods

All test sites were sampled in accordance with standard WEMAP field methods (Peck *et al*. 2006). A sampling reach was defined as 40 times the average stream width at the center of the reach, with a minimum reach length of 150 m. Two BMI samples were collected from each reach with standard 500-µm D-frame nets: 1) a RWB sample consisting of eleven 0.09-m^2 samples taken from equally spaced locations throughout the reach and 2) a TRB sample consisting of eight 0.09-m^2 samples taken from fast water habitat units within the reach (Hawkins *et al*. 2003).

All BMI samples used for the test datasets were processed at the California Department of Fish and Game's Aquatic Bioassessment Laboratory in Chico, CA. At least 500 individuals were identified to the standard taxonomic resolution targets described in Richards and Rogers (2006), i.e., those levels of taxonomic resolution that can be consistently achieved. A true, fixed 500-count random subsample was then

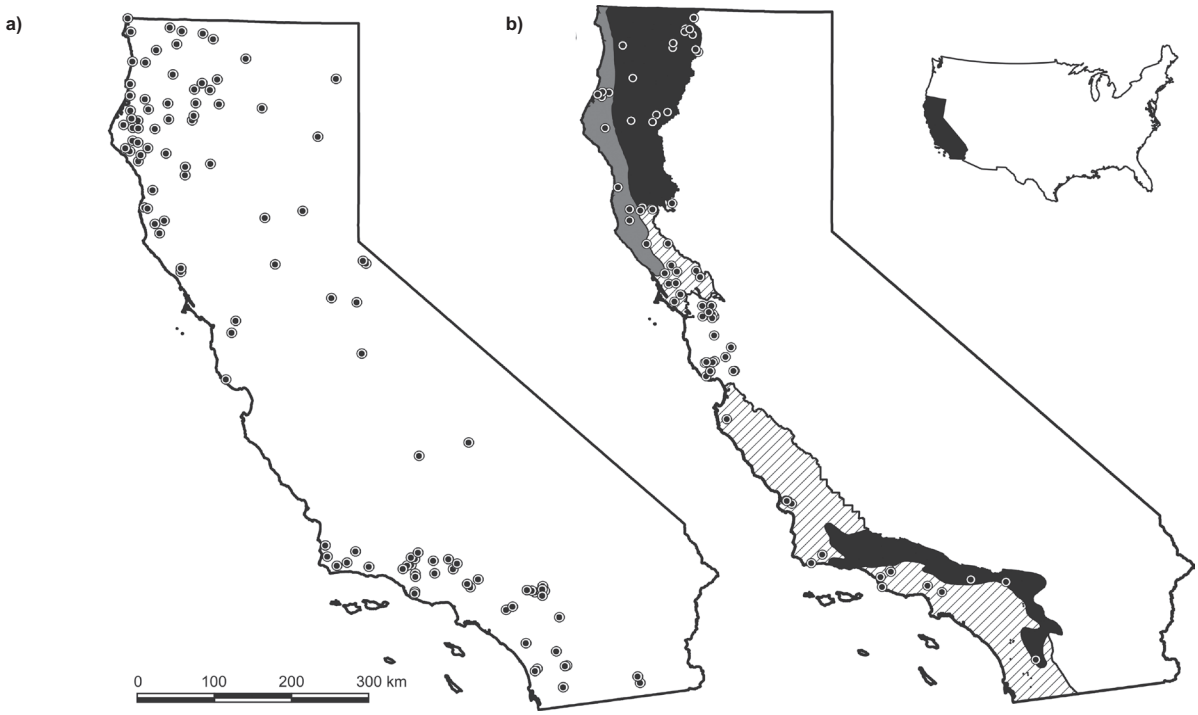

Figure 4. Location of test sites used for comparative analyses: 127 test sites used in O/E comparisons (a) and 68 test sites used in IBI comparisons (b). The three solid shaded regions correspond to mountain ecoregions used in the western and national models, whereas the three hatched regions correspond to the xeric ecoregions used in the western and national models. The two solid shaded regions in the northwest part of the state circumscribe the region for which the North Coast IBI was developed. The hatched regions and the continuously shaded region in southwestern part of the state circumscribe the region for which the South Coast IBI was developed. The inset shows the location of California in the United States.

obtained by computer resampling the sample data. Samples with between 450 and 500 individuals were retained in analyses. These raw data were then used to produce the standardized taxa lists and metrics needed for the various indices (Table 3). All analyses were based on the field methods, sample sizes and taxonomic levels used to develop each index (as indicated in Table 2).

Scoring Sites: Predictive Models

BMI taxonomic data

The raw subsample count data were further processed for use with the predictive models by: 1) converting the original identifications to the taxonomic levels used in the models (e.g., OTUs), 2) eliminating individuals that could not be assigned to an OTU (i.e., ambiguous individuals), and 3) resampling the remaining non-ambiguous individuals to 300-count samples. Samples with <300 individuals were retained in analyses.

Predictor variables

Geographic coordinates (latitude and longitude) were obtained via GPS measurements taken during sample collection. Watershed area were calculated after delineating upstream watershed boundaries for each site with automated GIS scripts or manual delineation where necessary. Long-term MMP, MMT, and MMA values for each site were estimated from GIS grids of (1961-1990) obtained from the Oregon Climate Center (http://www.ocs.orst.edu/prism). Site elevations were derived from 30-meter digital elevation models (http://ned.usgs.gov). Channel (reach) slope was measured in the field (as it was in model development).

Geographic and environmental attributes were used to assign each site to the appropriate WEMAP and CA models. Assignment of sites to the five WEMAP models was based on latitude, longitude, elevation, MMP, MMT, watershed area, and channel slope. These assignments were made prior to model building during the k-means analysis (MacQueen 1967). Assignment of test sites to the appropriate CA model was conducted after model development. This study used a simple classification and regression tree model based on long-term

precipitation and air temperature to assign sites to the CA submodels.

O/E values were calculated based on just those taxa with site probabilities of capture ≥0.5 because these values result in more precise O/E values that are also usually more sensitive to stress (Hawkins *et al*. 2000a, Ostermiller and Hawkins 2004, Van Sickle *et al*. 2007) than O/E values based on all taxa in the reference calibration data set. When reporting impairment decisions for the test sites, impairment thresholds were set at two standard deviations below the mean value of reference sites for all O/E models (Table 4).

Scoring Sites: MMIs

BMI taxonomic data

Because the MMIs differed with respect to organism count and taxonomic resolution, MMI scores were calculated based on the sample counts and taxonomy used when developing each index (Table 2). MMI values were then calculated for test samples that had been collected in a standard manner to avoid confounding comparisons with inter-method variability. All sites were assigned to either the xeric or mountain aggregate ecoregions, with mountain ecoregions being further divided into Southern California Mountains, Klamath Mountains, Coast Ranges, and Southern and Central California Chaparral and Oak Woodlands for the CA MMIs (Omernik 1987). MMI values were then calculated based on the specific scoring ranges developed for each individual metric and region and rescaled these MMI values from 0 to 100. As for O/E models, impairment thresholds for all MMIs were set at two standard deviations below the mean value at reference sites (Table 4).

TRB was used as the default sample type, although RWB samples were used at six sites where TRB samples were unavailable or had low sample counts (<450 organisms). Because it was found elsewhere that RWB samples on average scored 7.8 points lower on the CA IBIs than TRB samples (Rehn *et al*. 2007), 7.8 points were added to CA IBI scores for these RWB samples. To evaluate the potential effect of using TRB samples instead of RWB samples (the method used in national and western model development; Table 2) in comparisons, an additional analysis was performed in which both the TRB and RWB data from 21 sites with all three MMIs were scored. If paired t-tests indicated significant differences between methods, RWB scores were adjusted by a correction factor corresponding to the difference between mean site scores.

Comparison of Index Scores

The CA index values were used as a benchmark for comparing the performance of the WSA-West and WEMAP indices. Comparisons were based on index precision, bias, responsiveness, and sensitivity.

O/E comparisons

Precision was measured as the standard deviation (sd) of reference site O/E values. Bias was measured as the tendency for reference site O/E values to vary systematically with one or more of four

Table 4. Standard deviation (sd) values and impairment thresholds (IT) for each predictive model (O/E) and coefficients of variation (CV) and impairment thresholds (IT) for MMIs. Note that only WEMAP sub-models 2 through 5 apply to California.

California O/E			WEMAP O/E			WSA O/E		
Sub-model	**sd**	**IT**	**Sub-model**	**sd**	**IT**	**Model**	**sd**	**IT**
1	0.13	0.74	1	0.24	0.52	West	0.20	0.59
2	0.17	0.66	2	0.15	0.70			
3	0.16	0.68	3	0.20	0.60			
			4	0.20	0.60			
			5	0.17	0.66			

California MMI			WEMAP MMI			WSA MMI		
Sub-model	**CV**	**IT**	**Sub-model**	**CV**	**IT**	**Sub-model**	**CV**	**IT**
NCIBI	0.14	52	Mountain	0.13	55	Mountain	0.26	28
SCIBI	0.19	39	Xeric	0.23	36	Xeric	0.25	34

natural gradients (reach slope, elevation, watershed area, and percent of reach with fast water habitats). The study also assessed relative bias between pairs of O/E indices using linear regression; slopes were tested for significant differences from 1, and intercepts were were tested for significant differences from 0. The consequences of these types of biases were illustrated by plotting the pair-wise differences in index scores against these natural gradients. Responsiveness was measured as the mean difference between reference and test sites in O/E values. Sensitivity was measured as the proportion of test sites assessed as impaired by the models. This measure of sensitivity is a joint function of precision, bias, and responsiveness. For these assessments, the threshold values for inferring impairment were defined as 2 sds below the reference (calibration) means (Table 4). Binomial tests (Zar 1999) were used on sites with disagreeing impairment decisions to determine if the indices were equally likely to detect impairment. This test was performed within each of the three CA submodels, as well as on all sites combined. In addition to comparison of impairment determinations based on 2 sds thresholds, two different threshold corrections for ecoregional differences were also evaluated. In the WSA, impairment thresholds were established separately for xeric and mountain ecoregions at the 5th percentile of the calibration reference population (estimated as 1.64 standard deviations below the reference mean; Herlihy *et al*. 2008). We also estimated separate thresholds for mountain and xeric regions at 2 sd below the mean for each ecoregion, an approach consistent with previous comparisons. For all relevant analyses, Bonferroni adjustments were applied for multiple comparisons when the correction was conservative. That is, the correction was not applied when we were screening natural gradients as potential drivers of bias, but was applied for hypothesis tests of index agreement (e.g., impairment decisions, responsiveness tests).

Multimetric index comparisons

MMI analyses paralleled the O/E comparisons. However, raw MMI scores were not directly comparable because the scores at calibration reference sites differed among the MMIs. Therefore, MMI scores were rescaled by dividing the raw score by the index's reference mean. These adjusted scores were then used as a "common currency" in all analyses in which scores were compared directly. Thus, the MMI scaling in these analyses was similar to the ~1.0 reference mean in O/Es. Only the comparisons of impairment decisions were based directly on the raw MMI scores.

Results

O/E Comparisons

Precision

The predictions of the WSA-West and WEMAP models were less precise (reference site O/E sd = 0.17 to 0.20) than those of the CA models (sd = 0.13 to 0.17; Table 4). Imprecision in model predictions contributed, in part, to weak relationships between the CA O/E indices and the WSA-West and WEMAP O/E indices (CA vs. WSA-West $r^2 = 0.32$, CA vs. WEMAP $r^2 = 0.35$; Figure 5). However, the stronger agreement between the less precise WSA-West and WEMAP O/E indices (WSA-West vs. WEMAP $r^2 = 0.58$) indicates that factors other than precision (e.g., bias) must also be affecting differences in agreement (Figure 5).

Bias

The WSA-West and WEMAP O/E values were biased predictors of the CA O/E values and each other, with slopes and y-intercepts significantly different ($p < 0.001$) than 1 and 0, respectively, for all comparisons (Figure 5). Differences were large, with slopes as low as 0.58 and intercepts as high as 0.36. These results showed that the nature of the bias was not constant across all sites. Instead, differences in index scores depended on the site-specific differences among models in how they either over- or under-estimated E (the expected number of predicted taxa) relative to one another. The reason that the O/E indices were biased predictors of one another occurred, at least in part, because the WSA-West and WEMAP models failed to adjust predictions of E for the effects of at least one natural gradient. This failure is illustrated by systematic variation in reference site O/E values produced by the WSA-West and WEMAP models across percent slope (WSA-West score = 0.025% slope + 0.80, $p = 0.001$; WEMAP score = 0.023% slope + 0.67, $p = 0.002$) and percent fast water habitat gradients (WSA-West score = 0.0051% fast water + 0.747, $p < 0.001$; WEMAP score = 0.0045% fast water + 0.63, $p < 0.001$). No such relationships were evident for CA O/E values (CA score = 0.0086% slope + 0.78, $p = 0.259$; CA score = 0.0016% fast water + 0.77, $p = 0.205$). The reason the CA O/E indices were unrelated to reach slope is probably related to the fact that, within CA, channel

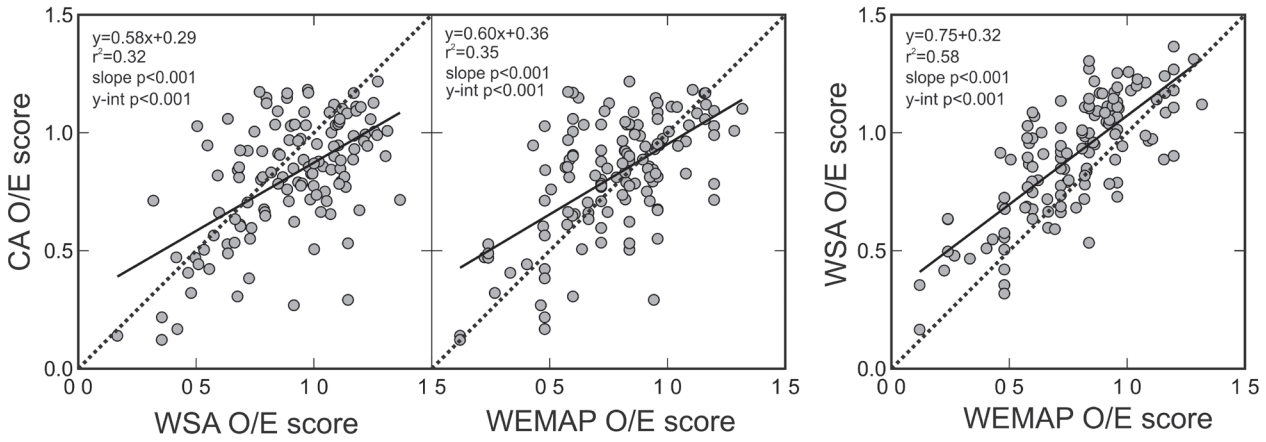

Figure 5. Regressions between O/E scores at CA test sites for all combinations of models. The dotted diagonal lines represent perfect 1:1 relationship between the models, and the thick solid lines indicate linear best-fit relationships. Significance tests are for y-intercept (y-int) = 0 and slope = 1.

slope was associated with watershed area (Area), a predictor in all three CA models (square root slope = 4.11 -0.531*log Area – 0.040*latitude across all reference sites, n = 209, r^2 = 0.14, model p <0.001). It is therefore possible that watershed area was a surrogate predictor of reach slope within CA. Percent fast water was measured at too few sites to determine its relationship with watershed area within CA. As a consequence of the bias between the WSA-West and WEMAP model predictions, pair-wise differences between O/E values for both the WSA-West and WEMAP indices and the CA indices were significantly related to channel slope and percent fast-water habitat (Figure 6). Similar biased predictions associated with either elevation or watershed area were not observed, nor were any of these relationships observed for pair-wise differences in values between WSA-West and WEMAP (Figure 6; Table 5). Furthermore, correlation coefficients were low for all of these relationships (Table 5), indicating that very little variance in differences between the indices was explained by these natural gradients. Although not related to the four natural gradients we examined, there was a tendency for the WSA-West model to produce higher O/E scores than the WEMAP submodels, especially at lower O/E values (p <0.005; Table 5; Figures 5 and 6).

Responsiveness

The WEMAP models tended to produce the lowest O/E values and the WSA-West models the highest O/E values at the test sites (Table 6). O/E values based on the CA models tended to be intermediate in magnitude. This pattern generally occurred for both mountain and xeric ecoregions, although differences were not always statistically significant. However, the magnitude of difference in mean test site O/E values between mountain and xeric test sites varied with the models used. The CA models resulted in lower average O/E values for xeric than for mountain sites (Table 6), whereas both the WEMAP and WSA-West models produced statistically similar mean O/E values at xeric and mountain test sites.

Index sensitivity and concordance among assessments

The WSA-West O/E was much less likely to lead to inferences of impairment (16 of 127 sites; Table 7) than either the WEMAP O/E (43 of 127 sites) or the CA O/E (35 of 127 sites, binomial tests, p <0.001). When an ecoregion correction based on 2 sds (consistent with primary analyses) was applied, there was no effect on any impairment decision (16 out of 127 sites impaired) because the separate xeric and mountain thresholds were within 2 points on a 100 point scale of their combined threshold. However, when an ecoregion correction based on the 5th percentile threshold used for the national wadeable streams assessment (Herlihy *et al*. 2008) was applied, the number of sites determined to be impaired by the WSA-West index (27 of 127 sites) was not significantly different from the 35 impairment decisions produced by the CA O/E index (binomial test, p = 0.081; Table 7).

Multimetric Index Comparisons: Comparison of TRB vs. RWB for WSA and WEMAP MMIs

MMI scores derived from the TRB and RWB sampling methods were highly correlated for both

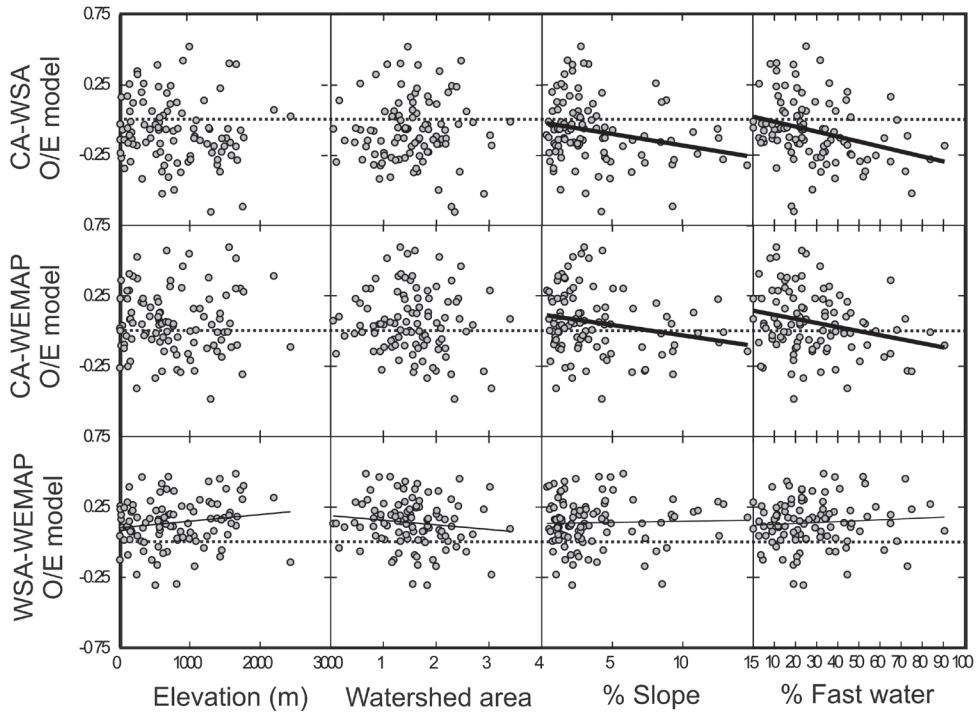

Figure 6. Scatterplots and regressions between the pair-wise differences in O/E values for three different O/E indices and four environmental gradients at CA test sites. The dashed horizontal lines represent zero difference. Thick solid lines denote regressions with r^2 and slopes significantly different from 0; thin solid lines denote those with intercepts significantly different from 0 but non-significant slope.

WSA and WEMAP indices (WSA $r^2 = 0.75$, WEMAP $r^2 = 0.73$), as has been shown elsewhere for CA MMIs (Rehn *et al*. 2007). For WEMAP MMIs, RWB samples collected in the mountain ecoregion scored on average 7.2 points lower than TRB samples (paired t test, $p < 0.001$), but samples based on the two methods collected in the xeric ecoregion produced statistically indistinguishable scores ($p = 0.65$). Mountain WSA MMI values were also lower for RWB samples (4 points, $p = 0.02$), but RWB MMI values from the xeric region were higher than TRB samples by 6 points ($p = 0.002$). For the purpose of inter-index comparisons, these scoring biases were corrected by adding 7.2 points to the MMI values for those mountain WEMAP RWB samples used in comparisons (three sites), adding 4 points to values for the mountain WSA RWB samples, and subtracting 6 points from values for xeric WSA RWB samples (three sites). However, only TRB samples were used in remaining MMI analyses.

Precision

The northern and southern CA MMIs were more precise (reference site CVs = 0.14 and 0.19) than the WSA-West mountain and xeric MMIs (CVs = 0.26, 0.25), but comparable to those of the WEMAP mountain and xeric MMIs (CVs = 0.13, 0.23; Table 4). Associations among the rescaled MMI indices (CA vs. WSA-West $r^2 = 0.70$, CA vs. WEMAP $r^2 = 0.76$, and WSA-West vs. WEMAP $r^2 = 0.75$; Figure 7) were much stronger than we observed for the O/E indices (Figure 5).

Bias

The rescaled WSA-West MMI was a biased predictor of both the CA and WEMAP MMIs, with slopes significantly different ($p < 0.001$) from 1 (Figure 7). In addition, the WEMAP MMI on average produced higher scores at test sites than the CA MMI (Table 6). The WEMAP MMI rated low-scoring sites higher than the WSA-West MMI and high-scoring sites lower than the WSA-West MMI (Figure 7). However, most of these differences in MMI values were not associated with the natural gradients we considered, except for the significant relationships between CA and WEMAP pairwise differences and both elevation and watershed area (Figure 8).

Responsiveness

On average, the rescaled CA MMIs scored test sites lower than the rescaled WEMAP MMIs, which

Table 5. Regressions (y = a + bx) for relationships shown in Figures 6 and 10 where y is the difference between the index scores of two models and x is a natural gradient variable. Asterisks indicate significant slopes, y-intercepts, or r^2 values at p = 0.05 level (significance threshold not adjusted for multiple comparisons).

Index	Natural Gradient	Model Difference	b	p-value for b	a	p-value for a	r^2
O/E (n = 101)	Elevation	CA-WSA	-0.000043	0.283	-0.043	0.259	0.01
		CA-WEMAP	0.0000042	0.918	0.059	0.132	0
		WSA-WEMAP	0.000048	0.112	0.1	<0.001*	0.03
	log Watershed Area	CA-WSA	0.0029	0.928	-0.081	0.125	0
		CA-WEMAP	-0.025	0.424	0.1	0.06	0.01
		WSA-WEMAP	-0.028	0.23	0.18	<0.001*	0.01
	Percent Slope	CA-WSA	-0.016	0.019*	-0.017	0.606	0.05*
		CA-WEMAP	-0.015	0.035*	0.12	<0.001*	0.04*
		WSA-WEMAP	0.0015	0.77	0.13	<0.001*	0
	Percent Fastwater	CA-WSA	-0.0035	0.002*	0.023	0.543	0.09*
		CA-WEMAP	-0.0029	0.012*	0.14	0.001*	0.06*
		WSA-WEMAP	0.00064	0.458	0.12	<0.001*	0.01
MMI-rescaled (n = 68)	Elevation	CA-WSA	0.000047	0.586	-0.24	<0.001*	0
		CA-WEMAP	0.00012	0.041	-0.15	<0.001*	0.06
		WSA-WEMAP	0.000073	0.415	0.086	0.028*	0.01
	log Watershed Area	CA-WSA	-0.043	0.19	-0.13	0.105	0.03
		CA-WEMAP	-0.057	0.01	0.011	0.832	0.1
		WSA-WEMAP	-0.014	0.674	0.14	0.095	0
	Percent Slope	CA-WSA	0.0024	0.832	-0.23	<0.001*	0
		CA-WEMAP	0.011	0.151	-0.14	<0.001*	0.03
		WSA-WEMAP	0.0085	0.46	0.09	0.020*	0.01
	Percent Fastwater	CA-WSA	0.0021	0.182	-0.28	<0.001*	0.03
		CA-WEMAP	-0.00071	0.518	-0.1	0.004*	0.01
		WSA-WEMAP	-0.0028	0.086	0.18	<0.001*	0.04

in turn scored test sites lower than rescaled WSA-West MMIs (Table 6). This trend generally held for both mountain and xeric ecoregions, although the WSA-West vs. WEMAP mountain contrast was not significantly different. All MMIs tended to score test sites in the xeric ecoregion lower than test sites in the mountain ecoregion, although the difference in mean values based on the WSA-West MMI was not significant (Table 6).

Index sensitivity and concordance among assessments

As with the O/E indices, impairment decisions differed considerably among the rescaled MMI indices (Table 8). The number of sites assessed as impaired was far fewer for the WSA-West and WEMAP MMIs (21 and 17 sites of 68 total sites, respectively) than the CA MMI (39 of 68 sites, binomial tests, p <0.001). This pattern occurred in both xeric and mountain ecoregions but was only significant in the xeric ecoregions (binomial tests: mountain p = 0.219, xeric p <0.001).

Summary of WEMAP and WSA-WEST indices performance relative to CA indices

Differences in index precision, bias, and responsiveness can each contribute to differences in index performance as measured by index sensitivity, the likelihood that an assessment will identify impairment. In this study, assessment differences between WEMAP or WSA-West indices and CA indices depended on the type of index examined and specific differences in index precision, bias, and responsiveness (Table 9). Although the large-scale indices tended to lead to different inferences regarding biological condition than the CA indices, the specific differences

Table 6. Results of t-test comparisons for differences in index responsiveness between sets of mountainous and xeric test sites, or between model pairs. Mean 1 and Mean 2 indicate the mean scores of the first and second members of each tested pair. All MMI scores were rescaled by dividing scores by the appropriate reference mean.

Test Dataset	Comparison Type	Test group	Indices in Test	Mean 1	Mean 2	Difference	p (*significant α =0.0167)	Test (2-tailed)
O/E	***Index Comparison***	*Both Ecoregions (n = 127)*	CA vs. WSA	0.82	0.90	0.09	<0.001*	paired t-test
			CA vs. WEMAP	0.82	0.77	0.04	0.032	
			WSA vs. WEMAP	0.90	0.77	0.13	<0.001*	
		MTN only (n = 74)	CA vs. WSA	0.87	0.93	0.06	0.023	paired t-test
			CA vs. WEMAP	0.87	0.80	0.07	0.002*	
			WSA vs. WEMAP	0.93	0.80	0.13	<0.001*	
		XER only (n = 53)	CA vs. WSA	0.75	0.87	0.12	0.005*	paired t-test
			CA vs. WEMAP	0.75	0.74	0.00	0.938	
			WSA vs. WEMAP	0.87	0.74	0.12	<0.001*	
	Ecoregion Comparison	*MTN vs. XER*	CA	0.87	0.75	0.12	0.006*	2 sample t-test
			WSA	0.93	0.87	0.06	0.156	
			WEMAP	0.80	0.74	0.05	0.248	
MMI	***Index Comparison***	*Both Ecoregions (n = 68)*	CA vs. WSA	0.65	0.88	0.23	<0.001*	paired t-test
			CA vs. WEMAP	0.65	0.77	0.12	<0.001*	
			WSA vs. WEMAP	0.88	0.77	0.11	<0.001*	
		MTN only (n = 30)	CA vs. WSA	0.80	1.00	0.20	<0.001*	paired t-test
			CA vs. WEMAP	0.80	0.88	0.07	0.009*	
			WSA vs. WEMAP	1.00	0.88	0.13	0.018	
		XER only (n = 38)	CA vs. WSA	0.53	0.78	0.24	<0.001*	paired t-test
			CA vs. WEMAP	0.53	0.69	0.15	<0.001*	
			WSA vs. WEMAP	0.78	0.69	0.09	0.006*	
	Ecoregion Comparison	*MTN vs. XER*	CA	0.80	0.53	0.27	<0.001*	2 sample t-test
			WSA	1.00	0.78	0.23	0.0219	
			WEMAP	0.88	0.69	0.19	0.001*	

Table 7. Counts of CA sites declared impaired (I) and not impaired (NI) by CA O/E estimates and corresponding WEMAP and WSA O/E estimates. WSA-Adjusted: Impairment thresholds set at 5th percentile for each ecoregion.

		CA Sub-model 1 (n = 58)		CA Sub-model 2 (n = 44)		CA Sub-model 3 (n = 25)		Total (n = 127)		All Sites
		I	NI	I	NI	I	NI	I	NI	
CA	I	13	-	16	-	6	-	35	-	35
	NI	-	45	-	28	-	19	-	92	92
WEMAP	I	10	7	11	8	4	3	25	18	43
	NI	3	38	5	20	2	16	10	74	84
WSA	I	5	1	7	2	0	1	12	4	16
	NI	8	44	9	26	6	18	23	88	111
WSA-Adjusted	I	9	4	9	4	0	1	18	9	27
	N	4	41	7	24	6	18	17	83	100

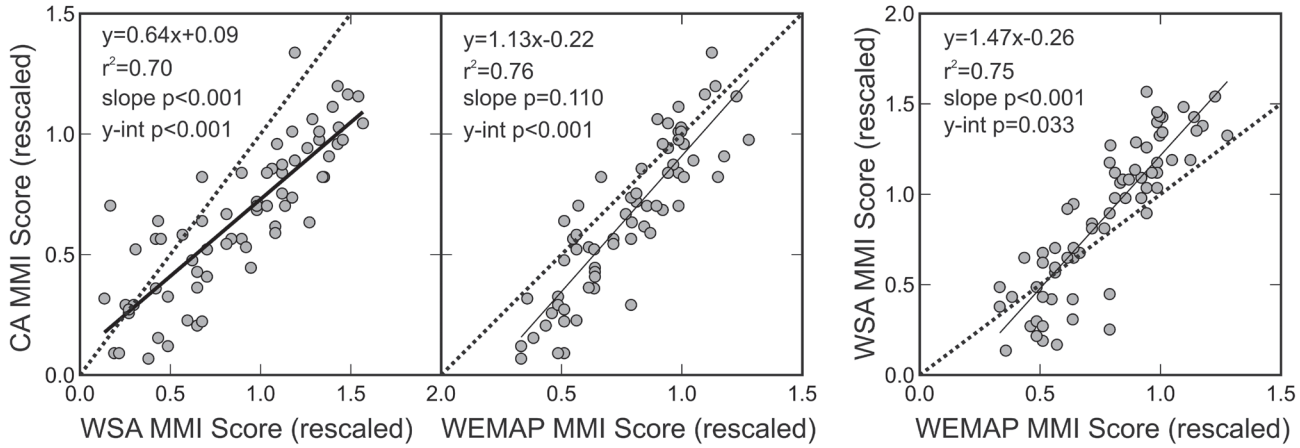

Figure 7. Regressions between rescaled scores at CA test sites between rescaled index scores for different combinations of the MMIs. The dashed diagonal lines represent perfect 1:1 relationship between the models, and the thick and thin solid lines indicate linear best-fit relationships. Significance tests are for y-intercept (y-int) = 0 and slope = 1.

among indices were variable. These differences lead to the WEMAP O/E index having similar sensitivity to the CA O/E indices, whereas the WSA-West O/E index was less sensitive. The difference between these two large-scale indices appeared to be largely associated with differences in their responsiveness. The MMI comparisons showed the opposite response in that the WEMAP MMI was slightly more sensitive than the CA MMI in mountain regions while the WSA-West MMI was less sensitive than the CA MMI in xeric regions. As we saw for the O/E comparisons, the differences between the WEMAP and WSA-West MMI sensitivities were also most clearly associated with differences in their responsiveness.

Discussion

The multiple spatial scales over which environmental gradients influence the taxonomic and functional composition of freshwater assemblages has been the focus of considerable interest in recent

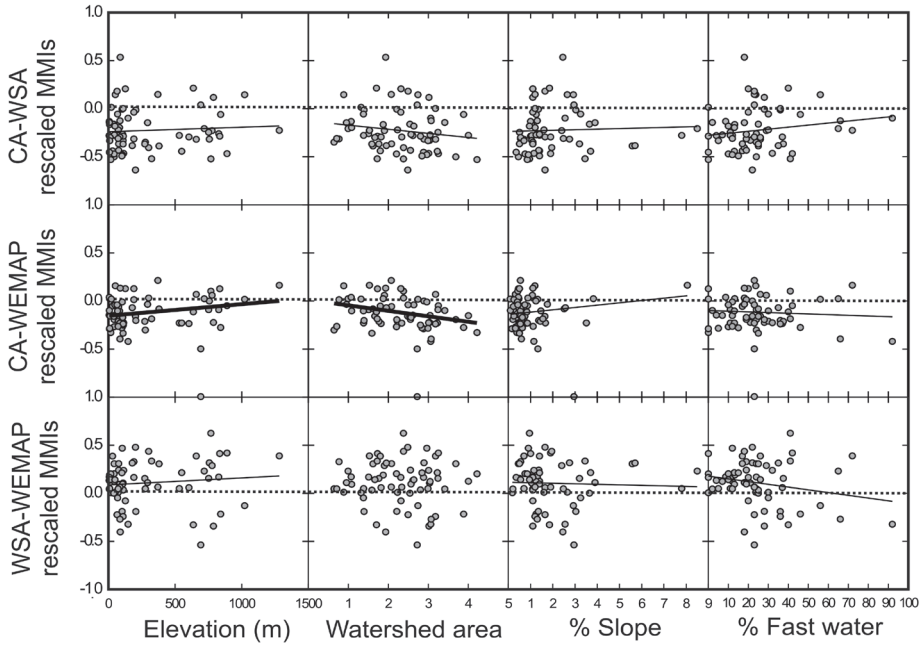

Figure 8. Scatterplots and regressions between the pair-wise differences in rescaled MMI values for the three different MMIs and four environmental gradients at CA test sites. The dashed horizontal lines represent zero difference. Thick solid lines denote regressions with r^2 and slopes significantly different from 0; thin solid lines denote those with intercepts significantly different from 0 but non-significant slope.

Table 8. Counts of CA sites declared impaired (I) and not impaired (NI) by CA MMI estimates and corresponding WEMAP and WSA MMI estimates.

		CA Mountain (n = 30)		CA Xeric (n = 38)		Total (n = 68)		All Sites
		I	NI	I	NI	I	NI	
CA	I	10	-	29	-	39	-	39
	NI	-	20	-	9	-	29	29
WEMAP	I	5	1	15	0	20	1	21
	NI	5	19	14	9	19	28	47
WSA	I	5	1	11	0	16	1	17
	NI	5	19	18	9	23	28	51

years (Poff 1997, Johnson *et al*. 2004, Johnson *et al*. 2007, Heino *et al*. 2007, Hoeinghaus *et al*. 2007, Mykrä *et al*. 2007, Mykrä *et al*. 2008). At the heart of all of these studies is a desire to clarify understanding of the factors that determine species distribution limits, one of the central goals of ecological theory (Levins 1966, Wiens 1989, Peters 1991, Brown *et al*. 1996, Guisan and Zimmermann 2000). This issue has significant implications for the utility of biotic indices because their effectiveness depends understanding how distribution patterns of individual taxa are influenced by landscape and waterway environmental heterogeneity, and how those effects are expressed at different scales of observation.

Index Comparability

O/E indices

Matching test sites with appropriate reference condition is a critical element of all bioassessments

Table 9. Summary of differences in precision, bias, responsiveness, and sensitivity of the WEMAP and WSA indices relative to CA indices. M = mountain ecoregion, X = xeric ecoregion. The term "similar" indicates no statistical difference; the terms "lower" and "higher" indicate the direction of a significant difference.

Performance Measure	O/E		MMI	
	WEMAP	WSA	WEMAP	WSA
Precision	Lower	Lower	Similar	Lower
Bias	Yes	Yes	Yes	Yes
Responsiveness	Lower	Lower	Lower	Lower
Sensitivity	Similar	Lower	Lower	Lower

(Moss *et al*. 1987, Hughes *et al*. 1995, Stoddard *et al*. 2007). Errors in specifying the correct reference condition can lead to either under- or over-estimates of the true biological condition at individual sites. Our results show that the failure of the large-scale predictive models to account for the effects of some naturally occurring environmental factors caused substantial systematic differences among the O/E values derived from these models relative and those derived from the CA models. The fact that the most spatially extensive models (WEMAP and WSA-West models) did not adjust for the effects of local environmental heterogeneity (i.e., slope, percent fast-water habitats) on E, and hence O/E, shows that such spatially extensive models may have limited applicability for site-specific assessments and use of these assessments to generate regional assessments. There are several reasons the more spatially extensive models may have failed to account for the effects of reach slope and percent fast water on assemblage composition. First, available map-derived variables may not have been good surrogates for these variables when used at large scales. For example, watershed area is likely related to one or more factors that influence taxa presence at a site, including channel slope and amount of fast-water habits (Hynes 1970, Allan and Castillo 2007). However, watershed area might not be consistently associated with channel slope across a region the size of the western United States. In the three sets of models we examined, watershed area appeared to account for differences among sites in channel reach for only the spatially less extensive CA models. Even in those models that used direct measures of channel slope as a predictor variable (e.g., the WSA-West model), the relationship between invertebrate taxa and slope may be obscured by strong relationships between invertebrate composition and predictors that vary markedly across regions, such as temperature and precipitation. Furthermore, a predictive model based on linear relationships between biotic composition and predictor variables will fail to accurately describe any non-linear relationships and hence inaccurately predict the taxa that should occur under specific states of that variable. In contrast, over a smaller range of environmental conditions, surrogate predictors such as watershed area, temperature, or precipitation may adequately capture differences between sites in local habitat features such as channel slope and type of habitat. In general, these problems of prediction bias might be reduced in the future by both improving how well reference site networks represent all

streams of interest (in terms of both sample size and types of streams) and by using robust predictors such as Random Forests (Cutler *et al*. 2007) that do not assume linear relationships.

The fact that the WSA-West model strongly underestimated impairment relative to the CA model has at least two possible explanations: 1) poorer precision in the WSA-West model resulted in lower impairment thresholds and thus fewer impairment decisions, and 2) WSA underestimated the probabilities of capture of some of the taxa that contribute to the O/E calculations. The second result could have arisen if the reference sites used to predict the fauna in California streams were less rich on average than the otherwise similar California sites assessed. Vinson and Hawkins (1996) reported that invertebrate taxa richness in streams draining mountainous regions of California (Coast Range Mountains and Sierra Nevada) was higher than streams draining other mountainous regions in the western USA. Models based on a mix of reference sites from across the western United States might therefore be expected to under-predict richness at CA mountain sites. This explanation seems plausible for the WSA-West model, because average WSA-West O/E values for CA mountainous reference sites were greater than 1 on average (Sierra Nevada = 1.04, Southern Coastal Mountains = 1.11, and Klamath Mountains = 1.04). However, WEMAP reference site O/E values did not exhibit this trend. It seems prudent that we should refine models to explicitly account for the effects of biogeographic history on taxa richness. Such modeling might be accomplished through the use of categorical predictive variables that classify sites by their relevant zoogeographic region rather than general purpose ecoregions (Hawkins and Vinson 2000, Hawkins *et al*. 2000b). The contrasting result for the WEMAP model (i.e., that WEMAP model did not underestimate impairment relative to the CA model despite precision values intermediate between the CA and WSA models) is likely the consequence of the tendency of the WEMAP model to score sites lower than the WSA model.

Multimetric indices

Although, agreement among the MMI scores was considerably stronger than for the O/E indices, the relationships between scores were not consistent across the scoring range, indicating differences in responsiveness of the indices at low vs. high biotic condition sites. Also, although the WEMAP and WSA-West MMIs were derived from nearly identical datasets, there were numerous differences in the performance of the two larger MMIs, including precision, responsiveness and sensitivity. These differences reflect the different approaches used to develop the MMIs (Ode *et al*. 2005; Rehn *et al*. 2005; Stoddard *et al*. 2005, 2008).

Differences in MMI responsiveness were likely caused by one or more of the following: 1) differences in how metrics were scaled in the separate indices, 2) differences in the quality of sites used to calibrate the indices, or 3) differences in how individual metrics in each MMI respond to stress. Because there was considerable overlap in metrics among the indices, much of the difference among the MMIs in their assessments probably lies in differences in the scoring ranges of specific metrics. For example, although the number of EPT taxa is a nearly ubiquitous metric in MMIs (Karr and Chu 1999), the scoring range for this metric varies among regions. An EPT scoring range established from reference site data combined across a large spatial extent will not necessarily reflect local reference conditions. In some regions, test sites will be underscored; in others they will be overscored. We found evidence of this effect in the number of disagreements in impairment decisions made under the different MMIs. Furthermore, the WSA-West MMI did not indicate a difference in biotic condition between mountain and xeric test sites, whereas the CA and WEMAP MMI did. This finding was echoed in the way impairment decisions differed between WEMAP and WSA-West indices in xeric and mountain regions. Both WEMAP and WSA-West MMIs tended to overestimate impairment at mountain sites relative to the CA MMI, whereas the WSA-West MMI underestimated impairment at xeric sites relative to the CA MMI.

A final potential explanation is that differences in MMI performance were related to differences in the calibration sets used to derive the metric scoring ranges. Because MMIs are calibrated with both reference and test data, any difference in the biological quality of either set of calibration sites can affect a site's scoring, just as they can in O/E models (Hawkins 2006). Because of incomplete information regarding the quality of reference and test sites used to calibrate the different indices, how seriously such differences affected index performance could not be addressed at this time.

Effects of spatial scale on index performance

It has been long known that taxonomic composition is influenced by natural environmental gradients. How these relationships are expressed at different spatial scales, and hence affect biological indices, is much less clear, but is of increasing interest (Finn and Poff 2005, Heino *et al*. 2007, Cao *et al*. 2007, Mykrä *et al*. 2008). MMIs and predictive models use different methods for accounting or adjusting for natural gradients. Predictive models are explicitly designed to describe how natural environmental gradients affect the distribution of individual taxa (Wright *et al*. 1989, 2000). However, some natural gradients may be important at certain geographic scales, but cease to matter at other scales, as shown in this study and elsewhere (Mykrä *et al*. 2008).

In contrast to O/E indices, MMIs attempt to minimize the effects of natural gradients by a priori classification of reference sites into environmentally homogeneous sets of sites. In addition, metrics are selected to be insensitive to natural gradients, or by adding correction factors that adjust for scoring differences along gradients (Karr and Chu 1999). In this study, for example, scoring ranges for the EPT richness metric varied little across spatial scales within ecoregions (Ode *et al*. 2005; Rehn *et al*. 2005; Stoddard *et al*. 2005, 2008), and the CA MMI for the North Coast explicitly corrects for watershed area in affected metrics (Rehn *et al*. 2005).

In this study, the large-scale predictive models were not completely successful in adjusting for two of the gradients (percent slope and percent fastwater habitats) we examined. Likewise, the CA and WSA-West MMIs were not completely effective at controlling for an elevation gradient.

Index performance and model traits

All the biological indices in our evaluations produce scores by comparing biological expectations to observed biology. Although E in O/E is explicitly modeled (i.e., predicted), MMI expectations are derived from a set of reference sites that are grouped (by ecoregion, stream size, etc.) to maximize similarity of the biological assemblages at reference sites. Thus, both O/E and MMI are indices based on modeled expectations. Levins (1966) postulated that there is an inherent tradeoff among three desirable model traits: reality (i.e., accuracy, or lack of bias), precision, and generality (see also Guisan and Zimmermann 2000). Although these model traits are not necessarily mutually exclusive, we cannot expect the models used to predict biotic conditions to optimize each trait. In creating standardized indices applicable across a large range of geoclimatic conditions, generality was improved at the expense of both reality and precision. This tradeoff points to the need to develop more localized models for bioassessment programs, especially those that use biocriteria to infer if streams are supporting their designated aquatic life uses. However, the fact that impairment decisions can be very sensitive to the thresholds used to define impaired conditions (as seen when an ecoregion-based correction was applied to the WSA-West model for O/E comparisons), suggests that it may be possible to adjust for some of the systematic differences among the models. Larger models could be rendered more suitable for local application by calibrating impairment thresholds to local reference conditions. In practice, a local regulatory entity could recalculate the standard deviations for O/E or MMI models based only on local reference sites and use these to set locally relevant thresholds.

Concluding Remarks

The answer to the central question of whether indices developed from geoclimatically extensive data can substitute for more locally produced indices depends both on their intended use and the type of indicator. In regional condition assessments, accuracy (lack of bias) is more important than precision. That is, for low precision can be compensated by looking at large numbers of samples with the expectation that the estimated average condition will still be accurate. For the purpose of regional assessments, use of the WEMAP O/E index produced results that were generally comparable to the CA indices. In contrast, because of its strong bias, the WSA-West O/E index would probably underestimate regional impairment. Likewise, lower precision and differences in responsiveness across the scoring range make the WSA-West MMIs less desirable for regional condition assessments.

For site-specific assessments, where both accuracy and precision are important, it seems clear that locally derived indices should outperform large-scale indices for both types of index (see also Mykrä *et al*. 2008). Because most applications of bioassessment tools are site-specific, there is a clear need to continue to develop regional models that explicitly take locally important gradients into account (Heino *et al*. 2007). However, because the WEMAP MMI had similar precision and WEMAP MMI scores were

highly correlated with CA MMI scores, the WEMAP MMI might provide an acceptable substitute in California (and potentially other regions in the western US) until local MMIs are developed, assuming care is taken to adjust impairment thresholds to reflect local reference conditions.

Finally, these results suggest three related applied research needs: 1) identifying the geographic or geoclimatic scale that optimizes index performance, 2) determining the factors that most strongly influence index performance and identifying the geographic scales at which they vary, and 3) identifying ways of more accurately specifying the reference condition from geoclimatically extensive sets of reference site data. It is not known much about which factors influence the optimal geographic scale for producing either predictive models or multimetric indices, but the rapidly expanding field of bioassessment would benefit greatly from the ability to predict these factors.

Literature Cited

Allan, J.D. and M.M. Castillo. 2007. Stream Ecology: Structure and Function of Running Waters. 2nd Edition. Kluwer. Dordrecht, The Netherlands.

Andrewartha, H.G. and L.C. Birch. 1954. The Distribution and Abundance of Animals. The University of Chicago Press. Chicago, IL.

Bailey, C.R., R.H. Norris and T.B. Reynoldson. 2004. Bioassessment of Freshwater Ecosystems: Using the Reference Condition Approach. Kluwer. Dordrecht, The Netherlands.

Barbour, M.T. and C.O. Yoder. 2000. The multimetric approach to bioassessment, as used in the United States. pp. 281-292 *in*: J.F. Wright, D.W. Sutcliffe and M.T. Furse (eds.), Assessing the Biological Quality of Fresh Waters: RIVPACS and Other Techniques. Freshwater Biological Association. Ambleside, United Kingdom.

Bonada, N., N. Prat, V.H. Resh and B. Statzner. 2006. Development in aquatic insect biomonitoring: a comparative analysis of recent approaches. *Annual Review of Entomology* 51:495-523.

Brown, J.H., G.C. Stevens and D.M. Kaufman. 1996. The geographic range: size, shape, boundaries, and internal structure. *Annual Review of Ecology and Systematics* 27:597-623.

Cao, Y., C.P. Hawkins, J.Olson and M. Kosterman. 2007. Modeling natural environmental gradients improves the accuracy and precision of diatom-based indicators. *Journal of the North American Benthological Society* 26:566-585.

Clarke, R.T., J.F. Wright and M.T. Furse. 2003. RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. *Ecological Modelling* 160:219-233.

Cutler, D.R., T.C. Edwards, Jr., K.H. Beard, A. Cutler, K.T. Hess, J. Gibson and J.J. Lawler. 2007. Random forests for classification in ecology. *Ecology* 88:2783-2792.

Environmental Protection Agency (EPA). 2006. Wadeable Streams Assessment: A Collaborative Survey of the Nation's Streams. EPA 841-B-06-002. EPA Yosemite. Yosemite, CA.

Finn, D.S. and N.L. Poff. 2005. Variability and convergence in benthic communities along the longitudinal gradients of four physically similar Rocky Mountain streams. *Freshwater Biology* 50:243-261.

Guisan, A. and N.E. Zimmermann. 2000. Predictive habitat distribution models in ecology. *Ecological Modeling* 135:147-186.

Hawkins, C.P. 2006. Quantifying biological integrity by taxonomic completeness: evaluation of a potential indicator for use in regional- and global-scale assessments. *Ecological Applications* 16:1251-1266.

Hawkins, C.P., R.H. Norris, J.N. Hogue and J.M. Feminella. 2000a. Development and evaluation of predictive models for measuring the biological integrity of streams. *Ecological Applications* 10:1456-1477.

Hawkins, C.P., J. Ostermiller, M. Vinson, R.J. Stevenson and J. Olson. 2003. Stream algae, invertebrate, and environmental sampling associated with biological water quality assessments. Western Center for Monitoring and Assessment of Freshwater Ecosystems. Utah State University. Logan, UT.

Hawkins, C.P., R.H. Norris, J. Gerritsen, R.M. Hughes, S.K. Jackson, R.H. Johnson and R.J. Stevenson. 2000b. Evaluation of landscape classifications for biological assessment of freshwater ecosystems: synthesis and recommendations.

Journal of the North American Benthological Society 19:541-556.

Hawkins, C.P. and M.R. Vinson. 2000. Weak correspondence between landscape classifications and stream invertebrate assemblages: implications for bioassessments. *Journal of the North American Benthological Society* 19:501-517.

Heino, J., H. Mykrä, J. Kotanen and T. Muotka. 2007. Ecological filters and variability in stream macroinvertebrate communities: do taxonomic and functional structure follow the same path? *Ecography* 30:217-230.

Herlihy, A.T., S. Paulsen, J. Van Sickle, J. Stoddard, C.P. Hawkins and L.L. Yuan. 2008. Striving for consistency in a national assessment: the challenges of applying a reference condition approach at a continental scale. *Journal of the North American Benthological Society* 27:860-877.

Hoeinghaus, D.J., K.O. Winemiller and J.S. Birnbaum. 2007. Local and regional determinants of stream fish assemblage structure: inferences based on taxonomic vs. functional groups. *Journal of Biogeography* 34:324-338.

Hughes, R.M. 1995. Defining acceptable biological status by comparing with reference conditions. pp. 31-47 *in*: W.S. Davies and T.P. Simon (eds.), Biological assessment and criteria: Tools for water resource planning and decision making. Lewis Publishers. Ann Arbor, MI.

Hughes, R.M., P.R. Kaufmann, A.T. Herlihy, T.M. Kincaid, L. Reynolds and D.P. Larsen. 1998. A process for developing and evaluating indices of fish assemblage integrity. *Canadian Journal of Fisheries and Aquatic Sciences* 55:1618-1631.

Hutchinson, G.E. 1959. Homage to Santa Rosalia; or, why are there so many kinds of animals. *American Naturalist* 93:145-159.

Hynes, H.B.N. 1970. The Ecology of Running Waters. University of Toronto Press. Toronto, Canada.

Johnson, R. K., M.T. Furse, D. Hering and L. Sandin. 2007. Ecological relationships between stream communities and spatial scale: implications for designing catchment-level monitoring programs. *Freshwater Biology* 52:939-958.

Johnson, R.K, W. Goedkoop and L. Sandin. 2004. Spatial scale and ecological relationships between the macroinvertebrate communities of stony habitats of streams and lakes. *Freshwater Biology* 49:1179-1194.

Karr, J.R. 1981. Assessment of biotic integrity using fish communities. *Fisheries* 6:21-27.

Karr, J.R. and E.W. Chu. 1999. Restoring Life in Running Waters: Better Biological Monitoring. Island Press. Washington, DC.

Kerans, B.L. and J.R. Karr. 1994. A benthic index of biotic integrity (B-IBI) for rivers of the Tennessee Valley. *Ecological Applications* 4:768-785.

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, M.B. Griffith and W.S. Davis. 2003. Development and evaluation of a macroinvertebrate biotic integrity index (MBII) for regionally assessing Mid-Atlantic Highland streams. *Environmental Management* 31:656-669.

Levins, R. 1966. The strategy of model building in population ecology. *American Scientist* 54:421-431.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. pp. 281-297 *in*: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. Berkeley, CA.

McCormick, F.H., R.M. Hughes, P.R. Kaufmann, D.V. Peck, J.L. Stoddard and A.T. Herlihy. 2001. Development of an index of biotic integrity for the Mid-Atlantic Highlands Region. Transactions of the *American Fisheries Society* 130:857-877.

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.

Mykrä, H., J. Aroviita, J. Kotanen and H. Hämäläinen. 2008. Predicting the stream macroinvertebrate fauna across regional scales: influence of geographical extent on model performance. *Journal of the North American Benthological Society* 27:705-716.

Mykrä, H., J. Heino and T. Muotka. 2007. Scale-related patterns in spatial and environmental compo-

nents of stream macroinvertebrate assemblage variation. *Global Ecology and Biogeography* 16:149-159.

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.

Omernik, J.M. 1987. Ecoregions of the conterminous United States Map (scale 1:7,500,000). *Annals of the Association of American Geographers* 77:118-125.

Ostermiller, J.D. and C.P. Hawkins. 2004. Effects of sampling error on bioassessments of stream ecosystems: application to RIVPACS-type models. *Journal of the North American Benthological Society* 23:363-382.

Peck, D.V., A.T. Herlihy, B.H. Hill, R.M. Hughes, P.R. Kaufmann, D. Klemm, J.M. Lazorchak, F.H. McCormick, S.A. Peterson, P.L. Ringold,T. Magee and M. Cappaert. 2006. Environmental Monitoring and Assessment Program-Surface Waters Western Pilot Study: Field Operations Manual for Wadeable Streams. EPA/620/R-06/003. US Environmental Protection Agency. Washington, DC.

Peters, R.H. 1991. A Critique for Ecology. Cambridge University Press. Cambridge, MA.

Poff, N.L. 1997. Landscape filters and species traits: towards mechanistic understanding and prediction in stream ecology. *Journal of the North American Benthological Society* 16:391-409.

Rehn, A.C., P.R. Ode and C.P. Hawkins. 2007. Comparisons of targeted-riffle and reach-wide benthic macroinvertebrate samples: implications for data sharing in stream-condition assessments. *Journal of the North American Benthological Society* 26:332-348.

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) assessment. Report to the State Water Resources Control Board. California Department of Fish and Game Aquatic Bioassessment Laboratory. Rancho Cordova, CA.

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.

Richards, A.B. and D.C. Rogers. 2006. List of freshwater macroinvertebrate taxa from California and adjacent states including standard taxanomic effort levels. Southwest Association of Freshwater Invertebrate Taxonomists. http://www.waterboards.ca.gov/swamp/docs/safit/ste_list.pdf.

Simpson, J.C. and R.H. Norris. 2000. Biological assessment of river quality: development of AUSRIVAS models and outputs. pp. 125-142 *in*: J.F. Wright, D.W. Sutcliffe and M.T. Furse (eds), Assessing the Biological Quality of Fresh Waters: RIVPACS and Other Techniques. Freshwater Biological Association. Amblesite, United Kingdom.

Statzner, B., B. Bis, S. Dolédec and P. Usseglio-Polatera. 2001. Perspectives for biomonitoring at large spatial scales: a unified measure for the functional classification of invertebrate communities in European running waters. *Basic and Applied Ecology* 2:73-85.

Stoddard, J.L., A.T. Herlihy, D.V. Peck, R.M. Hughes, T.R. Whittier and E. Tarquino. 2008. A process for creating multi-metric indices for large scale aquatic surveys. *Journal of the North American Benthological Society* 27:878-891.

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. 2006. Environmental Monitoring and Assessment Program (EMAP) Western Streams and Rivers Statistical Summary. EPA 620/R-05/006. US Environmental Protection Agency. Washington, DC.

Stoddard, J.L., D.V. Peck, S.G. Paulsen, J. Van Sickle, C.P. Hawkins, A.T. Herlihy, R.M. Hughes, P.R. Kaufmann, D.P. Larsen, G. Lomnicky, A.R. Olsen, S.A. Peterson, P.L. Ringold and T.R. Whittier. 2005. An Ecological Assessment of Western Streams and Rivers. EPA 620/R-05/005. US Environmental Protection Agency. Washington, DC.

Van Sickle, J., C.P. Hawkins, D.P. Larsen and A.T. Herlihy. 2005. A null model for the expected macroinvertebrate assemblage in streams. *Journal of the North American Benthological Society* 24:178-191.

Van Sickle, J., D.P. Larsen and C.P. Hawkins. 2007. Exclusion of rare taxa affects performance of the O/E index in bioassessments. *Journal of the North American Benthological Society* 26:319-331.

Van Sickle, J. and S.G. Paulsen. 2008. Assessing the attributable risks, relative risks, and regional extents of aquatic stressors. *Journal of the North American Benthological Society* 27:920-931.

Vinson, M.R. and C.P. Hawkins. 1996. Effects of sampling area and subsampling procedure on comparisons of taxa richness among stream. *Journal of the North American Benthological Society* 15:392-399.

Wiens, J. 1989. Spatial scaling in ecology. *Functional Ecology* 3:385-397.

Wright, J.F., P.D. Armitage, M.T. Furse and D. Moss. 1989. Prediction of invertebrate communities using stream measurements. *Regulated Rivers: Research and Management* 4:147-155.

Wright, J.F., D.W. Sutcliffe and M.T. Furse. 2000 Assessing the Biological Quality of Fresh Waters: RIVPACS and Other Techniques. Freshwater Biological Association. Ambleside, United Kingdom.

Yoder, C.O. and E.T. Rankin. 1995. The role of biological criteria in water quality monitoring, assessment and regulation. Technical Report MAS/1995-1-3. Ohio EPA. Columbus, OH.

Yuan, L.L., C.P. Hawkins and J. VanSickle. 2008. Effects of regionalization decisions on an O/E index for the national assessment. *Journal of the North American Benthological Society* 27:892-205.

Zar, J.H. 1999. Biostatistical Analysis (4th edition). Prentice Hall. Englewood Cliffs, NJ.

Acknowledgements

The authors would like to thank staff of the California Department of Fish and Game Aquatic Bioassessment Laboratory who collected the field data and identified the benthic invertebrates used in the test datasets and Alan Herlihy who provided helpful information about WEMAP and WSA-West model development. Funding for the EMAP and CMAP programs was provided by the EPA (Assistance Agreement No. CR82823801) and the California State Water Resources Control Board's (SWRCB) Non Point Source Program (Agreement Number 03-273-250-2). This study was supported by the SWRCB's Surface Water Ambient Monitoring Program. CPH was supported, in part, by USEPA Science To Achieve Results (STAR) grants R-82863701 and R-83059401 and a contract with Region 5 of the USDA Forest Service. This manuscript was greatly improved by comments from John Van Sickle, two anonymous reviewers, and Kerry Ritter.